ELUCID. VI. Cosmic Variance of the Galaxy Distribution in the Local Universe

Yangyao Chen1, H. J. Mo1,2 ☄️, Cheng Li1, Huiyuan Wang1,4 ☄️, Xiaohu Yang5,6 ☄️, Shuang Zhou1, and Youcai Zhang7 ☄️
1 Center for Astrophysics and Physics Department, Tsinghua University, Beijing 100084, People’s Republic of China; yangyao-17@mails.tsinghua.edu.cn
2 Department of Astronomy, University of Massachusetts, Amherst MA 01003-9305, USA
3 Key Laboratory for Research in Galaxies and Cosmology, Department of Astronomy, University of Science and Technology of China, Hefei, Anhui 230026, People’s Republic of China
4 School of Astronomy and Space Science, University of Science and Technology of China, Hefei 230026, People’s Republic of China
5 Department of Astronomy, Shanghai Jiao Tong University, Shanghai 200240, People’s Republic of China
6 IFSA Collaborative Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, People’s Republic of China
7 Shanghai Astronomical Observatory, Shanghai 200030, People’s Republic of China

Received 2018 September 3; revised 2019 January 7; accepted 2019 January 24; published 2019 February 25

Abstract

Halo merger trees are constructed from ELUCID, a constrained N-body simulation in the Sloan Digital Sky Survey (SDSS) volume. These merger trees are used to populate dark matter halos with galaxies according to an empirical model of galaxy formation. Mock catalogs in the SDSS sky coverage are constructed, which can be used to study the spatial distribution of galaxies in the low-z universe. These mock catalogs are used to quantify the cosmic variance (CV) in the galaxy stellar mass function (GSMF) measured from the SDSS survey. The GSMF estimated from the SDSS magnitude-limited sample can be affected significantly by the presence of an underdense region at \( z < 0.03 \), so that the low-mass end of the function can be underestimated significantly. Several existing methods designed to deal with the effects of the CV in the estimate of the GSMF are tested, and none is found to be able to fully account for the CV. We propose a method based on conditional stellar mass functions in dark matter halos, which can provide an unbiased estimate of the global GSMF. The application of the method to SDSS data shows that the GSMF has a significant upturn at \( M_\ast < 10^{9.5} \, h^{-1} M_\odot \), which has been missed in many earlier measurements of the local GSMF.

Key words: dark matter – galaxies: halos – large-scale structure of universe – methods: statistical

1. Introduction

The universe contains prominent structures up to \( \sim 100 \) Mpc, only reaching homogeneity on much larger scales (e.g., Peebles 1980; Davis et al. 1985). The properties of galaxies and other objects, which form and evolve in the cosmic web, are expected to be affected by their large-scale environments. Thus, astronomical observations, which are always made in limited volumes in the universe, can be affected by the cosmic variance (CV) caused by spatial variations of the statistical properties of cosmic objects, such as galaxies, due to the presence of large-scale structure. Because of CV, statistics obtained from a sample that covers a specific volume in the universe may be different from those expected for the universe as a whole. Erroneous inferences would then be made if such biased observational data were used to constrain models.

CV is a well-known problem (e.g., Somerville et al. 2004; Jha et al. 2007; Driver & Robotham 2010; Moster et al. 2011; Keenan et al. 2013; Marra et al. 2013; Whitbourn & Shanks 2014, 2016; Wojtak et al. 2014), and various attempts have been made to deal with it. One way is to analyze different (sub)samples, e.g., that obtained from the jackknife sampling of observational data, and to use the variations among them to have some handle on the CV. However, this can only provide information about the variance within the total sample itself, but not that of the total sample relative to a fair sample of the universe. Another way is to use the spatial distribution of bright galaxies (a density-defining population), which can be observed in a large volume, to quantify the CV expected in subvolumes (e.g., Driver & Robotham 2010), or to re-scale (or correct) the number density of faint galaxies observed in a smaller volume, as was done by Baldry et al. (2012) in their estimate of galaxy stellar mass functions (GSMFs) in the GAMA sample. However, this method relies on the assumption that galaxies of different luminosities/masses have similar spatial distributions, which may not be true. The same problem also exists in the maximal likelihood method (e.g., Efstathiou 1988), where the galaxy luminosity function (GLF) is explicitly assumed to be independent of environment. Yet another way is to estimate the CV expected from a given sample using simple, analytic models for the clustering properties of galaxies on large scales. Along this line, Somerville et al. (2004) tested the effects of the CV on different scales and proposed the use of either the two-point correlation function of galaxies, or the combination of the linear density field with halo bias models (e.g., Mo & White 1996; Sheth et al. 2001), to predict the CV of different surveys. Similarly, Moster et al. (2011) carried out an investigation of the CV expected in observations of the galaxy populations at different redshifts, using the linear density field predicted by the \( \Lambda \)CDM model combined with a bias model that takes into account the dependence of galaxy distribution on galaxy mass and redshift. Unfortunately, such an approach does not take into account observational selection effects. More importantly, this approach only gives a statistical estimate of the CV but does not measure the deviation of a specific sample from a fair sample. Finally, one can also use a large number of mock galaxy samples, either obtained directly from hydrodynamic simulations, or from \( N \)-body simulation-based semianalytic (SAM) and empirical models, to quantify how the sample-to-sample variation of the statistical measure in question depends on sample volume. However, this needs a large set of simulations for each model, analyzed in a way that takes into account the observational selection effects in the data, which in practice is costly and time consuming.
Furthermore, the same as the approach based on galaxy clustering statistics, this approach can only provide a statistical statement of the expected CV, but does not provide a way to correct the variance of a specific sample.

Can one develop a systematic method to study the CV, and to quantify and correct biases that are present in observational data? The answer is yes, and the key is to use constrained simulations. Indeed, if one can accurately reconstruct the initial conditions for the formation of the structures in which the observed galaxy population resides, one can then carry out simulations with such initial conditions in a sufficiently large box that contains the constrained volume, so that the large box can be used as a fair sample, while the constrained region can be used to model the observational data. By comparing the statistics obtained from the mock samples with those obtained from the whole box, one can quantify and correct the CV in the observational data.

Over the past few years, the ELUCID collaboration has embarked on the development of a method to accurately reconstruct the initial conditions responsible for the density field in the low-\(z\) universe (Wang et al. 2014). As demonstrated by various tests (Wang et al. 2014, 2016), the reconstruction method is much more accurate than other methods that have been developed and works reliably even in highly nonlinear regimes. The initial conditions in a 500 \(h^{-1}\)Mpc box that contains the main part of the Sloan Digital Sky Survey (SDSS) volume have already been obtained, and a high-resolution \(N\)-body simulation, run with (3072\(^3\) particles, has been carried out with these initial conditions in the current \(\Lambda\)CDM cosmology (Wang et al. 2016).

In the present paper, we use the dark matter halo merger trees constructed from the ELUCID simulation to populate simulated halos with model galaxies predicted by the empirical galaxy formation model developed by Lu et al. (2014b, 2015b, hereafter L14, L15). The model galaxies in the constrained volume are then used to construct mock catalogs that contain the same CV as the real SDSS sample. We compare the GSMF estimated from the mock catalogs with that obtained from the total simulation box to quantify the CV within the SDSS volume. Finally, we propose a method based on the conditional stellar mass or luminosity distribution in dark matter halos to correct for the CV in the observed GSMF. As we will see, the CV can be very severe in the low-mass end of the GSMF obtained from methods commonly adopted, and the slope of the low-mass end of the true GSMF in the low-\(z\) universe may be significantly steeper than that published in the literature.

The structure of this paper is as follows. In Section 2, we describe methods to implement Monte Carlo halo merger trees in simulated merger trees, so as to extend all trees down to a mass resolution sufficient for our purpose. In Section 3, we populate simulated halos with galaxies using an empirical model of galaxy formation and construct a number of mock catalogs to mimic the SDSS survey both in spatial distribution and physical properties. In Section 4, we examine in detail the CV in the estimates of the GSMF and show how commonly adopted methods to measure the GLF and GSMF fail to account for the CV. We also propose and test a new method to correct for CV in GLF and GSMF, and apply it to the real SDSS data to obtain the CV-corrected GLF and GSMF. Finally, a brief summary of our main results is presented in Section 5.

Throughout the paper, we define the GSMF as \(\Phi(M_\ast) = dN/dV/d \log M_\ast\), which is the number of galaxies per unit volume per unit stellar mass in logarithmic space, and define the GLF in the \(X\)-band as \(\Phi(X) = dN/dV/d(M_X - 5 \log h)\), which is the number of galaxies per unit volume per unit magnitude. The magnitude \(M_X\) is \(k\)-corrected to redshift 0.1 without evolution correction, unless specified otherwise.

2. Merger Trees of Dark Matter Halos from the ELUCID Simulation

2.1. The Simulation

We use the ELUCID simulation carried out by Wang et al. (2016) to model the dark matter halo population, their formation histories, and spatial distribution. This is an \(N\)-body simulation that uses L-GADGET, a memory-optimized version of GADGET-2 (Springel 2005), to follow the evolution of 3072\(^3\) dark matter particles (each with a mass of \(3.088 \times 10^8 h^{-1}M_\odot\)) in a periodic cubic box with side length of 500 \(h^{-1}\)Mpc in comoving units. The cosmology used is the one based on WMAP5 (Dunkley et al. 2009; Komatsu et al. 2009): a flat universe with \(\Omega_\Lambda = 0\), a matter density parameter \(\Omega_{m,0} = 0.258\), a cosmological constant \(\Omega_{\Lambda,0} = 0.742\), a baryon density parameter \(\Omega_{\text{B,0}} = 0.044\), a Hubble constant \(H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}\) with \(h = 0.72\), and a Gaussian initial density field with power spectrum \(P(k) \propto k^n\), with \(n = 0.96\) and with the amplitude specified by \(\sigma_8 = 0.80\). The simulation is run from redshift \(z = 100\) to \(z = 0\), with outputs recorded at 100 snapshots between \(z = 18.4\) and \(z = 0.0\).

The initial conditions (phases of Fourier modes) of the density field are those obtained from the reconstruction based on the halo-domain method of Wang et al. (2009) and the Hamiltonian Markov Chain Monte Carlo (HMC) method (Wang et al. 2013, 2014), constrained by the distributions of dark matter halos represented by galaxy groups and clusters selected from the SDSS redshift survey in the Northern Galactic Cap (NGC) region. This has a sky coverage of 7047 deg\(^2\) and a redshift range of 0.01 \(\leq z \leq 0.12\), constituting about 22\% of the simulation box (Yang et al. 2007, 2012). As shown in Wang et al. (2016) using mock catalogs, more than 95\% of the groups with masses above \(10^{14} h^{-1}M_\odot\) can be matched with simulated halos of similar masses, with a distance error tolerance of \(\sim 4\ h^{-1}\)Mpc, and massive structures such as the Coma cluster and the Sloan Great Wall can be well reproduced in the reconstruction. Thus, the use of the constrained simulation from ELUCID allows us not only to model accurately the large-scale environments within which observed galaxies reside, but also to recover, at least partially, the formation histories of the massive structures seen in the local universe.

2.2. The Construction of Halo Merger Trees

Halos and their subhalos with more than 20 particles are identified with the friend-of-friend (FOF) and SUBFIND algorithms (Springel et al. 2005). To be safe, we only use halos identified in the simulation with masses \(M_h \geq M_{\min} = 10^{10} h^{-1}M_\odot\). However, this mass resolution is not sufficient to resolve lower mass halos in which star formation may still be significant, particularly at high \(z\). In order to trace the star formation histories in halos to high redshifts, we need to reach a halo mass of about \(10^9 h^{-1}M_\odot\) below which star formation is expected to be unimportant due...
to photoionization heating (e.g., Babul & Rees 1992; Thoul & Weinberg 1996). Here we adopt a Monte Carlo method to extend the merger trees of the simulated halos down to a mass limit, $10^9 h^{-1} M_\odot$. Jiang & van den Bosch (2014) have tested the performances of several different methods of generating Monte Carlo halo merger trees, and found that the method of Parkinson et al. (2008, hereafter P08) consistently provides the best match to the halo merging trees obtained from N-body simulations. We therefore adopt the P08 method.

We join the P08 Monte Carlo trees to the halo merger trees obtained from the simulation through the following steps:

(i) For each simulated halo merger tree $T$, we eliminate halos that have masses below $M_{th} = 10^{10} h^{-1} M_\odot$, but have no progenitors more massive than $M_{th}$. The purpose of the second condition is to preserve halos that once had masses larger than $M_{th}$ but have become less massive later due to stripping and/or mass loss.

(ii) For each halo $H$ that is not eliminated in $T$, we generate a Monte Carlo tree $t$ (down to $10^9 h^{-1} M_\odot$), rooted in a halo $h$ that has the same mass and the same redshift as $H$, and eliminate all halos more massive than $10^{10} h^{-1} M_\odot$ in $t$.

(iii) We add $t$ to $H$. The procedure is repeated for all halos with masses above $10^{10} h^{-1} M_\odot$ in all trees in the ELUCID simulation, so that all such halos have merger trees extended to $10^9 h^{-1} M_\odot$.

(iv) For halos with masses below $10^{10} h^{-1} M_\odot$ at $z = 0$, their merger trees are entirely generated with the Monte Carlo method. Note that these halos are not identified from the simulation, but can be used to model galaxies in such low-mass halos when needed.

With these steps, we obtain “repaired” halo merger trees that have a mass resolution of $10^9 h^{-1} M_\odot$, with halos more massive than $10^{10} h^{-1} M_\odot$ sampled entirely by the simulation, and the less massive ones modeled by Monte Carlo trees. Figure 1 shows the conditional progenitor mass functions of dark matter halos, defined as the fraction of mass in progenitors per logarithmic mass, for merger trees rooted in different masses, and for progenitors at different redshifts. Our results, obtained by combining the simulated trees above the mass resolution $M_{th}$ with the Monte Carlo merging trees generated with the P08 model below the mass limit, are shown by the black solid lines and compared with the merger trees generated entirely with the P08 model. Overall, the progenitor mass distributions we obtain match well those obtained from the Monte Carlo method, indicating that our merger trees are reliable.

Because galaxies form and evolve in dark matter halos, our “repaired” halo merger trees from the ELUCID simulation provide the basis for linking galaxy properties to dark matter halos, and can be used in combination with halo-based methods of galaxy formation, such as abundance-matching and semianalytic and other empirical models, to populate halos with galaxies. The method can, in principle, be applied to simulated halos with any mass resolution and with any cosmology, to extend halo merger trees to a sufficiently low mass, as long as reliable Monte Carlo trees can be generated. We note that our merger trees do not include high-order subhalos, i.e., subhalos in subhalos. In the next section, we apply the empirical model, developed in L14 and L15, to follow galaxy formation and evolution in dark matter halos, based on our repaired halo merger trees.

3. Populating Halos with Galaxies

In this section, we describe the L14 and L15 empirical method, developed by Lu et al. (2014b, 2015b), to populate galaxies in the halo merger trees described in the previous section. Briefly, we assign a central galaxy to each distinctive halo and give it an appropriate star formation rate (SFR) according to the empirical model. We then evolve all galaxies in the current snapshot to the next, following the accretion of galaxies by dark matter halos and the mergers of galaxies. The stellar masses for both central and satellite galaxies are obtained by integrating the stellar contents along their histories. Finally, observable quantities, such as luminosity and apparent magnitude, are obtained from a stellar population synthesis model.

3.1. The Empirical Model of Galaxy Formation

In the model of L14 and L15, the SFR of a central galaxy is assumed to depend on the halo mass $M_{halo}$ and redshift $z$ as

$$\text{SFR}(M_{halo}, z) = \frac{\tau_B M_{halo}}{\tau_0} (1 + z)^{\alpha}(X + R)^{\gamma} \left(\frac{X + 1}{X + R}\right)^{\beta},$$

(1)

where $\tau_0 = 1/(10H_0)$, $\alpha$ and $\beta$ are time-independent model parameters, $\gamma$ is a time-dependent parameter, and $X = \frac{M_{halo}}{M_{tidal}}$. For $z < z_c$, $\gamma = \gamma_0$, $\alpha_0 = 1$, and $\beta = 0$; for $z \geq z_c$, $\gamma = \gamma_a(1 + z_c)^{\gamma_b}$ and $\alpha = \alpha_0(1 + z)^{\alpha_a}$.

All model parameters are determined by fitting the model predictions to a set of observational data (see the original papers for details). Here we adopt the parameters listed in L14 (denoted by “Model III SMF+CGLF” in this paper), which are based on a cosmology consistent with the WMAP5 cosmology (Dunkley et al. 2009; Komatsu et al. 2009) used here.

Once a dark matter halo hosting a galaxy is accreted by a bigger halo, the central galaxy in it is assumed to become a satellite galaxy and thus experience satellite-specific processes, such as tidal stripping and ram pressure stripping, which may reduce and quench its star formation. L14 modeled the SFR in satellites as

$$\text{SFR}(M_h, z) = \text{SFR}(t_{accr}) \exp \left(\frac{-t - t_{accr}}{\tau(M_h)}\right),$$

(4)

with

$$\tau(M_h) = \tau_{h,0}\exp(-M_h/M_{h,c}).$$

(5)
Here, $M_\text{sat}$ is the current mass of the satellite galaxy, $\tau_0$ and $M_\text{acc}$ are time-independent model parameters, and $t_{\text{acc}}$ is the cosmic time at which the host halo of the galaxy is accreted. After accretion, the satellite halo and the galaxies it hosts are expected to experience dynamical friction, which causes them to move toward the inner part of the new host halo. The satellites may then merge with the central galaxy located near the center. We follow L14 and use an empirical model to determine the time when the merger occurs:

$$\Delta t = 0.216 \frac{Y^{1.3}}{\ln(1 + Y)} \exp(1.9\eta) \frac{r_{\text{halo}}}{v_{\text{halo}}}$$  \hspace{1cm} (6)$$

where $Y = M_{\text{cen}}/M_{\text{sat}}$ is the ratio of the mass of the central halo to that of the satellite halo at the time when accretion occurs, and $r_{\text{halo}}$ and $v_{\text{halo}}$ are the virial radius and virial velocity of the central halo (e.g., Boylan-Kolchin et al. 2008). The parameter, $\eta$, describes the specific orbital angular momentum and is assumed to follow a probability distribution $P(\eta) = \eta^{1.2}(1 - \eta)^{1.2}$ (e.g., Zentner et al. 2005). After merger, a fraction of $f_{\text{TS}}$ of the stellar mass of the satellite is added to the central galaxy, with $f_{\text{TS}}$ a model parameter.

The ingredients given above can be used to predict the stellar mass and SFR of both central and satellite galaxies. In order to make predictions for galaxy luminosities in different bands, we also need the metallicities of stars. We use the mean metallicity—stellar mass relation given by Gallazzi et al. (2005) to assign metallicities to galaxies according to their masses. A simple stellar population synthesis model, based on Bruzual & Charlot (2003) with a Chabrier initial mass function (Chabrier 2003), is adopted to obtain the mass-to-light ratio of the formed star and the mass loss due to stellar evolution.
We note that the L14 model, which is based on Monte Carlo merger trees, does not take into account some special events that exist in numerical simulations. In simulated merger trees, some subhalos were main halos at some early times, accreted into other systems as satellites later, and were eventually ejected and became main halos again. For such cases, we treat the galaxy in the subhalo as a satellite galaxy even after the subhalo is ejected. The ejected subhalos are then treated as new main halos after ejection. This implementation does not make much physical sense, but best mimic the Monte Carlo merger trees in which subhalos are never ejected, and all halos at a given time are treated equally without depending on whether or not they have gone through a big halo. Such an implementation is necessary, as the model parameters given by L14 are calibrated using Monte Carlo merger trees.

Figure 2 shows the GSMF of model galaxies at redshift \( z = 0 \), in black solid line, in comparison with the result of L14 (purple solid line). As one can see, the L14 result is well reproduced over wide ranges of stellar masses, which demonstrates that our implementation of the L14 model with the ELUCID halo merger trees is reliable, as long as the general galaxy GSMF is concerned. For reference, we also include the observational data points (green dots with error bars) that were used in L14 to constrain their model parameters.

As a more demanding test, we compare in Figure 3 the conditional galaxy stellar mass functions (CGSMFs) in halos of different masses at redshift \( z = 0 \) obtained from the ELUCID halo merger trees with those given by L14. Here again we see a good agreement between the two. Since the CGSMF gives the average number of galaxies of a given stellar mass in a halo of a given mass, a good match in CGSMFs also implies that the spatial clustering of galaxies as a function of stellar mass is also reproduced.

3.2. Galaxy Occupation in Dark Matter Halos

To use our model galaxies to construct mock catalogs, we need to assign spatial positions and peculiar velocities to galaxies in each halo in the simulation according to the halo occupation distributions obtained from the empirical model described above. Here we adopt a subhalo abundance-matching method that links galaxies in a halo to the subhalos in it. As shown in Wang et al. (2016), the subhalo population can be identified reliably from the ELUCID simulation for subhalos with masses down to \( \sim 10^{10} h^{-1} M_\odot \). The abundance matching goes as follows. For a given halo, we first rank galaxies in descending order of stellar mass and subhalos in descending order of halo mass. Here the mass of a subhalo is that at the time when the subhalo was first accreted into its host. Note that subhalos both identified directly from the simulation and added using Monte Carlo merger trees (see Section 2.2) are used. For subhalos identified in the simulation, their positions and velocities are those given by SUNFIND. For the Monte Carlo subhalos that are joined to the simulated halos, on the other hand, we assign random positions according to the NFW profile (Navarro et al. 1997) with concentration parameters given by Zhao et al. (2009), and their velocities are drawn from a Gaussian distribution with dispersion appropriate for the density profile assumed.

This method can be used to construct volume-limited samples within the entire simulation box down to stellar masses.
~10^8 h^{-1}M_\odot, with full phase space information obtained from the simulated subhalos. This is sufficient for most of our purposes.

3.3. The SDSS Mock Catalog

With full information about the luminosities and phase space coordinates for individual galaxies, it is straightforward to make mock catalogs using galaxies in the constrained volume and applying the same selection criteria as in the observation. For each model galaxy in the simulation box, we assign to it a cosmological redshift, $z_{\cos}$, according to its distance to a virtual observer, and the observed redshift, $z_{\text{obs}}$, is given by $z_{\cos}$ together with its line-of-sight (los) peculiar velocity, $v_{\text{los}}$:

$$z_{\text{obs}} = z_{\cos} + (1 + z_{\cos}) \frac{v_{\text{los}}}{c},$$

(7)

with $c$ the speed of light. Here, the location of the virtual observer and the coordinate system are determined by the orientation of the SDSS volume in the simulation box. SDSS apparent magnitudes in $u$, $g$, $r$, $i$, and $z$ are assigned to each galaxy according to its luminosities in the corresponding bands. For our SDSS mock sample, we select all galaxies in the SDSS NGC region with redshifts $0.01 < z < 0.12$ and with magnitude $r \leq 17.6$.

Figure 4 shows the real SDSS galaxies (left) and the mock galaxies (right) in the same slice in the SDSS sky coverage. It is clear that the distribution of the mock galaxies is very similar to that of the real galaxies. The large-scale structures in the local universe, such as the Sloan Great Wall at redshift $\approx 0.08$, are well reproduced. Thus, the mock catalog can be used to investigate both the properties of the galaxy population in the cosmic web and the large-scale clustering of galaxies. In particular, because all galaxies above our mass resolution limit, which is about $10^8 h^{-1}M_\odot$, are modeled in the entire simulation box, a comparison of the statistical properties between the SDSS mock catalog and the whole simulation box carries information about the CV of the SDSS sample.

4. CV in GSMFs

The realistic model catalogs described above have many applications, such as to study the relationships between galaxies and the mass density field, and to investigate the galaxy population in different components of the cosmic web. Here we use them to analyze and quantify the CVs in the measurements of the GSMF and luminosity function (GLF). We first use model galaxies in the whole simulation box to quantify the CV as a function of sample volume and galaxy mass. We then use the SDSS mock catalog to examine the CV in the SDSS, and to investigate different estimates of the GSMF/GLF in their abilities to account for the CV. We propose and test a new method that can best correct for the CV. Finally, we apply our method to the SDSS catalog to obtain GLF and GSMF that are free of the CV.

4.1. CV as a Function of Sample Volume and Galaxy Mass

To quantify the effects of CV, we partition the whole $500^3 h^{-1}$ Mpc$^3$ simulation box into subboxes, each with a given size $L_s$, without overlaps. For each subbox $i$, we calculate the galaxy number density $n_{g,i}(M_s; L_s)$. Figure 5 shows the GSMFs obtained for 100 subboxes with sizes $L_s = 25, 50, 100,$ and $250$ \, $h^{-1}$Mpc, respectively. The results of individual subboxes are shown by the green lines, while the average and 2$\sigma$ variance (96\%) among the GSMFs are shown by the red curve and bars, respectively. As expected, the scatter among the subboxes decreases as the subbox size increases. For instance, the scatter for $L_s = 50$ \, $h^{-1}$Mpc is about $\approx 0.3$ dex over almost the entire stellar mass range, while it is smaller than 10\% for $L_s = 250$ \, $h^{-1}$Mpc.

Theoretically, the galaxy number density $n_g$ is related to the mass density $\rho_m$ by a stochastic bias relation:

$$\delta_g = b\delta_m + \epsilon,$$

(8)

where $\delta_g = (n_g/\bar{n}_g) - 1$ and $\delta_m = (\rho_m/\bar{\rho}_m) - 1$, with $\bar{n}_g$ and $\bar{\rho}_m$ being the mean number density of galaxies and the mean density of mass in the universe. The coefficient, $b$, is the bias parameter, which characterizes the deterministic part of the bias relation, and $\epsilon$ is the stochastic part. If the galaxy number density field is Poisson sampling of the mass density field, then the variance in the galaxy density can be written as

$$\sigma_g^2 = \sigma_{CV}^2 + \sigma_p^2,$$

(9)

where $\sigma_p = N^{-1/2}$ is due to Poisson fluctuation. Assuming linear bias, the deterministic part, which we refer to as the CV, can be written as

$$\sigma_{CV}^2(M_s; L_s) = b^2\sigma_m^2(L_s),$$

(10)

where $L_s$ is the characteristic size of the sample, and $\sigma_m(L_s)$ is the rms of the mass fluctuation on the scale of $L_s$.

Motivated by this, we model $\sigma_{CV}$ using the GSMF obtained from simulated galaxies. The number density $n_{\delta,g}$ of all subboxes is synthesized to give the mean value, $\bar{n}_g(M_s; L_s)$, and the variance, $\sigma_g^2(M_s; L_s)$. We use $\bar{n}_g$ to estimate the expected Poisson variance, $\sigma_p^2$, and use Equation (9) to estimate
\(\sigma_{\text{CV}}^2(M_\star; L_s)\) by subtracting the Poisson part from the total variance. Equation (10) is then used to fit the dependence of the CV on the stellar mass and the size of the subbox. We find that the \(L_s\) dependence can be well described by

\[
\log \sigma_m(x) = p_0 + p_1 x + p_2 x^2 + p_3 x^3,
\]

where \(x = \log(L_s/h^{-1}\text{Mpc})\), and \(p_0 = 1.53, p_1 = -2.02, p_2 = 0.92,\) and \(p_3 = -0.25,\) while the \(M_\star\) dependence is well described by

\[
\log b(y) = q_0 + q_1 y + q_2 y^2 + q_3 y^3,
\]

where \(y = \log(M_\star/h^{-1}\text{M}_\odot)\), and \(q_0 = -16.04, q_1 = 5.79, q_2 = -0.68,\) and \(q_3 = 0.026.\)

Figure 6 shows the comparison between \(\sigma_{\text{CV}}\) obtained directly from the simulated galaxy sample and the model prediction as a function of \(L_s\) for galaxies of different \(M_\star\), as indicated in the panel. Solid lines are the \(\sigma_{\text{CV}}\) estimated from the mock sample, while dashed lines are from the fitting formula. Lower panel: ratio of \(\sigma_{\text{CV}}\) between the mock sample and model prediction. The black solid line indicates the ratio of 1.0.

Figure 5. GSMFs at \(z = 0\) in subboxes in the 500 \(h^{-1}\text{Mpc}\) box of the ELUCID simulation. For each subbox size \(L_{\text{box}} \leq 100 h^{-1}\text{Mpc},\) 100 subboxes without overlap are randomly chosen in the simulation box, while for \(L_{\text{box}} = 25 h^{-1}\text{Mpc},\) all eight subboxes are used. The GSMFs of individual subboxes are shown by the green curves in each panel. The average over the subboxes of the same size is given by the red line in each panel. Error bars covering the 96\% (2\(\sigma\)) range among different subboxes are also plotted.

Figure 7. The covariance, \(\text{Cov}(M_{\star,1}, M_{\star,2}; L_s)\), of the GSMF between two stellar masses, \(M_{\star,1}\) and \(M_{\star,2}\), as a function of the characteristic sample size \(L_s\). Symbols show results obtained from the mock sample. Different \(M_{\star,1}\) are represented by different symbols: \(M_{\star,1}[h^{-1}\text{M}_\odot] = 10^{9.1}, 10^{10.3},\) and \(10^{11.5}\), from bottom up, scaled by 0.01, 1, and 100, respectively, for clarity. Different \(M_{\star,2}\) with the same \(M_{\star,1}\) are rescaled by 0.3, 1, and 1.2, for \(M_{\star,2}[h^{-1}\text{M}_\odot] = 10^{9.1}, 10^{10.3},\) and \(10^{11.5}\), respectively. The solid curves are model predictions.
The model of cosmic variance compared with SDSS data. Upper panel: cosmic variance $\sigma_{CV}$ of the GSF as a function of the characteristic sample size $L_\ast$ for galaxies of different stellar masses, $M_\ast$, shown by different colors. Solid lines are the $\sigma_{CV}$ estimated from the SDSS sample, while dashed lines are predictions of the fitting model. Lower panel: the ratio of $\sigma_{CV}$ between the SDSS sample and the model. The black horizontal line indicates the ratio of 1.0.

To summarize, the simple model presented above provides a useful way to estimate the level of CV expected in measurements of the GSF. This variance, which is produced by the fluctuations of the cosmic density field, should be combined with the Poisson variance from number counting to estimate the total variance in the uncertainty of the GSF. This is particularly the case where the galaxy population is observed in a small volume and the CV is larger than the counting error. In real applications, other types of uncertainties, such as errors in photometry, redshift, and stellar mass estimate, should also be modeled properly along with the CV described here.

4.2. CVs in the SDSS Volume

In this subsection, we examine in detail the CV in the SDSS using the mock samples constructed for the SDSS. Here we only consider galaxies and model galaxies in the SDSS NGC (hereafter the SDSS sky coverage) with redshift $0.01 \leq z \leq 0.12$ (hereafter the SDSS volume). We construct four different types of samples:

(i) SDSS sample: SDSS DR7 observed galaxies in the SDSS volume, with $r$-band magnitude selection $r \leq 17.6$.
(ii) SDSS mock sample: model galaxies in the SDSS volume, with $r$-band magnitude selection $r \leq 17.6$.
(iii) SDSS magnitude-limited mock samples: model galaxies in the SDSS volume, that are brighter than a given magnitude limit.
(iv) SDSS volume-limited mock sample: all model galaxies in the SDSS volume. This sample served as the benchmark of the GSF, because it is almost free of CV, as compared with the entire simulation box.

To investigate potential CV in the SDSS volume, we first examine the galaxy number density, $n_\ast$, as a function of redshift $z$, in the SDSS volume-limited mock sample. Model galaxies in a given stellar mass bin are binned in redshift intervals with bin size $\Delta z = 0.005$, and the galaxy count in each bin is used to estimate the galaxy number density. The results are shown in Figure 9. The redshift distribution of model galaxies in the SDSS volume shows two peaks, one around $z_1 \approx 0.03$, due to the presence of the large-scale structure known as the CfA Great Wall (Geller & Huchra 1989), and the other around $z_2 \approx 0.075$, due to the presence of the Sloan Great Wall (Gott et al. 2005). Below $z \approx 0.03$, the number densities show a sharp decline as $z$ decreases, and the effect is stronger for massive galaxies, indicating the presence of a local void (see also, for example, Whitbourn & Shanks 2014, 2016). For comparison, we also show the redshift distribution of SDSS...
galaxies, obtained by using subsamples complete to given absolute magnitude limits. We see that the observed distribution follows that in the mock sample well, indicating our mock sample can be used to study the CV in the SDSS sample. For reference, we also plot the number densities of simulated dark matter halos in the SDSS volume versus redshift. Here again we see structures similar to those seen in the galaxy distribution. In particular, there is a marked decline of halo density at $z < 0.03$, and the decline is more prominent for more massive halos.

The presence of the local low-density region shown above can have strong impact on the statistical properties of the galaxy population derived from the SDSS, especially for faint galaxies, which can be observed only within the local volume in a magnitude-limited sample. Indeed, the measurement of the GSMF, which describes the number density of galaxies as a function of galaxy mass, can be biased if the local low-density region is not properly accounted for.

As an illustration, Figure 10 shows the GSMFs derived from SDSS magnitude-limited mock samples with different $r$-band magnitude limits, using the standard $V$-max method. For reference, we also plot the GSMF obtained from the SDSS volume-limited mock sample (the thick dashed line), which matches well the "global" GSMF obtained from the whole $500^3\, h^{-3}\text{Mpc}^3$ simulation box. As one can see, the GSMF can be significantly underestimated if the magnitude limit is shallow (corresponding to a low value of the $r$-band magnitude limit, $r_{\text{lim}}$). Only a sample as deep as $r_{\text{lim}} = 20$ can provide an unbiased estimate of the GSMF down to $M_* \approx 10^9\,h^{-1}\text{M}_\odot$. For the SDSS limit, $r_{\text{lim}} = 17.6$, the measurement starts to deviate from the global GSMF at $M_* \approx 10^8\,h^{-1}\text{M}_\odot$, and the difference between them reaches a factor of about 5 at around $10^8\,h^{-1}\text{M}_\odot$. In Appendix A, we also check possible variances produced by using Monte Carlo halo merger trees to extend the simulated trees. As one can see, the bias revealed by the SDSS mock sample for low-mass galaxies is not produced by the Monte Carlo merger trees used for low-mass halos.

The underestimation of the GSMF at the low-mass end is produced by the presence of a low-density region at $z < 0.03$ in the SDSS volume. To show this more clearly, we define a "break" mass, $M_{\text{break}}(r_{\text{lim}})$, so that galaxies with stellar masses $M_* > M_{\text{break}}(r_{\text{lim}})$ are complete to $z = 0.03$ for the given magnitude limit, $r_{\text{lim}}$. Here we have used the mean mass-to-light ratio, obtained from the mock sample, to convert the stellar mass to an absolute magnitude. As one can see, for each $r_{\text{lim}}$ the GSMF obtained from the sample starts to deviate from the global benchmark at $M_{0.03}(r_{\text{lim}})$, shown by the vertical line, and is substantially lower at $M_* < M_{0.03}$. All these demonstrate that the faint end of the GSMF can be underestimated significantly in the SDSS due to the presence of the local low-density region at $z < 0.03$.

4.3. The Correction of Cosmic Variance

4.3.1. Conventional Methods

The results described above indicate that CV is a serious issue in the measurements of the GSMF, even for a sample as large as the SDSS. Corrections have to be made in order to obtain an unbiased result that represents the true GSMF in the low-$z$ universe. In the literature, some estimators other than the standard V-max method have been proposed, such as the maximum likelihood method (e.g., Efstathiou 1988; Blanton et al. 2001; Cole 2011; Whitbourn & Shanks 2016) and scaling with bright galaxies (e.g., Baldry et al. 2012). These methods were designed, at least partly, to correct for the effects of large-scale structure in the measurements of the GSMF from an observational sample. Here we test their performances using our SDSS mock samples.

In the maximum likelihood method, one starts with an assumed functional form, either parametric or nonparametric, for the GSMF, and then use a maximum likelihood method to match the model prediction with the data, thereby obtaining the parameters that specify the functional form of the GSMF. In our analysis here, we choose a triple-Schechter function to model the GSMF,

$$
\Phi(M_*)d\log M_* = \sum_{k=1}^{3} \Phi_{*,k} \left( \frac{M_*}{\mu_i} \right)^{\alpha_i + 1} e^{-M_*/\mu_i}d\log M_* ,
$$

where $\Phi_{*,k}$, $\mu_i$, and $\alpha_i$ are the amplitude, the characteristic mass, and the faint-end slope of the $i$th Schechter component, respectively. This function is assumed to be defined over the
domain \([M_\text{*,min}, M_\text{*,max}]\). For a galaxy \(i\) with stellar mass \(M_\text*\) at redshift \(z_i\) in the sample, the probability for it to be observed at this redshift is

\[
\mathcal{L}_i = \frac{\Phi(M_i)}{\int_{M_\text{*,min}}^{M_\text{*,max}} \Phi(M_\text*) d \log M_\text*}.
\]

The total likelihood \(\mathcal{L}\) that the GSMF takes the assumed \(\Phi\) is then given by

\[
\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_i,
\]

where \(N\) is the number of galaxies in the sample. The model parameters can be adjusted so as to maximize the likelihood \(\mathcal{L}\). In our application to the SDSS mock sample, we fit the GSMF obtained from the \(V\)-max method with the triple-Schechter function and use the parameters as the initial input of the maximization process. Since the bright end is free of CV, we fix the three parameters characterizing the Schechter component at the brightest end, leaving the remaining six parameters to be constrained by the maximum likelihood process. As the maximum likelihood method does not provide information about the overall amplitude of \(\Phi(M_\text*)\), the bright end is also used to fix the amplitude of \(\Phi(M_\text*)\). The GSMF estimated in this way from the SDSS mock sample is plotted in Figure 11 as the green line, in comparison with that estimated with the \(V\)-max method (dashed line) and with the benchmark GSMF (black line). It is clear that the maximum likelihood method works better than the \(V\)-max method, but it still underestimates the GSMF at the low-mass end. The underlying assumption of the maximum likelihood method is that the relative distribution of galaxies with respect to \(M_\text*\) is everywhere the same. This is in general not true, given that galaxy clustering depends on \(M_\text*\). This explains the failure of this method in correcting the CV.

In an attempt to control the CV in the GAMA survey, Baldry et al. (2012) proposed using the number density of brighter galaxies estimated in a larger volume to scale the number density of fainter galaxies that are observed only in a smaller volume. This method will be referred to as the “density scaling” method. Our implementation of this method is as follows.

(i) Choose a “cosmic-variance-free (CVF)” sample, including only bright galaxies that have \(z_{\text{max}}\) larger than 0.12. In our SDSS mock sample, this corresponds to select galaxies with \(M_\text* > 3 \times 10^{10} h^{-1} M_\odot\). This sample will be used as the density tracer at different redshifts to scale the density at the fainter end.

(ii) Compute the cumulative number density of the CVF sample, \(n_{\text{CVF}}(<z)\), as a function of redshift \(z\). In practice, the cumulative number density is calculated in the redshift range [0.01, \(z\)].

(iii) Compute the GSMF, \(\Phi_{V_{\text{max}}}(M_\text*)\), using the \(V\)-max method.

(iv) For each stellar mass bin of \(\Phi_{V_{\text{max}}}(M_\text*)\), find the largest redshift, \(z_{\text{max}}(M_\text*)\), below which galaxies in this bin can be observed in the sample.

(v) Obtain the corrected GSMF, \(\Phi_\text{sc}\), by scaling the \(V\)-max estimate with a correction factor:

\[
\Phi_\text{sc}(M_\text*) = \Phi_{V_{\text{max}}}(M_\text*) \frac{n_{\text{CVF}}(<0.12)}{n_{\text{CVF}}(<z_{\text{max}}(M_\text*))},
\]

where \(n_{\text{CVF}}(<0.12)\) is the number density of the CVF sample in the full redshift range, [0.01, 0.12], and \(n_{\text{CVF}}(<z_{\text{max}}(M_\text*))\) is that in the redshift range [0.01, \(z_{\text{max}}(M_\text*\)]).

The GSMF estimated in this way from the SDSS mock sample is plotted in Figure 11 as the purple line. This method appears to work better than both the \(V\)-max method and the maximum likelihood method in the low-mass end, but the underestimation is still substantial. Furthermore, this method leads to a dip around \(M_\text* = 10^{10.8} h^{-1} M_\odot\), because of the density enhancement associated with the CfA Great Wall. The failure of this scaling method has an origin similar to that of the maximum likelihood method. The underlying assumption here is that the bright galaxies can serve as a tracer of the cosmic density field, and that the distributions of bright and faint galaxies are both related to the underlying density field by a similar bias factor. In general, this assumption is not valid.

### 4.3.2. Methods Based on the Joint Distribution of Galaxies and Environment

Since galaxies form and reside in the cosmic density field, the number density of galaxies is expected to depend on the local environment of galaxies. Suppose the local environment is specified by a quantity or a set of quantities \(E\). The joint distribution of galaxy mass and \(E\) obtained from a given sample, “\(S\),” can be written as

\[
\Phi_S(M_\text*, E) = \Phi_S(M_\text*|E) P_S(E),
\]

Figure 11. The GSMFs estimated from the SDSS mock catalog with different methods designed to account for the cosmic variance. The upper panel shows the GSMFs, and the lower panel shows the ratio of each GSMF to the benchmark. Black solid line: the benchmark GSMF obtained from the SDSS volume-limited mock sample. Black dashed line: the GSMF obtained with the \(V\)-max method, with error bars calculated using 100 bootstrap samples. Purple line: the GSMF obtained from the density scaling method of Baldry et al. (2012). Green line: the GSMF obtained from the maximum likelihood method assuming a triple-Schechter function form.

\[
\Phi(V_{\text{max}}) = \frac{n_{\text{CVF}}(<0.12)}{n_{\text{CVF}}(<z_{\text{max}}(M_\text*))}.
\]

where \(n_{\text{CVF}}(<0.12)\) is the number density of the CVF sample in the full redshift range, [0.01, 0.12], and \(n_{\text{CVF}}(<z_{\text{max}}(M_\text*))\) is that in the redshift range [0.01, \(z_{\text{max}}(M_\text*\)]).

The GSMF estimated in this way from the SDSS mock sample is plotted in Figure 11 as the purple line. This method appears to work better than both the \(V\)-max method and the maximum likelihood method in the low-mass end, but the underestimation is still substantial. Furthermore, this method leads to a dip around \(M_\text* = 10^{10.8} h^{-1} M_\odot\), because of the density enhancement associated with the CfA Great Wall. The failure of this scaling method has an origin similar to that of the maximum likelihood method. The underlying assumption here is that the bright galaxies can serve as a tracer of the cosmic density field, and that the distributions of bright and faint galaxies are both related to the underlying density field by a similar bias factor. In general, this assumption is not valid.
where $\Phi_S(M_h|E)$ is the conditional distribution of galaxy mass in a given environment estimated from sample “S,” and $P_S(E)$ is the probability distribution function of the environmental quantity given by the sample. If galaxy formation and evolution is a local process so that $\Phi(S|M_h|E)$ is independent of the galaxy sample, then the CV in the stellar mass function derived from the sample can all be attributed to the difference between $P_S(E)$ and the global distribution function, $P(E)$, expected from a large sample where the distribution of $E$ is sampled without bias. An unbiased estimate of the GSMF $\Phi(M_*)$ is then

$$\Phi(M_*) = \int \Phi_S(M_h|E)P(E)dE.$$  

Thus, the unbiased GSMF is obtained from the conditional distribution function, $\Phi_S(M_h|E)$, derived from the sample “S,” and the unbiased distribution $P(E)$ of environment variable.

The environmental quantity has to be chosen properly so that it can be estimated from observations, while the unbiased distribution function, $P(E)$, can, in principle, be obtained from large cosmological simulations. Here, we analyze a method that uses the masses of dark matter halos as the environmental quantity. In this case, $E$ is represented by halo mass, $M_h$, $\Phi_E(M_h|E)$ is the CGSMF, and $P(E) = n(M_h)$ is the halo mass function estimated directly from the constrained simulation (Yang et al. 2003). The advantage here is that the unbiased estimates are only needed for the conditional functions, $\Phi(M_h|E)$. The disadvantage is that it is model-dependent through $n(M_h)$, and that one has to identify galaxy systems to represent dark matter halos.

Figure 12 shows the conditional stellar mass functions of galaxies in halos of different masses, estimated from the SDSS mock sample, in comparison with the benchmarks obtained from the total SDSS volume-limited sample. As one can see, for a given halo mass, the CGSMF obtained from the SDSS mock sample matches the benchmark well only in the massive end. This happens because of the absence of massive halos at small distances in the local underdense region, so that their faint member galaxies are not observed in the magnitude-limited sample. The total GSMF, obtained using Equation (19), is shown in Figure 13 by the green line, in comparison to the benchmark of the total GSMF represented by the black solid line and to the GSMF obtained by the traditional $V$-max method represented by the black dotted line. Here the benchmark CGSMFs (dotted lines in Figure 12) are used for halos with $M_h < 10^{12} h^{-1} M_\odot$, while the CGSMFs estimated from the magnitude-limited sample (solid lines in Figure 12) are used for less massive halos. This is to mimic the fact that the total CGSMF for less massive halos can be obtained by other means (see Equation (21)), while the low-stellar-mass end of CGSMFs for massive halos cannot be obtained directly from the SDSS spectroscopic sample. Here again the stellar mass function at the low-mass end is underestimated, although the method works substantially better than the $V$-max method. The reason is clear from Figure 12. The stellar mass function at the low-mass end is only sampled by low-mass halos because of the absence of massive halos in the nearby volume, while the low-mass end in the benchmark stellar mass function is actually affected by the low-mass ends of the conditional stellar mass functions of massive halos.

These results demonstrate an important point. If the shape of the CGSMF depends significantly on halo mass, then one needs
to estimate all the conditional functions reliably down to a
given stellar mass limit, in order to get an unbiased estimate of
the total stellar mass function down to the same mass limit. The
SDSS redshift sample is clearly insufficient to achieve this goal
in the low-mass end.

In a recent paper, Lan et al. (2016) showed that the
conditional functions of galaxies can be estimated down to
$M_r \sim -14$ (corresponding to a stellar mass of about $10^8 M_\odot$)
for halos with mass $M_h > 10^{12} M_\odot$ by cross-correlating galaxy
groups (halos) selected from the SDSS spectroscopic sample
with SDSS photometric data. Thus, if we can estimate the
contribution by halos with lower masses to a similar
magnitude, then the total function can be obtained. Here we
test the feasibility of such an approach using the SDSS mock
sample. First, we obtain the CGSMFs down to a stellar mass of
$10^8 h^{-1} M_\odot$ for halos with $M_h \geq M_1 = 10^{12.5} h^{-1} M_\odot$ directly
from the total simulation volume. This step is to mimic the fact
that such CGSMFs can be obtained, as in Lan et al. (2016),
from observational data. The GSMF contributed by such halos is

$$\Phi_1(M_h) = \int_{M_1}^{\infty} \Phi(M_{\text{halo}} | M_h) n(M_h) dM_h. \tag{20}$$

To maximally reduce possible uncertainties introduced by this
procedure, we estimate the total CGSMF $\Phi_1$ for $M_h \geq M_1$
directly from a modified $V$-max method for the high-stellar-
mass end. Specifically, each galaxy is assigned a weight,
$n_{\text{halo,fi}}/n_{\text{halo}}(V_{\text{max}})$, the ratio between the number density of
$M_h \geq M_1$ halos in the universe and that in $V_{\text{max}}$. In practice, the
weighted $V$-max has little impact on the results, as the effect of
CV for high-mass galaxies is small. The procedure is included
only for maintaining consistency. Equation (20) is then used only
at the low-stellar-mass end, where the $V$-max method fails
because of incompleteness. The result for $\Phi_1$ obtained in this
way is shown by the purple solid curve in Figure 13.

To estimate the contribution by halos with $M_h < M_1$, in a
way that can be applied to real observation, we first eliminate
all galaxies that are contained in halos with $M_h \geq M_1$. For the
rest of the galaxies, we estimate the function by a modified version
of the $V$-max method,

$$\Phi_2(M_h) = \sum_{i=1}^{N_{\text{halo,fi}}} \frac{1}{V_{\text{max},i}} \frac{1}{1 + b^2(V_{\text{max},i})}, \tag{21}$$

where the summation is over individual galaxies, $b = 0.6$ is the
bias factor which is considered to be constant for low-mass
halos (e.g., Sheth et al. 2001), $\delta(V_{\text{max}}) = \rho(V_{\text{max}})/\rho_0 - 1$ is the
mean overdensity within $V_{\text{max}}, \rho_0$ is the universal mass density,
and $\overline{\rho}(V_{\text{max}})$ is the mean mass density within $V_{\text{max}}$. The function
$\Phi_2$ thus estimated is shown as the purple dashed curve in
Figure 13. Note that small groups can only be seen in the very
local region, so the CGSMF estimated for halos in a small mass
bin can be very noisy. Our method intends to avoid this
uncertainty by calculating the total CGSMF for all halos less
massive than $M_1$. The total GSMF, $\Phi = \Phi_1 + \Phi_2$, is shown by
the black dashed line in Figure 13, which is very close to the
benchmark, indicating that our method can indeed take care of
the bias produced by the local underdense region. We have
checked that the result depends only weakly on the choice of
the value of $M_1$.

![Figure 14. The galaxy luminosity function (GLF) estimated from the SDSS catalog using our method, in comparison to the results in the literature. The upper panel shows the GLFs, while the lower panel shows the ratio of each GLF to that obtained with the $V$-max method. The black solid line is the GLF obtained using our method. The gray solid line at the faint end (first two data points) is obtained by linear extrapolation. The black dashed line is the GLF using the $V$-max method. The gray shaded band indicates the cosmic variance of SDSS sample expected from Equation (10). The purple line is from Loveday et al. (2012) for the GAMA survey. The green line is from Whitbourn & Shanks (2016) using the SDSS "cmodel" magnitude.](image)

4.4. Applications to Observational Data

In this subsection, we apply the method described above to the
real SDSS sample. We first estimate the GLF using the
procedure based on the conditional distributions of galaxy
luminosity in dark matter halos, as described in Section 4.3.2.
Here, the CLF, $\Phi_1(M_{\text{halo}})$, for faint galaxies with magnitude
$M_r - 5 \log h > -17.2$ in halos more massive than $10^{12.5} h^{-1} M_\odot$ are obtained from Lan et al. (2016), while the
CLF for brighter galaxies in these halos is estimated directly from
the SDSS sample using the $V$-max method and the group
catalog of Yang et al. (2012; see also Yang et al. 2007).
For halos with masses below $10^{12.5} h^{-1} M_\odot$, the CLF, $\Phi_2(M_{\text{halo}})$, is
obtained from the SDSS sample using the modified $V$-max
method as described by Equation (21). The total GLF is then
obtained from $\Phi(M_h) = \Phi_1(M_h) + \Phi_2(M_h)$. Figure 14 shows
the result of the GLF obtained with the solid black line, in
comparison with that obtained from the traditional $V$-max
method. At the faint end, $M_r - 5 \log h \approx -15$, the GLF is
about twice as high as that given by the $V$-max method,
indicating that the CV can have a large impact on the estimate
of the GLF at the faint end. To show this more clearly, we plot
the CV expected from Equation (10) as the shaded band in
Figure 14, where the stellar mass is obtained from the
luminosity by using the mean mass-to-light ratio. The expected
CV is quite large at the faint end, indicating that the CV is an
important issue in estimating the faint end of the GLF. The
GLF in the local universe has been estimated by many
authors using various samples (e.g., Blanton et al. 2003;
Jones et al. 2006; Yang et al. 2009; Driver et al. 2012; Loveday et al. 2012; Whitbourn & Shanks 2016). For comparison, we plot the GLFs obtained by Loveday et al. (2012) from the GAMA survey and by Whitbourn & Shanks (2016), who applied a maximum likelihood method to the SDSS to account for the CV in their estimates. The result of Whitbourn & Shanks (2016) matches ours over a wide range of luminosity, but seems to still underestimate the GLF at the faint end. The result of Loveday et al. (2012) has a large discrepancy with our result, possibly due to the CV in the small sky coverage of the GAMA sample used, which is 144 deg$^2$, and the depth, $z = 0.1$, they used to define their local sample. Since many of the faint galaxies in the SDSS photometric data do not have reliable stellar mass estimates, CGSMFs are not available at the low-mass end. Because of this, we cannot estimate the GSMF down to the low-mass end directly from the data with the method above. As an alternative, we use the $M_\star - M_r$ relation obtained from the SDSS spectroscopic sample to convert the GLF obtained above to estimate a GSMF. We do this through the following steps. (i) Construct a large volume-limited Monte Carlo sample of galaxies with absolute magnitude distribution given by the GLF. (ii) Bin these galaxies according to their absolute magnitudes. (iii) For each Monte Carlo galaxy, we randomly choose a galaxy in the real SDSS spectroscopic sample in the same absolute magnitude bin and assign the stellar mass of the real galaxy to the Monte Carlo galaxy. (iv) Compute the GSMF of this volume-limited Monte Carlo sample.

The GSMF obtained directly from the GLF in this way is shown in Figure 15 by the black solid curve. Since the GLF is estimated only down to $M_r - 5 \log h = -15$, the first two data points in the low-mass end of the GSMF may be underestimated, as galaxies fainter than $M_r - 5 \log h = -15$ may contribute to these two stellar mass bins. To test this, we extrapolate the faint end of the GLF to $M_r - 5 \log h = -14.2$, which is sufficient to include all galaxies with stellar masses down to $10^8 h^{-1} M_\odot$. This extrapolation is shown by the gray extension of the black solid curve in Figure 14. The GSMF obtained from the extended GLF is shown by the gray line in Figure 15. As one can see, the extension of the GLF only slightly increases the GSMF at the lowest mass. The GSMF estimated in this way is compared with that estimated with the conventional V-max method. The gray shaded band shows the expected CV given by Equation (10) for the SDSS sample. The effect of CV is quite large at the low-mass end. The difference between our result and that obtained from the V-max method is even larger, indicating again that the local SDSS region is an unusually underdense region. The GSMF in the low-$z$ universe has been estimated in numerous earlier investigations using different samples and methods (e.g., Li & White 2009; Yang et al. 2009; Baldry et al. 2012; Bernardi et al. 2013; He et al. 2013; D’Souza et al. 2015). Several of the earlier results are plotted in Figure 15 for comparison. The result of Li & White (2009), who measured the GSMF of the SDSS sample directly from the stellar masses estimated by Blanton & Roweis (2007) with a Chabrier IMF (Chabrier 2003) and corrections for dust, matches well our V-max result and also misses the steepening of the GSMF at $M_\star < 10^{9.5} h^{-1} M_\odot$. Our measurement at $M_\star > 10^{10.5} h^{-1} M_\odot$ is significantly lower than that from Bernardi et al. (2013), because they included the light in the outer parts of massive galaxies that may have been missed in the SDSS NYU-VAGC used here (see also He et al. 2013 for a discussion of this effect). Such corrections do not affect the GSMF in the low-mass end. The overall shape of our GSMF is similar to that of Baldry et al. (2012), which was obtained from the GAMA sample, but the amplitude of their function is about 50% lower.

GAMA has a small sky coverage, 144 deg$^2$, although it is deeper, to $r = 19.8$. Baldry et al. (2012) obtained their GSMF using the local $z \leq 0.06$ sample. According to our test with mock samples of similar sky coverage and depth, the CV in the GSMF estimated from such a sample can be very large. The lower amplitude given by the GAMA sample may be produced by such CV.

In Appendix B, we check the effects of possible stochastic factors in the estimates of stellar masses of galaxies in our empirical model. As one can see, such factors can introduce some uncertainties in the GSMF at the low-mass-end uncertainty, but the effects are much smaller than the CV.

In conclusion, when CV is carefully taken into account, the low-mass-end slope of the GSMF in the low-$z$ universe, which is about $-1.5$ as indicated by the red dashed line in Figure 15, is significantly steeper than that published in earlier studies. In particular, there is a significant upturn at $M_\star < 10^{9.5} h^{-1} M_\odot$ in the GSMF that is missed in many of the earlier measurements. For reference, we list the GLF and GSMF estimated with our method in Table 1.
Table 1

| \( M_\odot - 5 \log h \) | \(-15.2\) | \(-15.6\) | \(-16.0\) | \(-16.4\) | \(-16.8\) | \(-17.2\) | \(-17.6\) | \(-18.0\) | \(-18.4\) | \(-18.8\) | \(-19.2\) | \(-19.6\) | \(-20.0\) | \(-20.4\) | \(-20.8\) | \(-21.2\) | \(-21.6\) | \(-22.0\) | \(-22.4\) | \(-22.8\) |
|-----------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| \( \log \Phi(M_\odot) \) \( [h^2 \text{Mpc}^{-3} \text{mag}^{-1}] \) | \(-1.179_{-0.043}^{+0.041}\) | \(-1.290_{-0.027}^{+0.022}\) | \(-1.395_{-0.023}^{+0.022}\) | \(-1.500_{-0.023}^{+0.022}\) | \(-1.536_{-0.015}^{+0.015}\) | \(-1.571_{-0.011}^{+0.011}\) | \(-1.659_{-0.009}^{+0.009}\) | \(-1.730_{-0.006}^{+0.006}\) | \(-1.802_{-0.006}^{+0.006}\) | \(-1.864_{-0.005}^{+0.005}\) | \(-1.919_{-0.004}^{+0.004}\) | \(-1.960_{-0.003}^{+0.003}\) | \(-2.019_{-0.002}^{+0.002}\) | \(-2.111_{-0.002}^{+0.002}\) | \(-2.244_{-0.002}^{+0.002}\) | \(-2.449_{-0.003}^{+0.003}\) | \(-2.743_{-0.004}^{+0.004}\) | \(-3.166_{-0.006}^{+0.006}\) | \(-3.770_{-0.011}^{+0.011}\) | \(-4.501_{-0.023}^{+0.023}\) |
| \( \log M_\odot \) \( [h^{-1} M_\odot] \) | 8.1 | 8.3 | 8.5 | 8.7 | 8.9 | 9.1 | 9.3 | 9.5 | 9.7 | 9.9 | 10.1 | 10.3 | 10.5 | 10.7 | 10.9 | 11.1 | 11.3 | 11.5 | 11.7 | 11.9 |
| \( \log \Phi(M_\odot) \) \( [h^2 \text{Mpc}^{-3} \text{dex}^{-1}] \) | \(-1.005_{-0.039}^{+0.041}\) | \(-1.111_{-0.039}^{+0.041}\) | \(-1.184_{-0.039}^{+0.041}\) | \(-1.272_{-0.039}^{+0.041}\) | \(-1.345_{-0.039}^{+0.041}\) | \(-1.417_{-0.039}^{+0.041}\) | \(-1.518_{-0.039}^{+0.041}\) | \(-1.576_{-0.039}^{+0.041}\) | \(-1.624_{-0.039}^{+0.041}\) | \(-1.678_{-0.039}^{+0.041}\) | \(-1.688_{-0.039}^{+0.041}\) | \(-1.770_{-0.039}^{+0.041}\) | \(-1.903_{-0.039}^{+0.041}\) | \(-2.163_{-0.039}^{+0.041}\) | \(-2.562_{-0.039}^{+0.041}\) | \(-3.053_{-0.039}^{+0.041}\) | \(-3.770_{-0.039}^{+0.041}\) | \(-4.501_{-0.039}^{+0.041}\) |

Note. Galaxy brighter than \( M_\odot - 5 \log h = -15 \) is not sufficient to calculate the GSMF down to \( M_\odot = 10^{10.1} h^{-1} M_\odot \). Extrapolation of the GLF is used to solve this. The left column of \( \Phi(M_\odot) \) is without extrapolation, while the right column is with extrapolation.

5. Summary and Discussion

In this paper, we use the ELUCID simulation, a constrained N-body simulation in the SDSS volume to study galaxy distribution in the low-\( z \) universe. Our main results can be summarized as follows:

(i) Dark matter halos are selected from different snapshots of the simulation, and halo merger trees are constructed from simulated halos down to a halo mass of \( \sim 10^{10} h^{-1} M_\odot \). A method is developed to extend all the simulated halo merger trees to a mass resolution of \( 10^8 h^{-1} M_\odot \), which is needed to model galaxies down to a stellar mass of \( 10^8 h^{-1} M_\odot \).

(ii) The merger trees are used to populate simulated dark matter halos with galaxies according to an empirical model of galaxy formation developed by Lu et al. (2014a, 2015a). The model galaxies follow the real galaxies in the SDSS volume both in spatial distribution and in intrinsic properties. The catalog of model galaxies, therefore, provides a unique way to study galaxy formation and evolution in the cosmic web in the low-\( z \) universe.

(iii) Mock catalogs in the SDSS sky coverage are constructed, which can be used to investigate the distribution of galaxies as measured from the real SDSS data and its relation to the global distribution expected from a fair sample of galaxies in the low-\( z \) universe. These mock catalogs can thus be used to quantify the CVs in the statistical properties of the low-\( z \) galaxy population estimated from a survey like SDSS.

(iv) As an example, we use the mock catalogs so constructed to quantify the CV in the GSMF. Useful fitting formulae are obtained to describe the CV and covariance matrix of the GSMF as functions of stellar mass and sample volume.

(v) We find that the GSMF estimated from the SDSS magnitude-limited sample can be affected significantly by the presence of the underdense region at \( z < 0.03 \), so that the low-mass end of the function can be underestimated significantly.

(vi) We test several existing methods that are designed to deal with the effects of CV in the estimate of the GSMF, and find that none of them is able to fully account for the CV effects.

(vii) We propose and test a method based on the conditional stellar mass functions in dark matter halos, which is found to provide an unbiased estimate of the global GSMF.

(viii) We apply the method to SDSS data and find that the GSMF has a significant upturn at \( M_\odot < 10^{9.5} h^{-1} M_\odot \), which was missed in many earlier measurements of the local GSMF.

Our GSMF results have important implications for galaxy formation and evolution. The presence of an upturn in the GSMF at \( M_\odot < 10^{9.5} h^{-1} M_\odot \) suggests that there is a characteristic mass scale, \( \sim 10^{10} h^{-1} M_\odot \), corresponding to a halo mass of \( \sim 10^9 h^{-1} M_\odot \) (e.g., Lim et al. 2017a), below which star formation may be affected by processes that are different from those in galaxies of higher masses. The stellar mass function of galaxies at low \( z \) has been widely used to calibrate numerical simulations and semianalytic models of galaxy formation. The improved estimate of the GSMF presented here will clearly provide more accurate constraints on theoretical models.

The mock catalogs constructed here have other applications. For example, they can be used to analyze the CV in the measurements of other statistical properties of the galaxy population, such as the correlation functions (Zehavi et al. 2005; Wang et al. 2007) and peculiar velocities (e.g., Jing et al. 1998; Loveday et al. 2018) of galaxies of different luminosities/masses. Because of the presence of local large-scale structures, such as the underdense region at \( z < 0.03 \), the measurements for faint galaxies can be affected. A comparison between the results obtained from the mock sample and those from the benchmark sample can then be used to quantify the effects of CV. Another application is to Hi samples of galaxies. Current Hi surveys, such as HIPASS (Meyer et al. 2004) and ALFALFA (Giovanelli et al. 2005), are shallow, typically to \( z \sim 0.05 \), and so the Hi mass functions and correlation functions estimated from these surveys can be affected significantly by the CV in the nearby universe (e.g., Guo et al. 2017). The same method as described here can be used to construct mock catalogs for Hi galaxies and to quantify CVs in these measurements. We will come back to some of these problems in the future.

This work is supported by the National Key R&D Program of China (grant Nos. 2018YFA0404503 and 2018YFA0404502), the National Key Basic Research Program of China (grant Nos. 2015CB857002 and 2015CB857004), and the National Science Foundation of China (grant Nos. 11233005, 11621303, 11522324, 11421303, 11503065, 11673015, 11733004, and 11320101002). H.J.M. acknowledges the support from NSF AST-1517528.

Appendix A

Convergence Test of the Monte Carlo Tree

In Section 2.2, we introduced the method to extend the resolution of the simulated halo merger trees. The extended trees are consistent with the trees obtained using other methods,
as shown in Figure 1. Here we carry out another test to demonstrate the convergence of our method, by examining not only the dark halo merger trees but also the predicted galaxy population. We follow the steps described in Section 2.2 but only the dark halo merger trees but also the predicted galaxy histories of real galaxies. For low-mass galaxies, this may model predictions may not be identical to the formation model based on the assembly histories of halos. However, the uncertainty of the GSMF produced by the uncertainty in the scatter in the mass relation on the predicted GSMF. The red line is the original GSMF of the SDSS sample. The blue line is the result obtained by including the scatter in the $M_{	ext{subhalo}}-M_*$ relation, with the error bars indicating the variance among 100 realizations of the relation.

**Appendix B**

**Scatter in the Halo Mass–Stellar Mass Relation**

In Section 3.1, we introduce the empirical galaxy formation model based on the assembly histories of halos. However, the model predictions may not be identical to the formation histories of real galaxies. For low-mass galaxies, this may become important because the uncertainty in the prediction of the final stellar mass of a galaxy may be mixed with the CV that we see, thereby affecting our interpretation of the CV. Here, we test potential uncertainties in the stellar mass predictions, taking into account the scatter in the relation between galaxy mass, $M_*$, and the host subhalo mass, $M_{	ext{subhalo}}$.

We follow Moster et al. (2010); the dashed line in their Figure 6 and a variance of $\sigma_m = 0.15$ dex to model the the mean value and scatter for the $M_{	ext{subhalo}}-M_*$ relation. We first assign each galaxy in the SDSS sample a subhalo mass using its stellar mass together with the mean $M_{	ext{subhalo}}-M_*$ relation. We then convert the subhalo mass back to the stellar mass using the inverse of the $M_{	ext{subhalo}}-M_*$ relation together with the scatter. Finally, we use the newly assigned stellar masses to calculate the GSMF of the SDSS sample. This process is repeated 100 times, and the variance among them gives an estimate of the uncertainty of the GSMF produced by the uncertainty in the $M_{	ext{subhalo}}-M_*$ relation. Figure 17 shows the comparison of the GSMFs using the original stellar masses and the newly assigned stellar masses. There are some differences at the high-mass end produced by the stochastic factor introduced. The uncertainties at the low-mass end produced by the scatter in the $M_{	ext{subhalo}}-M_*$ relation are, however, much smaller than the CV, indicating that our results are not affected significantly by the scatter.

**ORCID iDs**

Yangyao Chen @ https://orcid.org/0000-0002-4597-5798
H. J. Mo @ https://orcid.org/0000-0002-9665-5380
Huiyuan Wang @ https://orcid.org/0000-0002-4911-6990
Xiaohu Yang @ https://orcid.org/0000-0003-3997-4606
Youcai Zhang @ https://orcid.org/0000-0003-1967-4091

**References**

Babul, A., & Rees, M. J. 1992, MNRAS, 255, 346
Baldry, I. K., Driver, S. P., Loveday, J., et al. 2012, MNRAS, 421, 621
Bernardi, M., Meert, A., Sheth, R. K., et al. 2013, MNRAS, 436, 697
Blanton, M. R., Dalal, C., Eisenstein, D., et al. 2001, AJ, 121, 2358
Blanton, M. R., Hogg, D. W., Bahcall, N. A., et al. 2003, ApJ, 592, 819
Blanton, M. R., & Roweis, S. 2007, AJ, 133, 734
Boylan-Kolchin, M., Ma, C.-P., & Quataert, E. 2008, MNRAS, 383, 93
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Chabrier, G. 2003, PASP, 115, 763
Cole, S. 2011, MNRAS, 416, 739
Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371
Driver, S. P., & Robotham, A. S. G. 2010, MNRAS, 407, 2131
Driver, S. P., Robotham, A. S. G., Kelvin, L., et al. 2012, MNRAS, 427, 3244
D’Souza, R., Vegetti, S., & Kauffmann, G. 2015, MNRAS, 454, 4027
Dunkley, J., Komatsu, E., Nolta, M. R., et al. 2009, ApJS, 180, 306
Efstathiou, G. 1988, MNRAS, 232, 431
Gallazzi, A., Charlot, S., Brinchmann, J., White, S. D. M., & Tremonti, C. A. 2005, MNRAS, 362, 41
Geller, M. J., & Huchra, J. P. 1989, Sci, 246, 897
Giovanelli, R., Haynes, M. P., Kent, B. R., et al. 2005, AJ, 130, 2598
Gott, J. R. III, Jurić, M., Schlegel, D., et al. 2005, ApJ, 624, 463
Guo, H., Li, C., Zheng, Z., et al. 2017, ApJ, 846, 61
He, Y. Q., Xia, X. Y., Hao, C. N., et al. 2013, ApJ, 773, 37
Jha, S., Riess, A. G., & Kirshner, R. P. 2007, ApJ, 659, 122
Jiang, F., & van den Bosch, F. C. 2014, MNRAS, 440, 193
