2010

A Bayesian approach to the semi-analytic model of galaxy formation: methodology

Yu Lu
University of Massachusetts Amherst

H.J. Mo
University of Massachusetts Amherst

Martin D. Weinberg
University of Massachusetts Amherst, weinberg@astro.umass.edu

N Katz
University of Massachusetts - Amherst, nsk@astro.umass.edu

Mon. Not. R. Astron. Soc. **000**, ??–?? (2010) Printed 16 April 2010 (MN LaTeX style file v2.2)

A Bayesian approach to the semi-analytic model of galaxy formation: methodology

Yu Lu*, H.J. Mo, Martin D. Weinberg, Neal S. Katz

Department of Astronomy, University of Massachusetts, Amherst, MA 01003-9305, USA

ABSTRACT

We believe that a wide range of physical processes conspire to shape the observed galaxy population but we remain unsure of their detailed interactions. The semi-analytic model (SAM) of galaxy formation uses multi-dimensional parameterizations of the physical processes of galaxy formation and provides a tool to constrain these underlying physical interactions. Because of the high dimensionality, the parametric problem of galaxy formation may be profitably tackled with a Bayesian-inference based approach, which allows one to constrain theory with data in a statistically rigorous way.

In this paper, we develop a generalized SAM using the framework of Bayesian inference. We show that, with a parallel implementation of an advanced Markov-Chain Monte-Carlo algorithm, it is now possible to rigorously sample the posterior distribution of the high-dimensional parameter space of typical SAMs. As an example, we characterize galaxy formation in the current ΛCDM cosmology using stellar mass function of galaxies as observational constraints. We find that the posterior probability distribution is both topologically complex and degenerate in some important model parameters. It is common practice to reduce the SAM dimensionality by fixing various parameters. However, this can lead to biased inferences and to incorrect interpretations of data owing to this parameter covariance. This suggests that some conclusions obtained from early SAMs may not be reliable. Using synthetic data to mimic systematic errors in the stellar mass function, we demonstrate that an accurate observational error model is essential to meaningful inference.

Key words: galaxies: formation — methods: numerical

* E-mail: luyu@astro.umass.edu

1 INTRODUCTION

In our current paradigm of structure formation, the matter density of the Universe is dominated by the cold dark matter (hereafter CDM), and galaxy formation is a two-stage process (e.g. White & Rees 1978). First, small perturbations in the density field originating from quantum fluctuations in the early universe grow and produce a population of virialized dark matter halos. Second, the baryonic matter associated with these halos accumulates at the halo centers due to cooling and cold flows, forming stars and galaxies. Because of the hierarchical nature of structure formation in a CDM cosmogony, dark matter halos merge. The halo mergers eventually lead to galaxy-galaxy mergers, resulting in the formation of elliptical galaxies.

The first stage of this process, the formation and virialization of dark matter halos, has been studied in great detail using the (extended) Press-Schechter formalism (e.g. Press & Schechter 1974; Bond et al. 1991; Lacey & Cole 1993), spherical and ellipsoidal collapse (e.g. Gunn & Gott 1972; Fillmore & Goldreich 1984; Bertschinger 1985; Sheth et al. 2001; Lu et al. 2006) and numerical simulations (e.g. Efstathiou et al. 1985; Navarro et al. 1997; Bullock et al. 2001a,b; Zhao et al. 2003a,b; Springel 2005; Macciò et al. 2007; Zhao et al. 2009). These studies have yielded the mass function, spatial distribution, formation history, and internal structure of the CDM halo population and serve as the backbone of study of galaxy formation. The knowledge of the second stage of galaxy formation is far less well established, mainly because the baryonic processes involved (cooling, star-formation and feedback) are poorly understood. Additional physical processes whose importance is not fully understood include dynamical friction, tidal stripping, black hole formation and accretion, and adiabatic contraction.

Hydrodynamic simulations can now be used to study galaxy formation and evolution in a full cosmological context (e.g. Katz 1992; Navarro & White 1993; Kereš et al. 2005; Oppenheimer & Davé 2006; Simha et al. 2009). However, computational power is still a severe limitation at the present, and one has to compromise between simulation resolution and box size. Because of this, an alternative approach, the semi-analytical model of galaxy formation, has been developed and widely adopted to study the statistical properties of the galaxy population (e.g. White & Frenk 1991; Kauffmann et al. 1993; Mo et al. 1998; Somerville & Kolatt 1999; Cole et al. 2000; Kang et al. 2005; Croton et al. 2006; Dutton & van den Bosch 2009). In the semi-analytical model (hereafter SAM), one adopts "recipes" to describe and parameterize the underlying physical ingredients, such as star formation and feedback. The free parameters in the models are then tuned to reproduce certain observational data of the galaxy population, such as stellar mass functions, color-magnitude relations, metallicity-stellar mass relations, Tully-Fisher relation, and two-point statistics that describe the spatial distribution of galaxies (e.g. two-point correlation function, pairwise peculiar velocity dispersion, etc.). However, the theory of galaxy formation and evolution still faces several outstanding problems (see Primack 2009, for an up-to-date review). For example, it remains challenging to fit the faint-end slope of the galaxy luminosity function (e.g. Benson & Madau 2003; Mo et al. 2005), and the models typically predict disk rotation velocities that are too high, unless adiabatic contraction and/or disk self-gravity are ignored (e.g. Cole et al. 2000; Dutton et al. 2007). In addition, the models have problems matching the evolution of the galaxy mass function with redshift (e.g. De Lucia & Blaizot 2007; Somerville et al. 2008; Fontanot et al. 2009), and typically overpredict the fraction of red satellite galaxies (Baldry et al. 2006; Weinmann et al. 2006; Kimm et al. 2009; Liu et al. 2010). There are three main reasons for these problems. First and foremost, current models most likely miss some vital ingredients or the recipes used do not properly implement the physical mechanism. Secondly, sub-space features and degeneracies in the model parameter space have been either missed or not sufficiently explored (Liu et al. 2010; Neistein & Weinmann 2009). Thirdly, the difficulties may actually reflect inconsistencies in the data themselves (so-called "systematic" errors). For example, it has been pointed out that the observed evolution in the stellar mass function is inconsistent with the observed cosmic star formation history (Fardal et al. 2007; Primack et al. 2008).

In order to address these problems, one must quantitatively characterize the model constraints implied by existing data sets as well as explore a wider range of models. The SAM approach provides a promising avenue to tackle these problems owing to its flexibility in implementation and its relatively fast speed in computation. However, significant changes in the methodology must be made in order to fully utilize the potential of the SAM. The main shortcoming in the current SAMs is that they are not probabilistically rigorous. In many published SAM applications, a subset of model parameters is held fixed while other parameters are adjusted to match some observational properties. If the match is unsatisfactory, one further adjusts some of the parameters or changes the model parameterization until a "good" fit is achieved. However, the goodness of fit is often assessed "by eye"; one overlays the model prediction, a luminosity function for example, on the observed result to see if the prediction is sufficiently close to the data. Since the statistical uncertainties in both the data and the model are not consistently computed, confidence levels do not follow. Similarly, since the model parameters are explored by hand, marginal probability can not be computed. As mentioned earlier, a number of physical processes in galaxy formation are still poorly understood, and so the parameterizations of these processes have to be made very general. This leaves a large parameter space to be probed. Given the high dimensionality of the parameter space and the complex covariance between parameters, it is almost impossible to find and delineate the dominant mode by hand-tuning model parameters. For example, if a good fit between model and data is not found, the relative significance of the parameter region relative to others remains unknown. Third, since the model parameters may be strongly covariant, the effect of changing one model parameter while keeping others fixed is conditional to the values of other parameters that have been kept fixed. Therefore, switching on and off a process in a fiducial model is unlikely to figure out its importance to galaxy formation. Indeed, in order to investigate the influence of a specific recipe, one should allow the parameters to range over all their entire a priori plausible domain. Unfortunately, this kind of analysis has been missing in the current SAMs. Fourth, because many processes in galaxy formation are still poorly understood, different SAMs may adopt different parameterizations for the same process. While all these models can be tuned to match a limited set of observational data, it is difficult to judge which model is actually preferred. In principle, each of the parameterizations should be considered as a subset of a more general parameterization, and model selections should be made according to statistical evidence. Again, such an analysis is not included in the current SAMs.

In summary, a variety of physical processes affecting galaxy formation are not yet well understood while copious observational data constrain the models. Thus, in order to derive meaningful constraints from observations, we would like to know the probability of the various model parameters and, indeed, entire model families given the data. This leads us directly to Bayesian inference! The semi-analytical model provides a very powerful tool to translate the theory of galaxy formation into a set of model parameters. The Bayesian approach will then allow us to obtain the posterior distribution of the model parameters for a given set of data and to assess how a particular model is supported by the data. Moreover, given different model families, Bayesian model comparison techniques such as Bayes Factors and Reversible Jump techniques (Green 1995) allow one to determine the relative odds for each model producing the observed data.

Some attempts have been made recently in this direction. For instance, Benson & Bower

(2010) have performed an exhaustive search of model parameter space using Latin hypercube sampling (Mckay et al. 1979) to find a region which reproduces good fit to particular observations. Bower et al. (2010) have explored the parameter space of GALFORM model (Bower et al. 2006) using model emulator technique. Kampakoglou et al. (2008) and Henriques et al. (2008) have adopted Markov-Chain Monte-Carlo (MCMC) technique to explore the capability of their adopted SAMs in accommodating multiple observational data sets. However, since these authors considered only a very restricted set of models and a subset of the total parameter space, their explorations are limited to specific cases, and no general conclusions can be reached from them. Indeed, in order to overcome the shortcomings in the current SAMs, one needs a general semi-analytical model with a wide range of variations both in the parameterizations of physical processes and in model parameters, and a rigorous Bayesian approach that can determine the posterior probability over a large parameter space.

In this paper, we develop a scheme to incorporate SAM into the framework of Bayesian inference. To this end, our parameterizations are designed so that the corresponding SAM contains a number of published SAMs as subsets. This is important, because we want to explore a large model space while incorporating the findings of previous investigations. We also show that, aided with advanced MCMC techniques and moderate computational facilities, it is now possible to build a Bayesian inference-based SAM to efficiently explore the high dimensional parameter space involved and to establish the posterior distribution of model parameters reliably.

The goal of the present paper is a description of our approach and a demonstration of key points with simple examples. We will illustrate the limitations of the conventional SAM approach and the advantages of a Bayesian inference-based SAM. In particular, we will show that the common practice of tuning some model parameters while keeping others fixed can lead to an incorrect inference and that our Bayesian inference-based SAM overcomes this problem. We will also demonstrate sensitivity of the resulting inference to error model for data. The paper is organized as follows. In Section 2, we describe our generalized SAM and its relations to other models. A brief introduction to the principle of the Bayesian inference and the MCMC technique is presented in Section 3. In Section 4, we show a case study using the stellar mass function of galaxies as the observational constraint. The impacts of prior assumptions and data modeling on the model inference are presented in Section 5 and 6, respectively. Finally, in Section 7, we discuss and summarize our main results.

2 SEMI-ANALYTIC MODEL

As all other SAMs, our model consists of two main parts, (i) the assembly of individual dark matter halos, and (ii) gas, radiative and star-formation processes relevant to galaxy formation. We first prepare a large set of halo merger trees with the currently favored cosmology and adopt it for all subsequent semi-analytical modeling of the baryonic processes. Since the formation of dark matter halos is now relatively well understood, we focus on the baryonic physics in our Bayesian analysis.

2.1 Halo merger history

Halo merger trees can either be extracted from cosmological N-body simulations (e.g. Kang et al. 2005; Croton et al. 2006), or generated by a Monte-Carlo method using the extended Press-Schechter formalism (Lacey & Cole 1993; Somerville & Kolatt 1999; Cole et al. 2000; van den Bosch 2002). Merger trees from simulations provide the dynamics and environments of the halo population, but their construction is computationally expensive and limited by numerical resolution. On the other hand, Monte-Carlo merger trees are computationally cheaper to generate and have, in principal, infinite resolution. In this paper, we adopt the algorithm proposed by Parkinson et al. (2008) to generate the merger trees for halos with a given final ($z = 0$) virial mass. This algorithm has been tuned to match the conditional mass functions found in N-body simulations. More specifically, as an demonstration we choose the control parameters $G_0 = 1$, $\gamma_1 = \gamma_2 = 0$, so that the resulting halo conditional mass functions are those predicted by the Extended Press-Schechter conditional mass function (Parkinson et al. 2008). We sample a certain number of merger trees in each halo mass bin from $10^{10}\,h^{-1}\mathrm{M}_{\odot}$ to $10^{15}\,h^{-1}\mathrm{M}_{\odot}$, the mass range relevant to the modeling in this paper. Since the halos and their merger trees are randomly sampled from the halo mass function and the conditional mass function, the model prediction based on a finite merger tree sample suffers from sampling variance. In order to reduce such sampling variance, we generate a sufficiently large number of halo merger trees in each mass bin, so that the variance in the model prediction induced by merger-tree sampling is much smaller than the error in the observational data used to constrain the model and can be ignored. Specifically, we use 1000 merger trees for halos with present masses in the range 10^{11} - $10^{12.5}\,h^{-1}\mathrm{M}_{\odot}$, 1500 merger trees in the range $10^{12.5}$ - $10^{13.5}\,h^{-1}\mathrm{M}_{\odot}$, 400 merger trees in the range 10^{10} - $10^{11}\,h^{-1}\mathrm{M}_{\odot}$, and about 100 merger trees in the range $10^{13.5}$ - $10^{15}\,h^{-1}\mathrm{M}_{\odot}$. Since massive halos are rare in the assumed cosmology, their contribution to the scatter of the stellar mass function is negligible. The mass resolution of merger tree changes with the final halo mass. For halos with final masses smaller than $10^{12}\ h^{-1}\mathrm{M}_\odot$, the mass resolution is $10^{9.3}\ h^{-1}\mathrm{M}_\odot$; for halos with final masses larger than $10^{14}\ h^{-1}\mathrm{M}_\odot$, it is $10^{11}\ h^{-1}\mathrm{M}_\odot$; and for the intermediate halos, it is $10^{10}\ h^{-1}\mathrm{M}_\odot$. All merger trees are sampled at 60 snapshots equally spaced in $\log(1+z)$ from $z = 7$ to $z = 0$. Throughout the paper, we use a ΛCDM cosmology with $\Omega_\mathrm{M} = 0.26$, $\Omega_{\mathrm{B},0} = 0.044$, $h = 0.71$, $n = 0.96$, and $\sigma_8 = 0.79$, which are consistent with the WMAP5 data (Dunkley et al. 2009; Komatsu et al. 2009).

2.2 Radiative cooling

Once the halo formation history is fixed, we model the radiative cooling of halo gas. As shown in Lu et al. (2010), the predictions of often-used cooling models do not agree. Since these models do not incorporate uncertainties in their cooling prescriptions, the model choice imposes a strong prior on the SAM. In order to compare with published results, our model follows Croton et al. (2006). An analysis of varying the cooling prescription will be presented in a future paper. In the Croton model, the halo hot gas is redistributed at every time-step, and the density profile of the hot gas is assumed to be a singular isothermal profile,

$$\rho_\mathrm{gas} = \frac{m_\mathrm{gas0}}{4\pi r_\mathrm{vir}} r^{-2},$$

where r_vir is the virial radius of the halo. The total mass of hot halo gas mass is $m_\mathrm{gas0} = f_\mathrm{b} m_\mathrm{vir} - \sum_i [m_*^i + m_\mathrm{cold}^i + m_\mathrm{out}^i]$, where $f_\mathrm{b} = \Omega_\mathrm{b}/\Omega_0$ is the universal baryon fraction, m_*, m_cold and m_out are the masses in stars, cold gas and ejected gas, respectively, and the summation is over all galaxies in the halo. The temperature of the hot gas is constant for each halo with $T_\mathrm{gas} = T_\mathrm{vir} = 35.9(\frac{v_\mathrm{vir}}{\mathrm{kms}^{-1}})^2\mathrm{K}$ where v_vir is the circular velocity of the halo at the virial radius. The cooling timescale of the gas at radius r is then estimated by

$$\tau_\mathrm{cool}(r) = \frac{3}{2} \frac{\mu m_\mathrm{H} k T_\mathrm{gas}}{\rho_\mathrm{gas}(r) \Lambda(T_\mathrm{gas}, Z_\mathrm{gas})}, \tag{1}$$

where μ is the mean molecular weight in units of the mass of hydrogen atom, and Λ is the cooling function from Sutherland & Dopita (1993). At each time-step, we calculate the cooling radius r_cool by equating the cooling timescale with the dynamical timescale, $\tau_\mathrm{cool} = \tau_\mathrm{dyn} \equiv r_\mathrm{vir}/v_\mathrm{vir}$. If the cooling radius is equal to or smaller than the virial radius, the cooling rate is defined as

$$\dot{m}_\mathrm{cool} = 0.5 m_\mathrm{hot} \frac{r_\mathrm{cool} v_\mathrm{vir}}{r_\mathrm{vir}^2}. \tag{2}$$

In words, half of the hot gas mass enclosed by the cooling radius cools and accretes onto the central object of the halo in a dynamical timescale. If the cooling radius is larger than the virial radius, we set the cooling rate equal to the total hot gas mass in the halo divided by the dynamical timescale. We implicitly assume that all hot gas is associated with the primary halo and only central galaxy can accrete cooling gas; that is, satellite subhalos contain no hot gas.

In some recent SAMs, Active Galactic Nuclei (AGN) feedback reduces the gas cooling in massive halos (e.g. Croton et al. 2006; Bower et al. 2006; Somerville et al. 2008). Equivalently, AGN feed back stops radiative cooling for halos with masses larger than a characteristic mass ($\sim 10^{12}$ M$_\odot$) (Cattaneo et al. 2006). To include this effect, we introduce a characteristic halo mass for radiative cooling, $M_{\rm CC}$, above which radiative cooling of the hot halo gas is assumed to be negligible. Since the exact value of $M_{\rm CC}$ is not known a priori, we treat it as a free parameter in a relatively large mass range, $10^{11.5}$ - $10^{14.5}$ h^{-1}M$_\odot$.

2.3 Star formation

We assume that the cooled-fraction of halo gas settles into the galaxy in an exponential disk with scale length $r_{\rm disc}$. This gas form stars when the gas disk has a surface density higher than a certain threshold, $\Sigma_{\rm SF}$, mimicking the critical surface gas density for star formation seen in disk galaxies (e.g. Kennicutt 1998; Kennicutt et al. 2007; Bigiel et al. 2008). The fraction of cold gas above the threshold is given by the ratio of the radius $r_{\rm crit}$ at which the cold gas density is $\Sigma_{\rm SF}$ to the disk scale length:

$$r_{\rm crit}/r_{\rm disc} = \ln \frac{m_{\rm cold}}{2\pi r_{\rm disc}^2 \Sigma_{\rm SF}}, \tag{3}$$

where $m_{\rm cold}$ is the total cold gas mass of the galaxy. Therefore, the cold gas mass enclosed by $r_{\rm crit}$ is determined by the ratio $r_{\rm disc}^2\Sigma_{\rm SF}/m_{\rm cold}$. Observationally, the threshold surface density is $\sim$ 10 M$_\odot$pc^{-2} (e.g. Martin & Kennicutt 2001), although the scale length may vary. Theoretically, the disk radius (the scale-length) is related to the virial radius and the spin parameter of its host halo: $r_{\rm disc} \approx \frac{\lambda}{\sqrt{2}} r_{\rm vir}$ (e.g. Mo et al. 1998). In cosmological N-body simulations, the spin parameters λ for dark matter halos follow a log-normal distribution with a median ~ 0.05 (e.g. Warren et al. 1992; Cole & Lacey 1996), but the distribution of λ for the baryonic component that forms galaxy disks is poorly understood (e.g. Bett et al. 2010; Navarro & Benz 1991). In our SAM, we adopt the fiducial value $\lambda_0 = 0.05$. This yields $r_{\rm disc,0} = 0.035 r_{\rm vir}$ and $\Sigma_{\rm SF,0} = 1$ M$_\odot$pc^{-2}. We then parameterize the term $r_{\rm disc}^2\Sigma_{\rm SF} = f_{\rm SF} r_{\rm disc,0}^2 \Sigma_{\rm SF,0}$. In the Croton et al. (2006) model, $r_{\rm disc}$ is set to be $3r_{\rm disc,0}$, and $\Sigma_{\rm SF} = 10\,{\rm M_\odot pc^{-2}}$, so that $f_{\rm SF} = 90$.

Using on our parameterization, the cold gas mass on the disk available for star formation is

$$m_{\rm sf} = m_{\rm cold}\left[1 - \left(1 + \ln\frac{m_{\rm cold}}{2\pi f_{\rm SF}\Sigma_{\rm SF,0} r_{\rm disc,0}^2}\right)\frac{2\pi f_{\rm SF}\Sigma_{\rm SF,0} r_{\rm disc,0}^2}{m_{\rm cold}}\right]. \tag{4}$$

We take the star formation rate proportional to the cold gas mass within $r_{\rm crit}$ and inversely proportional to the dynamical timescale of the disk, $\tau_{\rm disc} = \frac{r_{\rm disc}}{v_{\rm vir}}$, yielding

$$\dot{m}_* = \epsilon_* \frac{m_{\rm sf}}{\tau_{\rm disc}}, \tag{5}$$

where ϵ_* is the star formation efficiency. We assume that ϵ_* has a broken power-law dependence on the circular velocity of the host halo:

$$\epsilon_* = \begin{cases} \alpha_{\rm SF} & v_{\rm vir} \geq V_{\rm SF}; \\ \alpha_{\rm SF}\left(\frac{v_{\rm vir}}{V_{\rm SF}}\right)^{\beta_{\rm SF}} & v_{\rm vir} < V_{\rm SF}, \end{cases} \tag{6}$$

where $\alpha_{\rm SF}$ and $\beta_{\rm SF}$ are parameters. Early models adopted a pure power-law until $\epsilon_* \sim 1$ (e.g. Kang et al. 2005). The Croton et al. (2006) model assumes $\beta_{\rm SF} = 0$ and sets $\alpha_{\rm SF}$ so that 5–15% of the cold gas is converted into stars in a disk dynamical time. The GALFORM of Cole et al. (2000) considers cases with $\beta_{\rm SF} = 0$, 1.5 and 2.5. In our model, all the four parameters, $\alpha_{\rm SF}$, $\beta_{\rm SF}$, $V_{\rm SF}$ and $f_{\rm SF}$, are considered as free parameters when modeling star formation in quiescent disks.

2.4 Supernova feedback

We assume that supernova (SN) feedback affects the interstellar medium (ISM) and hot halo gas in three ways: (i) the energy feedback from SN reheats a fraction of the disk ISM from the cold phase to the hot phase, and the reheated gas is mixed with the hot halo gas; (ii) a fraction or all of the heated gas is directly ejected from the host halo without mixing with the hot halo gas; and (iii) if the SN energy from all galaxies in a halo is sufficiently large, the hot gas in the host halo can be heated, causing a fraction of the halo hot gas to be ejected from the halo. No SAM has incorporated all of these mechanisms and the strength of the feedback is usually chosen without strong prior justification. For example, the Croton model considered both mechanisms (i) and (iii) (Croton et al. 2006), while GALFORM incorporated (i) and (ii) (Benson et al. 2003). In these models, the total amount of SN feedback energy is assumed to be related to the star formation rate, and the feedback is assumed to be instantaneous. The feedback strength is controlled by a fixed number (e.g. Croton et al. 2006) or assumed to have a power-law dependence on the circular velocity of the host halo (e.g. Somerville & Kolatt 1999; Cole et al. 2000; Kang et al. 2005). Our model incorporates all three mechanisms, and their relative strengths are free parameters. We assume that for every solar mass of stars formed, the energy released by supernovae is $\eta_{\rm sn}E_{\rm sn}$, where $\eta_{\rm sn}$ is determined by the stellar initial mass function (IMF) and $E_{\rm sn}=10^{51}$erg. Our feedback model enforces energy conservation, so that the total energy to heat the gas cannot exceed the total energy released from supernovae.

We write the SN energy released by a mass of Δm_* of star formation as

$$E_{\rm fb}=\alpha_{\rm SN}\frac{1}{2}\Delta m_* V_{\rm SN}^2 \tag{7}$$

where $V_{\rm SN}=630$km/s and the free parameter $\alpha_{\rm SN}$ describes the uncertainties in the feedback energy and in the IMF. For a Scalo IMF ($\eta_{\rm sn}=5\times10^{-3}$) and with 20% of the SN energy in feedback (e.g. Kang et al. 2005), we find $\alpha_{\rm SN}=0.25$. We allow $\alpha_{\rm SN}$ to vary from 0.001 to 10, encompassing the uncertainty of this parameterization. To conserve energy, the total SN energy released by m_* of star formation and available for feedback, $E_{\rm fb}$, should be equal to the sum of the energies used for the reheating, ejection and powering the wind. Thus, we can write

$$E_{\rm fb}=\frac{1}{2}\left(1-f_{\rm ej}\right)f_{\rm rh}\Delta m_* v_{\rm vir}^2+\frac{1}{2}f_{\rm ej}f_{\rm rh}\Delta m_* v_{\rm esc}^2+\frac{1}{2}\Delta m_{\rm wind}v_{\rm esc}^2, \tag{8}$$

where the coefficients, $f_{\rm rh}$ and $f_{\rm ej}$, control the mass loading for the reheating and ejection, $v_{\rm esc}$ is the circular velocity of the current host halo characterizing its binding energy, and $v_{\rm vir}$ is the circular velocity of the host halo at the latest time when it was still a primary halo, characterizing the binding energy of the galaxy. Note that $v_{\rm esc}\neq v_{\rm vir}$ only for a satellite galaxies. We further assume that the fraction for reheating, $f_{\rm rh}$, has a power-law dependence on the circular velocity of the halo, $v_{\rm vir}$. If the galaxy is a satellite, we use the circular velocity of its host halo when it first became a subhalo. So we have

$$f_{\rm rh}=\alpha_{\rm RH}\left(\frac{V_0}{v_{\rm vir}}\right)^{\beta_{\rm RH}}, \tag{9}$$

where V_0 is an arbitrary factor and is set to be 220 km/s. The power-law has an upper limit given by energy conservation:

$$f_{\rm rh,max}=\alpha_{\rm SN}\left(\frac{V_{\rm SN}}{v_{\rm vir}}\right)^2. \tag{10}$$

When an amount of $f_{\rm rh}\Delta m_*$ cold gas is reheated, we assume a fraction $f_{\rm ej}$ escapes from the halo. For simplicity, we assume $f_{\rm ej}$ has a power-law dependence of the circular velocity of the current host halo:

$$f_{\rm ej} = \alpha_{\rm EJ} \left(\frac{V_0}{v_{\rm esc}} \right)^{\beta_{\rm EJ}} . \tag{11}$$

Again energy conservation sets an upper limit on $f_{\rm ej}$:

$$f_{\rm ej,max} = \left[\frac{f_{\rm rh,max}}{f_{\rm rh}} - 1 \right] \times \left[\left(\frac{v_{\rm esc}}{v_{\rm vir}} \right)^2 - 1 \right]^{-1} . \tag{12}$$

If there is still energy available after reheating and ejection, the surplus is assumed to power a wind, and the mass of the wind can be written as

$$\Delta m_{\rm wind} = \epsilon_{\rm W} \Delta m_* \left\{ \alpha_{\rm SN} \left(\frac{V_{\rm SN}}{v_{\rm esc}} \right)^2 - f_{\rm rh} \left[\left(\frac{v_{\rm vir}}{v_{\rm esc}} \right)^2 + f_{\rm ej} \right] \right\} . \tag{13}$$

We assume that a fraction of $f_{\rm RI}$ of the gas in the outflow, ejection and wind will come back to the halo as hot gas in a dynamical timescale, and we treat $f_{\rm RI}$ as a free parameter.

Thus, we model the SN feedback with 7 parameters: $\alpha_{\rm SN}$, $\alpha_{\rm RH}$, $\beta_{\rm RH}$, $\alpha_{\rm EJ}$, $\beta_{\rm EJ}$, $\epsilon_{\rm W}$ and $f_{\rm RI}$. Because the wind dominates the outflow, we find that $\alpha_{\rm EJ}$ and $\beta_{\rm EJ}$ are not constrained by the stellar mass function alone (see Section 4). Therefore we fix $\alpha_{\rm EJ} = 0$ and $\beta_{\rm EJ} = 0$ in the present paper.

2.5 Galaxy mergers

When two dark matter halos merge, we simply add the dark matter and hot gas of the smaller halo to the bigger one. The central galaxy of the more massive halo is then treated as the central galaxy of the new halo, and all other galaxies are considered as satellites. A satellite galaxy merges into the central galaxy in some fraction $f_{\rm DF}$ of the dynamical friction timescale. The dynamical friction timescale is parameterized as

$$t_{\rm fric} = \frac{1.17 r_{\rm vir}^2 v_{\rm vir}}{\ln \Lambda G M_{\rm sat}}, \tag{14}$$

where $r_{\rm vir}$ and $v_{\rm vir}$ are the virial radius and circular velocity of the new host halo, $M_{\rm sat}$ is the mass of the previous host halo of the satellite before it merges into the current halo, and $\ln \Lambda$ is the Coulomb logarithm, which is modeled as $\ln \Lambda = \ln(1 + M_{\rm vir}/M_{\rm sat})$ (e.g. Croton et al. 2006). This formula assumes that the satellite galaxy is hosted by a subhalo with mass $M_{\rm sat}$ and orbits in a central halo with a singular isothermal density profile of circular velocity $v_{\rm vir}$ from the virial radius (Binney & Tremaine 1987). Earlier SAMs adopted similar parameterizations, but used different prefactors. For example, some SAMs chose the galaxy mass for $M_{\rm sat}$ (e.g. Cole et al. 2000) and some others chose the subhalo mass for $M_{\rm sat}$

(e.g. Croton et al. 2006); this results in an order of magnitude difference in the prefactor. Other uncertainties include the value of the Coulomb logarithm, the effect of tidal stripping on orbital decay, and the initial velocity of the satellite. In our model, these uncertainties are absorbed into the prefactor, $f_{\rm DF}$, a free parameter.

The merging timescale is calculated when the host halo of the satellite merges into the host halo of the central galaxy. If the satellite was already a satellite before the merger, the dynamical fraction timescale for the satellite is recalculated based on the properties of the new host. When a satellite galaxy merges into the central galaxy, our treatment for the merger remnant depends on the mass ratio of the two galaxies, $m_{\rm sat}/m_{\rm central}$. Mergers are considered as a major or a minor merger depending on whether $m_{\rm sat}/m_{\rm central}$ is larger or smaller than a pre-selected $f_{\rm MG} < 1$. The values of $f_{\rm MG}$ adopted in earlier SAMs are ~ 0.3. As the choice of this parameter is not constrained by the stellar mass function of galaxies, we simply take $f_{\rm MG} = 0.3$ instead of treating it as a free parameter.

For a minor merger ($m_{\rm sat}/m_{\rm central} \leq 0.3$), the satellite's stars are added to the central bulge, and the satellite's gas is added to the central disk. A minor merger is assumed to trigger a star-burst in the disk, and all the stars formed in the burst is added to the disk component. For a major merger ($m_{\rm sat}/m_{\rm central} > 0.3$), we combine all the existing stars from the two merging galaxies into a central galaxy, which is now assumed to be an elliptical. Each major merger triggers a star-burst, and all stars formed in the burst are added into the central elliptical galaxy. A fraction $e_{\rm burst}$ of the combined cold gas in the two merging progenitors becomes stars, and the rest joins the gaseous disk. We assume that $e_{\rm burst}$ depends on the ratio of the baryon masses of the two galaxies: $e_{\rm burst} = \alpha_{\rm burst}(m_{\rm sat}/m_{\rm central})^{\beta_{\rm burst}}$.

Similar models for galaxy mergers were adopted by Somerville et al. (2001, 2008) and Croton et al. (2006) although different authors used different values for the model parameters. In our model, the four parameters in the parameterization, $f_{\rm DF}$, $f_{\rm MG}$, $\alpha_{\rm burst}$ and $\beta_{\rm burst}$, are all treated as free parameters. As mentioned above, since $f_{\rm MG}$ is not constrained by the stellar mass function considered in this paper, we simply fix its value to be 0.3.

2.6 Calculation of a single model

The flowchart shown in Figure 1 summarizes the calculation of the SAM described above. The code loads merger trees and other tables (e.g. cooling functions, stellar mass-to-light ratios for different star formation histories, and dust extinction) for subsequent calculations, and then reads the model parameters introduced above in this section (and summarized in Table 1). Each of the merger trees is walked from the top (initial time) to the bottom (present time). At each tree level, a galaxy grows in the center of a halo if the halo does not have any progenitor halos. If the halo is assembled through the mergers of progenitor halos, the central galaxy of the most massive progenitor is considered to be the central galaxy of the current halo, and all other existing galaxies are considered to be satellites.

At the initial time, we distribute hot gas in dark matter halos and radiatively cooling begins. We sub-divide each of the 60 time steps used to sample a merger tree into 5 finer time steps (equally speced in t) to compute the cooling and to evolve galaxies. In every time step, gas that is able to cool in the current time step is assigned to the central galaxy. For all galaxies in the halo, star formation continues until the cold gas is exhausted as the cold gas surface density goes below the threshold value. When a satellite galaxy merges into a central galaxy, the recipes for the morphological transformation and merger-triggered starburst are applied. For any star formation, quiescent- or burst-mode, the code calculates the effect of SN feedback. Chemical evolution of the ISM is modeled by the "instantaneous recycling approximation" (Cole et al. 2000): a fraction R of the newly formed stellar mass and a yield p of heavy elements are instantaneously returned to and uniformly mixed with the cold gas. Metals enrich the halo gas as the reheated gas mixes with the hot halo gas (assuming a one-zone model, see Subsection 2.4) and affect the cooling rate. Both R and p depend on the IMF. Since we have adopted a simplified model for gas cooling (see Subsection 2.2) and since the stellar mass function we are concerned here is affected by metallicity only through gas cooling, in this paper we simply fix $R = 0.3$ and $p = 0.03$ instead of treating them as free parameters. Our code uses the Stellar Population Synthesis (SPS) model of Bruzual A. & Charlot (1993) and a dust model of Kauffmann et al. (1999) to assign fluxes to galaxies.

The evolution continues until the root of the merger tree is reached. At this point, we have a realization of the model specified by the set of parameters. The results obtained from these realizations can then be used to compare with observational data to evaluate the likelihood of the data given the model.

Table 1. Model parameters

#	Parameter	Meaning	Prior	Posterior
1	$\log M_{\rm CC}(\,{\rm M}_\odot)$	cooling cut-off halo mass	[1.5, 4.5] [1.5 , 4.5] [1.5, 4.5]	[2.19, 2.67] [3.09, 4.47] [2.13, 2.49] [3.15, 4.47] [2.07, 2.49]
2	$\log \alpha_{\rm SF}$	star formation efficiency power-law amplitude	[-3, 0] [-3, 0] [-3, 0]	[-2.19, -0.03] [-2.97, -2.85] [-1.47, -0.03] [-2.97, -2.49]
3	$\beta_{\rm SF}$	star formation efficiency power-law index	[-1, 12] [-0.2, 0.2] [-0.2, 0.2]	[-0.87, 0.43] [3.29, 11.87] [-0.2, 0.2] [-0.2, 0.2]
4	$\log V_{\rm SF}$ (km/s)	star formation law turn-over halo circular velocity	[1.5, 3.0] [2.1, 2.3] [2.1, 2.3]	[1.52, 2.39] [2.1, 2.3] [2.1, 2.3]
5	$\log f_{\rm SF}(\,{\rm M}_\odot/{\rm pc}^2)$	star formation threshold gas surface density	[-1, 3] [1.8, 2.2] [1.8, 2.2]	[-0.96, -0.64] [-0.24, 2.16] [1.8, 2.2] [1.8, 2.2]
6	$\log \alpha_{\rm SN}$	SN feedback energy fraction	[-3, 1] [-3, 1] [-3, 1]	[-2.35, 0.85] [-1.35, -1.15] [-0.25, 1.05] [-1.75, 0.25]
7	$\log \alpha_{\rm RH}$	SN feedback reheating power-law amplitude	[-3, 2] [-3, 2] [-3, 2]	[-2.55, 1.95] [-2.65, -0.75] [0.260, 1.22] [-0.75, 1.95]
8	$\beta_{\rm RH}$	SN feedback reheating power-law index	[0, 14] [0, 14] [1.8, 2.2]	[0.14, 13.86] [5.46, 11.62] [1.8, 2.2]
9	$\log \epsilon_{\rm W}$	fraction of surplus SN feedback energy used for powering wind	[-3, 0] [-3, 0] [-3, 0]	[-2.97, -0.15] [-2.97, -0.81] [-2.97, -0.21]
10	$\log f_{\rm RI}$	fraction of re-infall ejected hot gas	[-2, 0] [-2, 0] [-2, 0]	[-1.98, -0.02] [-1.97, -0.03] [-1.94, -0.02]
11	$\log f_{\rm DF}$	merging time-scale in dynamical friction time-scale	[-1, 2] [-1, 2] [-1, 2]	[0.53, 1.97] [0.23, 0.59] [0.77, 1.97] [0.05, 0.65]
12	$\log \alpha_{\rm SB}$	merger triggered star burst efficiency power-law amplitude	[-2, 0] [-2, 0] [-2, 0]	[-1.98, -0.02] [-1.97, -0.09] [-1.97, -0.15]
13	$\beta_{\rm SB}$	merger triggered star burst efficiency power-law index	[0, 2] [0, 2] [0, 2]	[0.02, 1.98] [0.02, 1.98] [0.02, 1.98]
14	$\alpha_{\rm EJ}$ (fixed)	SN feedback cold gas ejection power-law amplitude	0.0	0.0
15	$\beta_{\rm EJ}$ (fixed)	SN feedback cold gas ejection power-law index	0.0	0.0
16	$f_{\rm MG}$ (fixed)	major merger minor merger threshold	0.3	0.3

3 BAYESIAN MODEL INFERENCE AND MCMC METHOD

3.1 Bayesian inference

Bayes Theorem states that the posterior probability of a set of parameters $\mathbf{\Theta}$ in a model (or hypothesis) H for given data $\mathbf{D}$ is

$$P(\mathbf{\Theta}|\mathbf{D}, H) \propto P(\mathbf{\Theta}|H)L(\mathbf{D}|\mathbf{\Theta}, H), \tag{15}$$

where $P(\mathbf{\Theta}|H)$ is the prior probability distribution which describes the knowledge about the parameters acquired before seeing the data, and $L(\mathbf{D}|\mathbf{\Theta}, H)$ is the likelihood of the data $\mathbf{D}$ for the given model parameter set $\mathbf{\Theta}$. As mentioned earlier, for the problem we are tackling here, the prior knowledge about the model parameters is limited. We therefore choose either uniform or exponential distributions between two physically chosen bounds, depending on the parameter in question. As a test, the priors for some of the parameters are made strongly restrictive to demonstrate the sensitivity to these choices. The prior information is summarized in Table 1. The problem-specific definition of the likelihood function is described in later sections.

3.2 The Markov-Chain Monte-Carlo algorithm

It is not possible to integrate the posterior probability distribution function analytically for our SAM. We resort to use a Monte-Carlo sampling approach to summarize the posterior distribution. We adopt a newly developed software package, the Bayesian Inference Engine (BIE, Weinberg 2010a,b), which includes a suite of advanced MCMC algorithms and supports parallel computation. As we will show later, the topological structure of the posterior probability distribution in our problem is high-dimensional and very complex. Not only does the posterior show multi-mode and strong degeneracy among model parameters, but also the high-probability regions only occupy a very small fraction of the entire parameter space.

Because of these features, it is technically challenging to sample the posterior efficiently with a standard Metropolis-Hastings MCMC algorithm. To expedite sampling and mixing, we employ a multiple-chain differential evolution algorithm with tempering (Ter Braak 2006; Weinberg 2010b). At each step, the algorithm randomly selects two other chains and uses a fraction of the vector connecting the current states of the two chains as a proposal. This strategy improves proposal efficiency and mixing by automatically "tuning" the proposals to the ensemble of states comprising the individual chains. For a multi-dimensional Gaussian posterior, the optimized fraction of the vector is $\gamma_0 = 2.38/\sqrt{N_{\rm dim}}$, where $N_{\rm dim}$ is the dimension of the parameter space (Ter Braak 2006). Since our posterior is expected to deviate significantly from a Gaussian, we use $\gamma = 0.1\gamma_0$ to maintain a good acceptance rate ($\approx 10\%$). For every 10 steps, we use the full vector as the proposal by temporally setting $\gamma = 1$ to allow chains to swap modes. As the simulation proceeds, the chains gradually settle into the high probability regions and the distribution can guide the chains to move along the ridges of the posterior or to jump between different modes. Moreover, all converged chains sample the posterior, further enhancing the overall efficiency.

The standard differential evolution algorithm depends on the initial distribution of chains for state-space exploration. Because the posterior distribution for our model is compact and complex, tempered simulation algorithm (Neal 1996) is used to enhance the exploration of state space. In short, tempered simulation proposes exchanges between the posterior distribution of P_0 and a "powered-up" distribution $P_j \propto P_0^{1/T_i}$ with $T_i \leq T_{\rm max}$. Each step begins by "melting", $T_{i+1} > T_i$ followed by "freezing", $T_{i+1} < T_i$. We perform a one tempered step for every 21 standard steps, with the maximum temperature $T_{\rm max}$ selected similar to the difference in the logarithmic posterior probability between a high-probability region and a low-probability valley: $T_{\rm max} \approx \ln P_{max}/P_{min}$. In the temperature range from 0 to $T_{\rm max}$, we set M temperature levels equally spaced in logarithmic scale. The default value of M is set to be $\sqrt{N_{\rm dim}+3}\ln T_{\rm max}$. For our problem, $N_{\rm dim} = 13$ and we set $T_{\rm max} = 64$, so that $M = 16$. At each temperature level T_i, 10 differential evolution steps are taken, with γ stretched by a factor of $T_i^{1/2}$. In total, it takes 320 differential evolution steps for a chain to go through the "melting" and "freezing" procedure for a single tempered step. Therefore, employing the tempered simulation steps significantly increases the computational load. However, the tempered simulation steps substantially improve the exploration of state space and dramatically speed up convergence.

3.3 A Bayesian-inference based SAM

The structure of our Bayesian-inference based SAM is outlined by Figure 2. The MCMC algorithm provides proposal parameter vectors for the SAM, and the SAM predicts the galaxy population using the proposed parameter set. The likelihood is evaluated by comparing the model prediction with data, and is returned back to the MCMC. The MCMC algorithm accepts or rejects the proposal based on the posterior probability, and generates a new proposal for the SAM. The MCMC-SAM loop continues until convergence is achieved. The convergence of the chains is monitored by the Gelman-Rubin $\hat{R}$ statistic (Gelman & Rubin 1992). In essence, $\hat{R}$ is the ratio of the variance between chains to the variance within chains. We declare convergence when $\hat{R} \leq 1.2$. When the chains are converged, we use post-convergence states (typically about 10^6) to study and characterize the posterior distribution. The converged states sample the full probability distribution of the model parameters given observational data, and can be used to estimate confidence regions for individual model-data comparison through marginalization and determine the relative posterior odds for different models. In the following sections, we use a simple example to demonstrate the power of our Bayesian-inference based SAM.

4 POSTERIOR OF THE SAM FOR STELLAR MASS FUNCTION

In this paper, we consider constraints on our SAM implied by the stellar mass function of galaxies, a fundamental property of the galaxy population that has been extensively used for model–data comparison. We choose the stellar mass function instead of the luminosity function simply because the stellar mass of galaxies is a direct prediction of our SAM, without the uncertainties in the stellar population synthesis model. However, these same uncertainties are present in the reduced data, since a stellar population synthesis model was used to convert the observed galaxy luminosities into stellar masses. These uncertainties should in principle be properly included in the error budget of the observational data, and we will examine the impact of incorrect error models on the Bayesian inference.

We study the constraints on the 13 free parameters characterizing our SAM (see Table 1) using the stellar mass function of Bell et al. (2003). Assuming that stellar mass bins are mutually independent, the likelihood function is

$$L(\Phi_{\rm obs}|\theta) = L_0 \exp\left\{-\sum_i \frac{[\Phi_{i,{\rm obs}} - \Phi_{i,{\rm mod}}(\theta)]^2}{2\sigma^2_{i,{\rm obs}}}\right\}, \tag{16}$$

where L_0 is an arbitrary normalization factor, $\Phi_{i,{\rm obs}}$ is the value of the observed stellar mass function in the ith bin, $\Phi_{i,{\rm mod}}$ is the corresponding value predicted by the model with the given parameter set θ, and $\sigma^2_{i,{\rm obs}}$ is the variance of the observed stellar mass function. The error estimation of Bell et al. (2003) only takes into account the sampling error, but we expect significant bias (systematic uncertainty) from assumptions made in the data reduction. To mimic the effect of large systematic uncertainty, we artificially inflate the statistical error bars by a factor of 3. Please note, we are not advocating this procedure, rather, we argue this is a *very bad* thing to do in general for at least two reasons: (i) this tends to imply greater support for a model than is truly admitted by the data, and conversely, tends to reduce the ability of the data to choose between competing hypotheses; and (ii) inflated error may hide serious problem with the data or inconsistencies with other data. Strictly speaking, the Bayesian approach applies equally well to systematic uncertainty as to sampling error. Mathematically, let systematic uncertainties be described by a parameter vector η. The likelihood now depends on η through $\Phi_{i,\mathrm{obs}}(\eta)$. We simply define a prior distribution for the uncertainty $P(\eta)$ by expert opinion or through an ancillary calibration. The inference continues as before, now with the augmented parameter vector $\mathbf{\Theta} = (\theta, \eta)$. In the end, we simply marginalize over η. For our problem specifically, we are aware our error inflation produce is ad hoc and does not correctly represent the bin-to-bin covariance in $\Phi_{i,\mathrm{obs}}$ induced by the stellar mass function. We will discuss how such covariance affects our results in Section 6. We will perform a luminosity function-based inference using a population synthesis model and an appropriately chosen prior uncertainty in a future paper. However, the lack of a stellar mass function with a suitably described error model forces us to made a crude error model approximation for the point of illustration in the next section. In addition, our Monte-Carlo evaluation of $\Phi_{i,\mathrm{mod}}$ has variance. However, it is typically 3 times smaller than that in the data and, therefore, not explicitly included in equation (16).

4.1 Physical implications

Figure 3 shows the one- and two-dimensional marginalized posterior probability distributions of the 13 free parameters. Three of these parameters, f_{RI}, α_{SB} and β_{SB}, are unconstrained by the stellar mass function and not shown. In the upper-right corner of the figure, we plot the predicted stellar mass function by marginalizing over the 95% confidence range of the posterior. Clearly, the stellar mass function is well reproduced by the model. Table 1 lists the 95% confidence bounds of all parameters.

The strength of the constraint varies widely. Some parameters are weakly constrained: for example, ϵ_{W}, the efficiency of SN feedback powering galactic wind, is very weakly constrained. In contrast, some parameters are tightly constrained: for example, V_{SF} is constrained to a narrow range (around $\sim$ 160km/s), so are β_{SF} (around 6) and β_{RH} (around 8). Our inferred values of β_{SF} and β_{RH} are much higher than those adopted in previous analyses. The posterior indicates a sharply suppressed star formation efficiency in halos with circular velocities below $\sim$ 160km/s. In addition, the posterior distribution in the $\beta_{\rm SF}$ - $\beta_{\rm RH}$ plane reveals bimodality: either the star formation efficiency or the SN reheating efficiency is a steep power-law of halo circular velocity. In other words, the shallow slope of the stellar mass function at the low-mass end requires the suppression of star formation in small halos. Since star formation efficiency directly controls the conversion of cold gas to stellar mass, high $\beta_{\rm SF}$ mode dominates high $\beta_{\rm RH}$ mode. We are unsure whether or not such high values of $\beta_{\rm SF}$ and $\beta_{\rm RH}$ are physical plausible. It is likely that some new physics in addition to SN feedback is required to suppress star formation in low-mass halos, as we will demonstrate in detail in a forthcoming paper.

Some model parameters are strongly correlated. These include the following pairs of parameters: $f_{\rm SF}$–$\alpha_{\rm SF}$; $\alpha_{\rm RH}$–$\beta_{\rm RH}$; and $M_{\rm CC}$–$f_{\rm DF}$. Both $\alpha_{\rm SF}$ and $f_{\rm SF}$ control the conversion of cold gas to stars and the degeneracy is expected. Similarly, the two parameters in the power-law parameterization of the SN reheating, $\alpha_{\rm RH}$ and $\beta_{\rm RH}$, are degenerate. And again, the parameters controlling the two mechanisms responsible for the formation of central galaxies in massive halos, $M_{\rm CC}$ and $f_{\rm DF}$ are correlated; massive central galaxies can either acquire their mass through gas cooling and *in situ* star formation, or through accretion of satellite galaxies. The observed sharp decline of the stellar mass function at the high-mass end requires either the AGN feedback be strong so that gas cooling and star formation in halos more massive than $\sim 10^{12}\,{\rm M}_{\odot}$ is effectively quenched, or the merger rate of satellite galaxies into a central galaxy by dynamical friction be slow.

4.2 Structure of the posterior distribution

The two-dimensional posterior distributions shown in Figure 3 are marginalized over 11 dimensions and wash out much of the intrinsic sub-dimensional structure that complicates the inference and renders unreliable tweaking by hand. To demonstrate this, Figure 4 shows some two-dimensional cuts through the posterior distribution. These cuts are made in the $\alpha_{\rm SF}$–$\alpha_{\rm RH}$ plane at four different values of $f_{\rm SF}$ after the posterior is first marginalized over the other 10 dimensions. The posterior distribution changes dramatically with $f_{\rm SF}$, from horizontally oriented at $\log(f_{\rm SF}) \sim 2.2$, to vertically oriented at $\log(f_{\rm SF}) \sim 0.5$. These suggest that the posterior is significantly more complex than the marginalized distributions shown in Figure 3. The fine structure and complex topology of the posterior also make it clear that it is extremely difficult to find the bast fit by tuning model parameters by hand.

It also explains why it is computationally challenging to properly sample the posterior. For example these results required approximately 5×10^4 2GHz Opteron CPU hours.

5 THE IMPACT OF PRIOR CHOICE

In this section, we study the affect of the prior distribution on the final inference by selectively applying narrow prior distributions in some dimensions. This mimics the practice of fixing some model parameters. In Case 1, three of the 13 parameters are given narrow priors. The value of $\beta_{\rm SF}$ is limited to the narrow range $[-0.2, 0.2]$, to mimic a flat power-law for star formation efficiency adopted in some earlier SAMs (e.g. Croton et al. 2006). The parameter $V_{\rm SF}$ has little effect so we set $\log(V_{\rm SF}/{\rm km\,s^{-1}})$ in the narrow range $[2.1, 2.3]$. Furthermore, we assign a narrow prior, $[1.9, 2.1]$, to $\log f_{\rm SF}$, corresponding to the choice $\Sigma_{\rm sf} \approx 10\,{\rm M_\odot/pc^2}$ and $r_{\rm disc} \approx 3r_{\rm disc0}$ that is often used in previous models (e.g. Croton et al. 2006). All other prior distributions are the same as in the fiducial case considered §4 (Case 0). The resulting marginalized posterior distribution in Figure 5 shows that the distribution in all parameters becomes more compact. The improvement of prior information on some parameters not only tightens the constraints on those parameters themselves but can also help break degeneracy in other dimensions. For example, since the star formation law is restricted to have weak dependence on halo mass, the efficiency of SN reheating is forced to be a steep power-law of halo circular velocity. For the same reason, degeneracies of other parameters with $\beta_{\rm RH}$ are all reduced. Note that the restrictive prior is not located near the dominant posterior mode in Case 0 (cf. Fig. 3). Moreover, the bulk of the Case 1 posterior has very low probability in the Case 0 posterior. Nevertheless, the quality of the fit does not change much, as one can see from the reproduced stellar mass functions shown in the upper-right panel of the plot. This illustrates the danger in fixing the values of parameters to plausible values especially when there is no prior reason for a strong constraint.

Case 1 requires that the SN feedback is a very steep function of the halo circular velocity when the star formation efficiency is forced to change slowly with halo mass. Early SAMs (e.g. Kauffmann et al. 1999; Kang et al. 2005) assumed a weak dependence of the SN feedback on halo mass ($\beta_{\rm RH} \sim 2$) and concluded that the number of faint galaxies are over-predicted if $\beta_{\rm SF} \sim 0$. However, whether or not a good fit can still be obtained by changing other parameters while keeping $\beta_{\rm SF} \sim 0$ and $\beta_{\rm RH} \sim 2$ requires a full exploration of the high-dimensional parameters space. Case 2 addresses this question by imposing the additional prior restriction, $\beta_{\rm RH} \in [1.8, 2.2]$, and in Figure 6 we show the resulting marginalized distributions. The modes have moved substantially with respect to those in Case 1. To compensate for the weakened SN reheating in small halos due to the assumed weak dependence of SN reheating on halo circular velocity, the model employs a much lower efficiency for star formation and a larger reheating amplitude; the mode moves from the lower-right to the upper-left in the $\log\alpha_{\rm SF} - \log\alpha_{\rm RH}$ plane. For similar reasons, the posterior mode in other dimensions also change.

The posterior-weighted stellar mass function is shown in the upper-right panel of Figure 6. Even though the power indices $\beta_{\rm SF}$ and $\beta_{\rm RH}$ are fixed to low values, the model can still reproduce the observed stellar mass function. This illustrates the importance of carefully specifying the prior distribution for each parameter, especially when a parameter is weakly constrained, and the necessity for fully characterizing the posterior distribution over its full domain.

In summary, the results obtained in this section show that assigning restrictive prior distributions to some parameters can significantly reduce the volume of the parameter space and tighten the constraints on all parameters in galaxy formation models. This has two important implications. Firstly, any prior knowledge, either from observation or physical consideration, can help model inference and hence improve our understanding of galaxy formation. Secondly, it is very dangerous to use unsubstantiated priors to fix model parameters. Since parameters are most likely correlated, fixing one parameter incorrectly can lead to erroneous inference for all other parameters.

6 THE IMPACT OF ERROR MODEL

The observational error model directly influences the value of likelihood and the shape of the cost function in parameter space. However, the impact of error model has not been carefully considered in SAMs. In this section, we explore the effect of incorrect error estimates on the resulting inference.

Astronomers traditionally differentiate two kinds of errors, *statistical errors* and *systematic errors*. Statistical errors result from well-understood processes with known parent distributions (e.g. sampling error) while systematic errors come from the underlying assumptions made for the measurements. From the Bayesian context, a *systematic error* is the result of poor prior information and often results in bias. For the stellar mass function considered here, the total error budget consists of the independent statistical errors of individual stellar mass bins due to the finite number of galaxies used in estimating the stellar mass function, and systematic errors, which arise from the uncertainties of the stellar population synthetic model used to estimate the stellar mass from the observed luminosity. These uncertainties correlate the bins. For example, the uncertainty in the assumed IMF will increase or decrease stellar masses of all galaxies in the same sense.

When errors in different mass bins are correlated, the likelihood function may be approximated as follows:

$$L(\mathbf{\Phi}_{\rm obs}|\theta) = \frac{L_0 \exp[-\frac{1}{2}(\mathbf{\Phi}_{\rm obs} - \mathbf{\Phi}_{\rm mod})^T \cdot \mathbf{\Sigma^{-1}} \cdot (\mathbf{\Phi}_{\rm obs} - \mathbf{\Phi}_{\rm mod})]}{(2\pi)^{I/2}|\det(\mathbf{\Sigma})|^{1/2}}, \tag{17}$$

where $\mathbf{\Phi}_{\rm obs}$ and $\mathbf{\Phi}_{\rm mod}$ are the vectors of the observed and predicted stellar mass functions over the stellar mass bins, Σ is the covariance matrix that describes the correlated error model, and I is the rank of the covariance matrix. For independent errors, all the off-diagonal terms vanish and the likelihood reduces to Eq.(16).

To test the effect of correlated error, we construct a synthetic stellar mass function with correlated errors that mimic the systematic uncertainties in real observation. We choose a truncated series of Chebyshev polynomials to represent the observed stellar mass function. The low-order coefficients specify the overall shape of the function, while the higher orders characterize small scale features. We find that Chebyshev polynomials up to order 4 nicely fits the logarithmic stellar mass function, $\log\Phi(\log m_*)$; the best-fit coefficients are $[-4.17, -1.26, -0.516, -0.274, -0.114]$. We choose the standard deviations of these coefficients $[0.05, 0.10, 0.12, 0.08, 0.03]$ to represent the correlated uncertainties in the measurements. Then, we calculate the covariance matrix of this synthetic data using 1000 Monte Carlo realizations of Chebyshev polynomials and use synthetic data and the derived covariance matrix to constrain the parameters in our SAM.

Figure 7 compares the marginalized posteriori for four pairs of model parameters obtained with the full covariance matrix (upper panels) and those obtained with the diagonal terms only (lower panels). Removing off-diagonal terms is equivalent to ignoring covariance. Clearly the contours produced with the full covariance matrix are more compact. This is expected because the correlation of error between different bins implies that the total independent error in the data is smaller. There are also noticeable changes in the shape and the orientation of the posterior distribution, indicating that it is important to model the covariance of the data properly in order to make correct model inferences.

7 SUMMARY AND DISCUSSION

Many of the physical processes parameterized in semi-analytical models of galaxy formation remain poorly understood and under specified. This has two critically important consequences for inferring constraints on the physical parameters: 1) prior assumptions about the size of the domain and the shape of the parameter distribution will strongly affect resulting inference; and 2) a very large parameter space must be fully explored to obtain an accurate inference. Moreover, both *must* be done together. Both of these issues are naturally tackled with a Bayesian approach that allows one to constrain theory with data in a probabilistically rigorous way. In this paper, we have presented a semi-analytic model of galaxy formation in the framework of Bayesian inference and illustrated its performance on a test problem. Our sixteen-parameter semi-analytic model incorporates the most commonly used parameterizations of important physical processes from existing SAMs including star formation, SN feedback, galaxy merger, and AGN feedback. We combined this model with the Bayesian Inference Engine developed at UMass. The BIE is an extensible MPI-based software package for developing, testing and running advanced Markov-Chain Monte-Carlo algorithms on large data sets. The resulting tool allows the exploration of the posterior distribution of a large number of model parameters, and to constrain models over multiple data sets in a statistically rigorous way.

To demonstrate the power of this approach, we used the observationally derived stellar mass function of galaxies to constrain a number of important model parameters. We find that the posterior distribution has very complex structure and topology, indicating that finding the best fit by tweaking model parameters is improbable. Thus, some of the conclusions drawn previously with the conventional SAM approach are most likely invalid. Moreover, the posterior clearly shows that many model parameters are strongly covariant, and therefore the inferred value of a particular parameter can be significantly affected by the priors used for other parameters. As a consequence, one *may not* tune a small subset of model parameters while keeping other parameters fixed and expect a valid result. This erroneous procedure ignoring the large uncertainties in the parameters that are fixed and is equivalent to imposing strong restrictive priors without scientific motivation. We have demonstrated here that this practice can lead to biased inferences and wrong interpretation of observational data. Finally, with the use of synthetic data to mimic systematic uncertainties in the reduced data, we have shown that resulting model parameter inferences can be significantly affected by the use of incorrect error model. This clearly demonstrates that an accurate analysis of error (both sampling error and systematic uncertainties) is a crucial part in observational data, and conversely, a data-model comparison without an accurate error model is likely to be erroneous.

The method developed here can be straightforwardly applied to other data sets and to multiple data sets simultaneously. Large galaxy surveys available now and in the near future will provide many more data sets to characterize the properties of the galaxy population not only at the present time but also at high redshifts. The Bayesian-inference based SAM described in this paper provides a framework for parameter estimation (e.g constraining the parameters in theoretical models given particular data sets), for hypothesis testing (e.g. comparing the support for particular models given the data), and for predicting the power of new observations to constraining models of interest. In addition, the Bayesian approach explicitly builds on previous results by incorporating the constraints from previous inferences into new data sets; the Bayesian Inference Engine is designed to do this automatically. The approach developed here will produce probabilistically rigorous constraints on theoretical models, and facilitate understanding underlying physical processes that shapes the observed galaxy population. For many processes in galaxy formation, competing models have been proposed but not quantitatively compared. The marginalized likelihood or *Bayes Evidence*, which can be directly derived from the posterior, and explicit model comparison techniques, such as the reversible jump algorithm (Green 1995), can provide a quantitative comparison between different models for given data. The Bayesian hypothesis test can therefore be used to discriminate models. Finally, the prediction for an observable including the inferential uncertainties can be obtained by marginalizing over the posterior. Such predictions can be used to assess the power of new observations. In a series of forthcoming papers, we will use the scheme developed here to make inferences from various data sets, focussing on a number of aspects discussed above.

ACKNOWLEDGMENT

This material is based upon work supported by the National Aeronautics and Space Administration AISR program AISR-126270 and the National Science Foundation under Grant No. III-0611948. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469

Bell E. F., McIntosh D. H., Katz N., Weinberg M. D., 2003, ApJS, 149, 289

Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38

Benson A. J., Bower R. G., 2010, ArXiv e-prints

Benson A. J., Madau P., 2003, MNRAS, 344, 835

Bertschinger E., 1985, ApJS, 58, 39

Bett P., Eke V., Frenk C. S., Jenkins A., Okamoto T., 2010, MNRAS, 256

Bigiel F., Leroy A., Walter F., Brinks E., de Blok W. J. G., Madore B., Thornley M. D., 2008, AJ, 136, 2846

Binney J., Tremaine S., 1987, Galactic dynamics. Princeton, NJ, Princeton University Press, 1987, 747 p.

Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440

Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645

Bower R. G., Vernon I., Goldstein M., Benson A. J., Lacey C. G., Baugh C. M., Cole S., Frenk C. S., ., 2010, ArXiv e-prints

Bruzual A. G., Charlot S., 1993, ApJ, 405, 538

Bullock J. S., Dekel A., Kolatt T. S., Kravtsov A. V., Klypin A. A., Porciani C., Primack J. R., 2001a, ApJ, 555, 240

Bullock J. S., Kolatt T. S., Sigad Y., Somerville R. S., Kravtsov A. V., Klypin A. A., Primack J. R., Dekel A., 2001b, MNRAS, 321, 559

Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, MNRAS, 370, 1651

Cole S., Lacey C., 1996, MNRAS, 281, 716

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168

Croton D. J., Springel V., White S. D. M., De Lucia G., Frenk C. S., Gao L., Jenkins A., Kauffmann G., Navarro J. F., Yoshida N., 2006, MNRAS, 367, 864

De Lucia G., Blaizot J., 2007, MNRAS, 375, 2

Dunkley J., Komatsu E., Nolta M. R., Spergel D. N., Larson D., Hinshaw G., Page L., Bennett C. L., Gold B., Jarosik N., Weiland J. L., Halpern M., Hill R. S., Kogut A., Limon M., Meyer S. S., Tucker G. S., Wollack E., Wright E. L., 2009, ApJS, 180, 306

Dutton A. A., van den Bosch F. C., 2009, MNRAS, 396, 141

Dutton A. A., van den Bosch F. C., Dekel A., Courteau S., 2007, ApJ, 654, 27

Efstathiou G., Davis M., White S. D. M., Frenk C. S., 1985, ApJS, 57, 241

Fardal M. A., Katz N., Weinberg D. H., Davé R., 2007, MNRAS, 379, 985

Fillmore J. A., Goldreich P., 1984, ApJ, 281, 1

Fontanot F., De Lucia G., Monaco P., Somerville R. S., Santini P., 2009, MNRAS, 397, 1776

Gelman A., Rubin D., 1992, Statistical Science, 7, 457

Green P. J., 1995, Biometrika, 82, 711

Gunn J. E., Gott III J. R., 1972, ApJ, 176, 1

Henriques B. M., Bertone S., Thomas P. A., 2008, MNRAS, 383, 1649

Kampakoglou M., Trotta R., Silk J., 2008, MNRAS, 384, 1414

Kang X., Jing Y. P., Mo H. J., Börner G., 2005, ApJ, 631, 21

Katz N., 1992, ApJ, 391, 502

Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999, MNRAS, 303, 188

Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201

Kennicutt Jr. R. C., 1998, ARA&A, 36, 189

Kennicutt Jr. R. C., Calzetti D., Walter F., Helou G., Hollenbach D. J., Armus L., Bendo G., Dale D. A., Draine B. T., Engelbracht C. W., Gordon K. D., Prescott M. K. M., Regan M. W., Thornley M. D., Bot C., Brinks E., de Blok E., de Mello D., Meyer M., Moustakas J., Murphy E. J., Sheth K., Smith J. D. T., 2007, ApJ, 671, 333

Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2

Kimm T., Somerville R. S., Yi S. K., van den Bosch F. C., Salim S., Fontanot F., Monaco P., Mo H., Pasquali A., Rich R. M., Yang X., 2009, MNRAS, 394, 1131

Komatsu E., Dunkley J., Nolta M. R., Bennett C. L., Gold B., Hinshaw G., Jarosik N., Larson D., Limon M., Page L., Spergel D. N., Halpern M., Hill R. S., Kogut A., Meyer S. S., Tucker G. S., Weiland J. L., Wollack E., Wright E. L., 2009, ApJS, 180, 330

Lacey C., Cole S., 1993, MNRAS, 262, 627

Liu L., Yang X., Mo H. J., van den Bosch F. C., Springel V., 2010, ApJ, 712, 734

Lu Y., Mo H. J., Katz N., Weinberg M. D., 2006, MNRAS, 368, 1931

Lu Y., Kereš D., Katz N., Mo H.J., Weinberg M.D., Fardal M., 2010, in preparation

Macciò A. V., Dutton A. A., van den Bosch F. C., Moore B., Potter D., Stadel J., 2007, MNRAS, 378, 55

Martin C. L., Kennicutt Jr. R. C., 2001, ApJ, 555, 301

Mckay M.D., Beckman R.J., Conover W.J., 1979, Technometrics, 21, 239, ArticleType: primary article / Full publication data: May, 1979 / Copyright 1979 American Statistical Association and American Society for Quality

Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319

Mo H. J., Yang X., van den Bosch F. C., Katz N., 2005, MNRAS, 363, 1155

Navarro J. F., Benz W., 1991, ApJ, 380, 320

Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493

Navarro J. F., White S. D. M., 1993, MNRAS, 265, 271

Neal R., 1996, Statistics and Computing, 6, 353

Neistein E., Weinmann S. M., 2009, ArXiv e-prints

Oppenheimer B. D., Davé R., 2006, MNRAS, 373, 1265

Parkinson H., Cole S., Helly J., 2008, MNRAS, 383, 557

Press W. H., Schechter P., 1974, ApJ, 187, 425

Primack J. R., 2009, in American Institute of Physics Conference Series, Vol. 1192, American Institute of Physics Conference Series, F. Roig, D. Lopes, R. de La Reza, & V. Ortega, ed., pp. 101–137

Primack J. R., Gilmore R. C., Somerville R. S., 2008, in American Institute of Physics Conference Series, Vol. 1085, American Institute of Physics Conference Series, F. A. Aharonian, W. Hofmann, & F. Rieger, ed., pp. 71–82

Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1

Simha V., Weinberg D. H., Davé R., Gnedin O. Y., Katz N., Kereš D., 2009, MNRAS, 399, 650

Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481

Somerville R. S., Kolatt T. S., 1999, MNRAS, 305, 1

Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, MNRAS, 320, 289

Springel V., 2005, MNRAS, 364, 1105

Sutherland R. S., Dopita M. A., 1993, ApJS, 88, 253

Ter Braak C. J. F., 2006, Stat. Comput., 16, 239

van den Bosch F. C., 2002, MNRAS, 331, 98

Warren M. S., Quinn P. J., Salmon J. K., Zurek W. H., 1992, ApJ, 399, 405

Weinberg M. D., 2010a, The UMass Bayesian Inference Engine, http://www.astro.umass.edu/BIE

Weinberg M. D., 2010b, in preparation

Weinmann S. M., van den Bosch F. C., Yang X., Mo H. J., 2006, MNRAS, 366, 2

White S. D. M., Frenk C. S., 1991, ApJ, 379, 52

White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2003a, ApJL, 597, L9

—, 2009, ApJ, 707, 354

Zhao D. H., Mo H. J., Jing Y. P., Börner G., 2003b, MNRAS, 339, 12

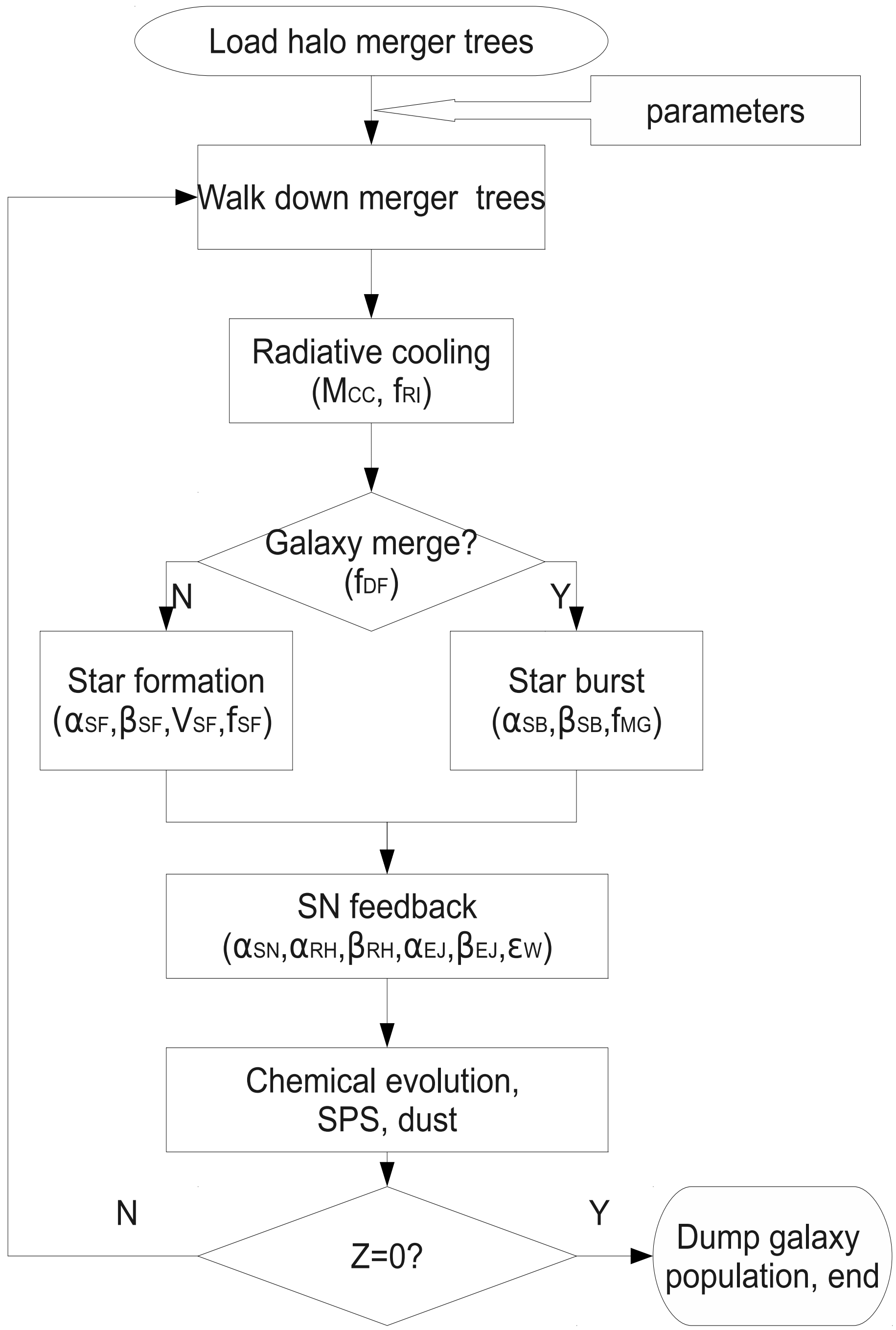


Figure 1. Flow chart describing the calculation of the semi-analytic model of galaxy formation. The parameters explored in the present paper are listed in the corresponding blocks.

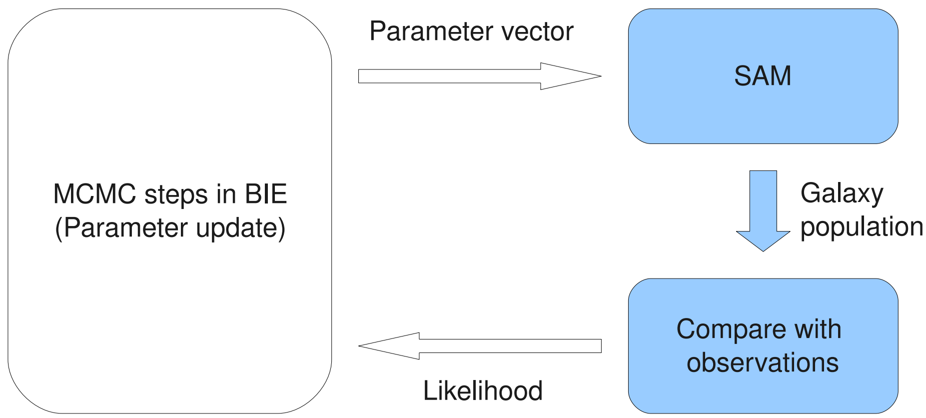


Figure 2. A flow chart describing the structure of our Bayesian-inference based semi-analytic model.

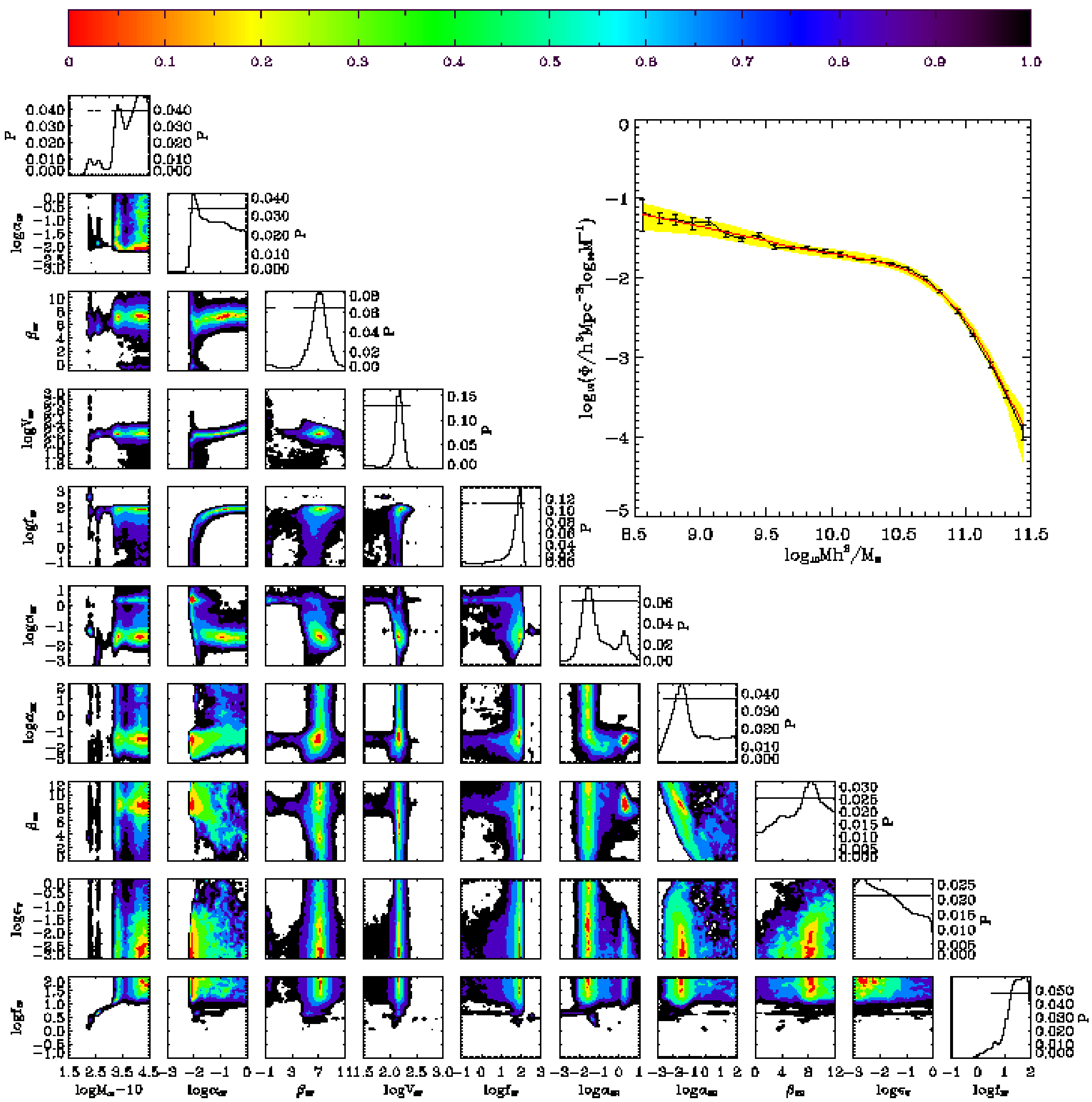
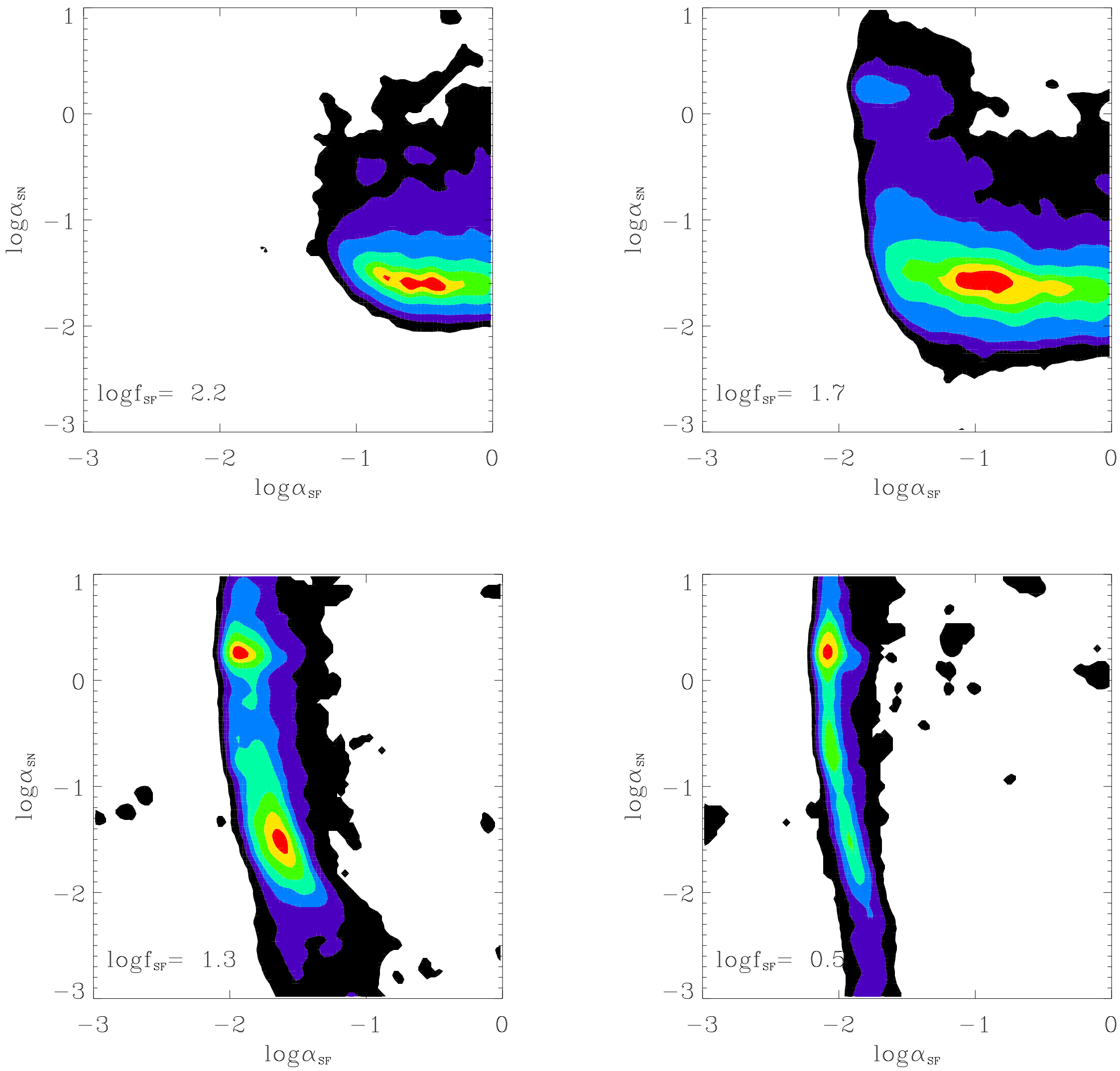


Figure 3. The marginalized posterior distribution for key parameters for our fiducial run (Case 0). The color coding represents confidence levels as shown by the color-bar on the top of the figure. The horizontal bars in the one-dimensional marginals indicate the 95% confidence interval. The observed stellar mass function of galaxies (black line and error bars) from (Bell et al. 2003) together with the marginalized model prediction is inset. The red solid line shows the median value of the prediction, while the yellow shaded region represents the 95% confidence interval.



[]

Figure 4. Four two-dimensional slices of the marginalized posterior at $\log f_{\rm SF} = 2.2$, 1.7, 1.3 and 0.5 with thickness ± 0.3. This figure demonstrates that the breadth of the marginalized distribution (cf. Fig. 3) is mainly due to projection. The structure of the posterior in the 13-dimensional parameter space is dramatically more compact and more complex than the marginalized posterior.

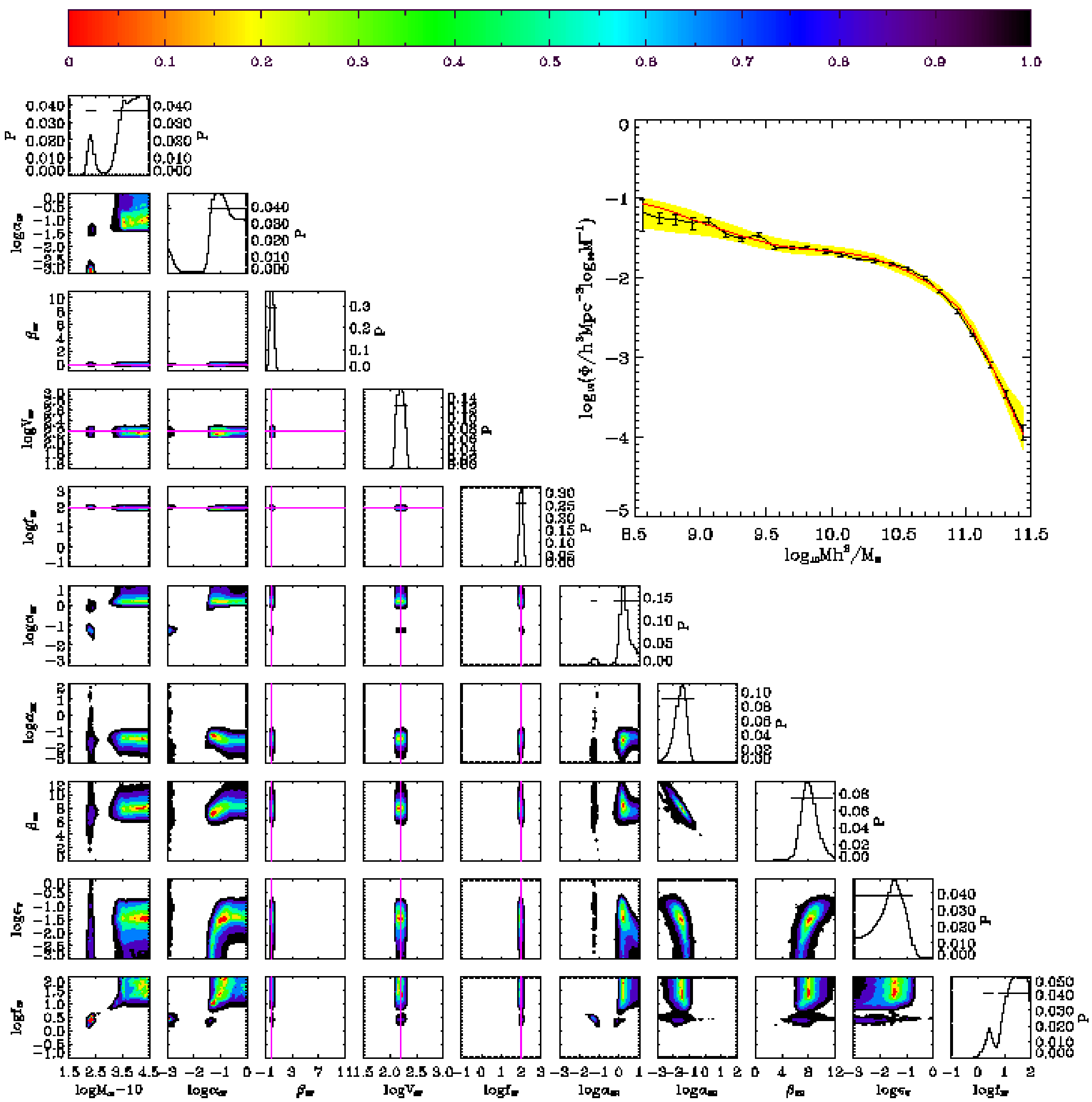


Figure 5. The marginalized posterior distribution for key parameters in Case 1.Very restrictive priors are assumed for the parameters, $\beta_{\rm SF}$, $V_{\rm SF}$ and $f_{\rm SF}$, whose central values are indicated by magenta lines.

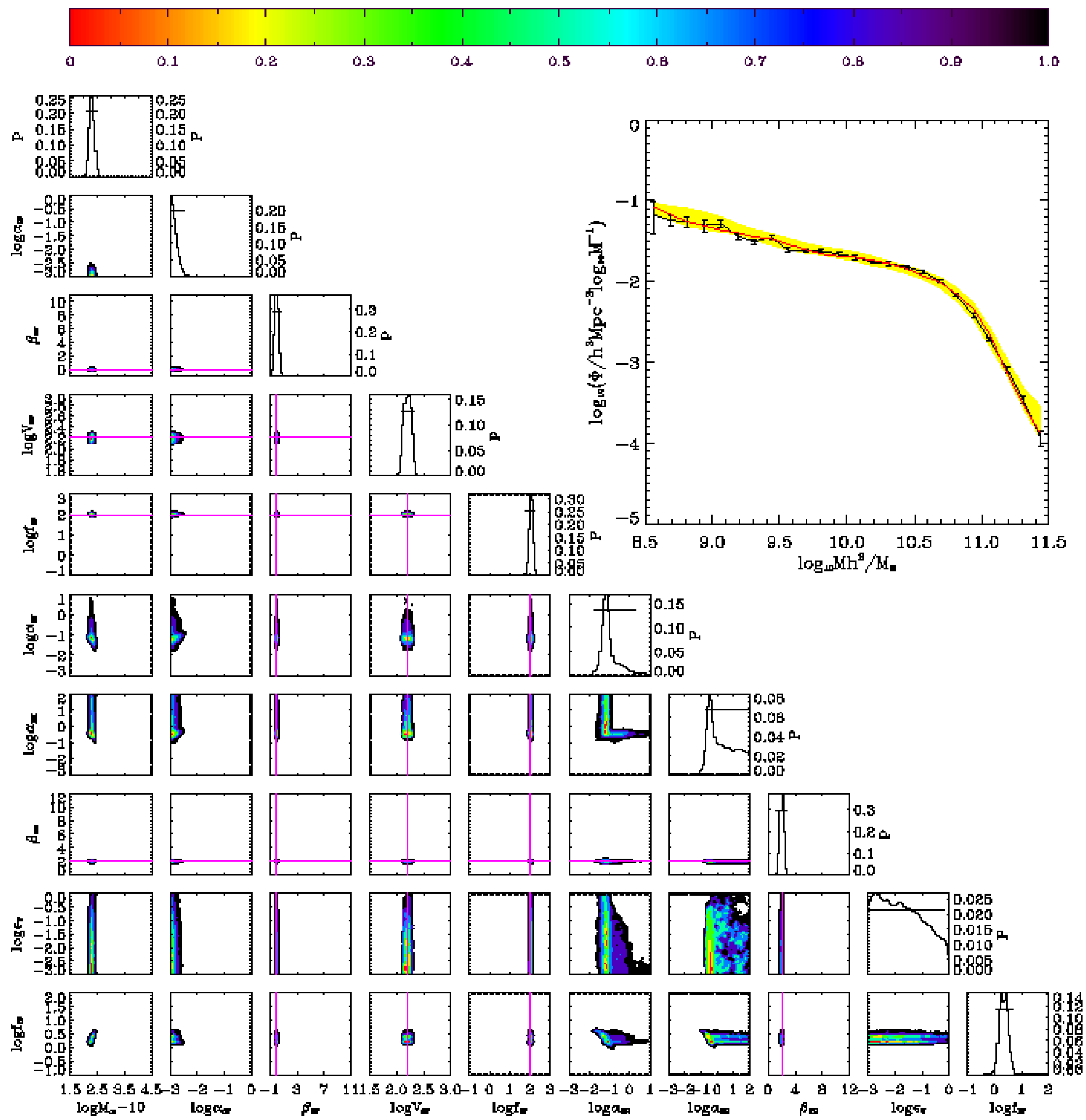


Figure 6. The marginalized posterior distribution of key parameters for Case 2. This includes the restrictions of Case 1 as in Fig. **??** with the additional restrictive prior for β_{RH}.

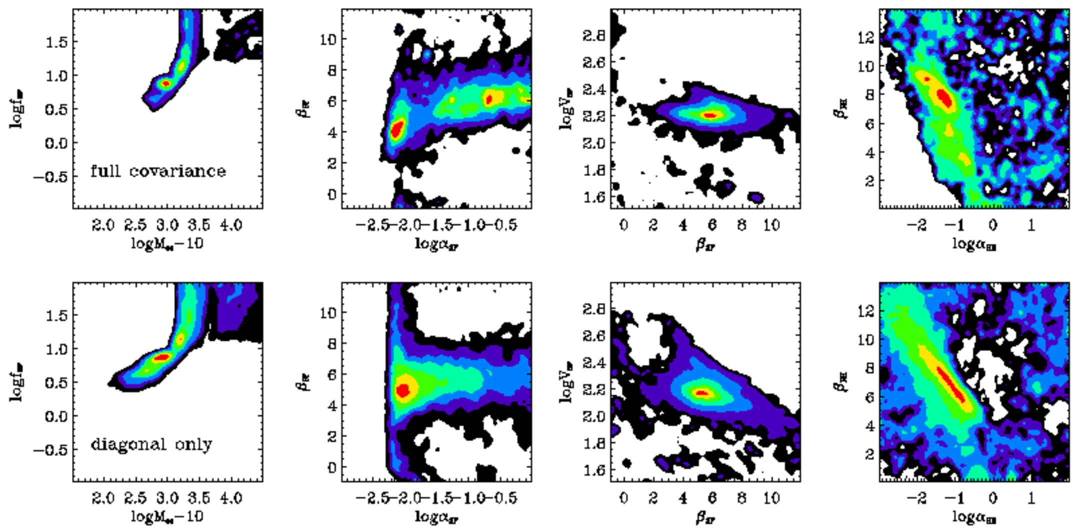


Figure 7. A comparison of the posterior distribution obtained for a likelihood function including covariance (upper row, eq. 17), and using only the diagonal terms of the covariance matrix (lower row).