Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **469,** 2791–2805 (2017)
Advance Access publication 2017 April 17

doi:10.1093/mnras/stx894

Approximate Bayesian computation in large-scale structure: constraining the galaxy–halo connection

ChangHoon Hahn,[1★†] Mohammadjavad Vakili,[1★†] Kilian Walsh,[1] Andrew P. Hearin,[2] David W. Hogg[1,3,4,5] and Duncan Campbell[6]

[1]*Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA*
[2]*Yale Center for Astronomy and Astrophysics, Yale University, New Haven, CT 06520, USA*
[3]*Flatiron institute, 160 Fifth Avenue, New York, NY 10010, USA*
[4]*Center for Data Science, New York University, 60 Fifth Ave, New York, NY 10011, USA*
[5]*Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany*
[6]*Department of Astronomy, Yale University, New Haven, CT 06511, USA*

Accepted 2017 April 10. Received 2017 March 14; in original form 2016 July 5

ABSTRACT

Standard approaches to Bayesian parameter inference in large-scale structure assume a Gaussian functional form (chi-squared form) for the likelihood. This assumption, in detail, cannot be correct. Likelihood free inferences such as approximate Bayesian computation (ABC) relax these restrictions and make inference possible without making any assumptions on the likelihood. Instead ABC relies on a forward generative model of the data and a metric for measuring the distance between the model and data. In this work, we demonstrate that ABC is feasible for LSS parameter inference by using it to constrain parameters of the halo occupation distribution (HOD) model for populating dark matter haloes with galaxies. Using specific implementation of ABC supplemented with population Monte Carlo importance sampling, a generative forward model using HOD and a distance metric based on galaxy number density, two-point correlation function and galaxy group multiplicity function, we constrain the HOD parameters of mock observation generated from selected 'true' HOD parameters. The parameter constraints we obtain from ABC are consistent with the 'true' HOD parameters, demonstrating that ABC can be reliably used for parameter inference in LSS. Furthermore, we compare our ABC constraints to constraints we obtain using a pseudo-likelihood function of Gaussian form with MCMC and find consistent HOD parameter constraints. Ultimately, our results suggest that ABC can and should be applied in parameter inference for LSS analyses.

Key words: methods: data analysis – methods: statistical – galaxies: haloes – dark matter – large-scale structure of Universe.

1 INTRODUCTION

Cosmology was revolutionized in the 1990s with the introduction of likelihoods – probabilities for the data given the theoretical model – for combining data from different surveys and performing principled inferences of the cosmological parameters (White & Scott 1996; Riess et al. 1998). Nowhere has this been more true than in cosmic microwave background (CMB) studies, where it is nearly possible to analytically evaluate a likelihood function that involves no (or minimal) approximations (Oh, Spergel & Hinshaw 1999; Eriksen et al. 2004; Wandelt, Larson & Lakshminarayanan 2004; Planck Collaboration XVI 2014; Planck Collaboration XIII 2016).

Fundamentally, the tractability of likelihood functions in cosmology flows from the fact that the initial conditions are exceedingly close to Gaussian in form (Planck Collaboration XVII 2016; Planck Collaboration XX 2016) and that many sources of measurement noise are also Gaussian (Knox 1995; Leach et al. 2008). Likelihood functions are easier to write down and evaluate when things are closer to Gaussian, so at large scales and in the early universe. Hence, likelihood analyses are ideally suitable for CMB data.

In large-scale structure (LSS) with galaxies, quasars and quasar absorption systems as tracers, formed through non-linear gravitational evolution and biasing, the likelihood *cannot* be Gaussian. Even if the initial conditions are perfectly Gaussian, the growth of structure creates non-linearities that are non-Gaussian (see Bernardeau et al. 2002 for a comprehensive review). Galaxies form within the density field in some complex manner that is modelled only effectively (Dressler 1980; Kaiser 1984; Santiago &

★ E-mail: chh327@nyu.edu (CHH); mjvakili@nyu.edu (MV)
† These authors have contributed equally to the paper.

Strauss 1992; Steidel et al. 1998; see Somerville & Davé 2015 for a recent review). Even if the galaxies were a Poisson sampling of the density field, which they are not (Mo & White 1996; Somerville et al. 2001; Casas-Miranda et al. 2002), it would be tremendously difficult to write down even an approximate likelihood function (Ata, Kitaura & Müller 2015).

The standard approach makes the strong assumption that the likelihood function for the data can be approximated by a pseudo-likelihood function that is a Gaussian probability density in the space of the two-point correlation function estimate. It is also typically limited to (density and) two-point correlation function (2PCF) measurements, assuming that these measurements constitute sufficient statistics for the cosmological parameters. As Hogg (in preparation) demonstrates, the assumption of a Gaussian pseudo-likelihood function cannot be correct (in detail) at any scale, since a correlation function, being related to the variance of a continuous field, must satisfy non-trivial positive-definiteness requirements. These requirements truncate function space such that the likelihood in that function space could never be Gaussian. The failure of this assumption becomes more relevant as the correlation function becomes better measured, so it is particularly critical on intermediate scales, where neither shot noise nor cosmic variance significantly influence the measurement.

Fortunately, these assumptions are not required for cosmological inferences, because high-precision cosmological simulations can be used to directly calculate LSS observables. Therefore, we can simulate not just the one- or two-point statistics of the galaxies but also any higher order statistics that might provide additional constraining power on a model. In principle, there is therefore no strict need to rely on these common but specious analysis assumptions as it is possible to calculate a likelihood function directly from simulation outputs.

Of course, any naive approach to sufficiently simulating the data would be ruinously expensive. Fortunately, there are principled, (relatively) efficient methods for minimizing computation and delivering correct posterior inferences, using only a data simulator and some choices about statistics. In this work, we use approximate Bayesian computation – ABC – which provides a *rejection sampling* framework (Pritchard et al. 1999) that relaxes the assumptions of the traditional approach.

ABC approximates the posterior probability distribution function (model given the data) by drawing proposals from the prior over the model parameters, simulating the data from the proposals using a forward generative model, and then rejecting the proposals that are beyond a certain threshold 'distance' from the data, based on summary statistics of the data. In practice, ABC is used in conjunction with a more efficient sampling operation like Population Monte Carlo (PMC; Del Moral, Doucet & Jasra 2006). PMC initially rejects the proposals from the prior with a relatively large 'distance' threshold. In subsequent steps, the threshold is updated adaptively, and samples from the proposals that have passed the previous iteration are subjected to the new, more stringent, threshold criterion (Beaumont et al. 2009). In principle, the distance metric can be any positive definite function that compares various summary statistics between the data and the simulation.

In the context of astronomy, this approach has been used in a wide range of topics including image simulation calibration for wide field surveys (Akeret et al. 2015), the study of the morphological properties of galaxies at high redshifts (Cameron & Pettitt 2012), stellar initial mass function modelling (Cisewski et al., in preparation) and cosmological inference with with weak-lensing peak counts (Lin & Kilbinger 2015; Lin, Kilbinger & Pires 2016), Type Ia Supernovae (Weyant, Schafer & Wood-Vasey 2013) and galaxy cluster number counts (Ishida et al. 2015).

In order to demonstrate that ABC can be tractably applied to parameter estimation in contemporary LSS analyses, we narrow our focus to inferring the parameters of a halo occupation distribution (HOD) model. The foundation of HOD predictions is the halo model of LSS, that is, collapsed dark matter haloes are biased tracers of the underlying cosmic density field (Press & Schechter 1974; Bond et al. 1991; Cooray & Sheth 2002). The HOD specifies how the dark matter haloes are populated with galaxies by modelling the probability that a given halo hosts N galaxies subject to some observational selection criteria (Lemson & Kauffmann 1999; Seljak 2000; Scoccimarro et al. 2001; Berlind & Weinberg 2002; Zheng et al. 2005). This statistical prescription for connecting galaxies to haloes has been remarkably successful in reproducing the galaxy clustering, galaxy–galaxy lensing and other observational statistics (Miyatake et al. 2015; Rodríguez-Torres et al. 2016), and is a useful framework for constraining cosmological parameters (van den Bosch, Mo & Yang 2003; Tinker et al. 2005; Cacciato et al. 2013; More et al. 2013) as well as galaxy evolution models (Conroy & Wechsler 2009; Tinker, Wetzel & Conroy 2011; Leauthaud et al. 2012; Behroozi, Wechsler & Conroy 2013b; Tinker et al. 2013; Walsh & Tinker, in preparation).

More specifically, we limit our scope to a likelihood analysis of HOD model parameter space, keeping cosmology fixed. We forward model galaxy survey data by populating pre-built dark matter halo catalogues obtained from high-resolution N-body simulations (Klypin, Trujillo-Gomez & Primack 2011; Riebe et al. 2013) using HALOTOOLS[1] (Hearin et al. 2016a), an open-source package for modelling the galaxy–halo connection. Equipped with the forward model, we use summary statistics such as number density, two-point correlation function, galaxy group multiplicity function (GMF) to infer HOD parameters using ABC.

In Section 2, we discuss the algorithm of the ABC-PMC prescription we use in our analyses. This includes the sampling method itself, the HOD forward model and the computation of summary statistics. Then in Section 3.1, we discuss the mock galaxy catalogue, which we treat as observation. With the specific choices of ABC-PMC ingredients, which we describe in Section 3.2, in Section 3.3, we present the results of our parameter inference using two sets of summary statistics, number density and 2PCF and number density and GMF. We also include in our results, analogous parameter constraints from the standard MCMC approach, which we compare to ABC results in detail, Section 3.4. Finally, we discuss and conclude in Section 4.

2 METHODS

2.1 Approximate Bayesian computation

ABC is based on rejection sampling, so we begin this section with a brief overview of rejection sampling. Broadly speaking, rejection sampling is a Monte Carlo method used to draw samples from a probability distribution, $f(\alpha)$, which is difficult to directly sample. The strategy is to draw samples from an instrumental distribution $g(\alpha)$ that satisfies the condition $f(\alpha) < Mg(\alpha)$ for all α, where $M > 1$ is some scalar multiplier. The purpose of the instrumental distribution $g(\alpha)$ is that it is easier to sample than $f(\alpha)$ (see Bishop 2007 and references therein).

[1] http://halotools.readthedocs.org

In the context of simulation-based inference, the ultimate goal is to sample from the joint probability of a simulation X and parameters $\boldsymbol{\theta}$ given observed data D, the posterior probability distribution. From Bayesian rule, this posterior distribution can be written as

$$p(\boldsymbol{\theta}, X|D) = \frac{p(D|X)p(X|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathcal{Z}}, \tag{1}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution over the parameters of interest and $\mathcal{Z}$ is the evidence,

$$\mathcal{Z} = \int \mathrm{d}\boldsymbol{\theta}\ \mathrm{d}X\ p(D|X)p(X|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{2}$$

where the domain of the integral is all possible values of X and $\boldsymbol{\theta}$. Since $p(\boldsymbol{\theta}, X|D)$ cannot be directly sampled, we use rejection sampling with instrumental distribution

$$q(\boldsymbol{\theta}, X) = p(X|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{3}$$

and the choice of

$$M = \frac{\max\ p(D|X)}{\mathcal{Z}} > 1. \tag{4}$$

Note that we do not ever need to know $\mathcal{Z}$. The choices of $q(\boldsymbol{\theta}, X)$ and M satisfy the condition

$$p(\boldsymbol{\theta}, X|D) < Mq(\boldsymbol{\theta}, X), \tag{5}$$

so we can sample $p(\boldsymbol{\theta}, X|D)$ by drawing $\boldsymbol{\theta}, X$ from $q(\boldsymbol{\theta}, X)$. In practice, this is done by first drawing $\boldsymbol{\theta}$ from the prior $\pi(\boldsymbol{\theta})$ and then generating a simulation $X = f(\boldsymbol{\theta})$ via the forward model. Then $\boldsymbol{\theta}, X$ is accepted if

$$\frac{p(\boldsymbol{\theta}, X|D)}{Mq(\boldsymbol{\theta}, X)} = \frac{p(D|X)}{\max\ p(D|X)} > u, \tag{6}$$

where u is drawn from `Uniform`$[0, 1]$. By repeating this rejection sampling process, we sample the distribution $p(\boldsymbol{\theta}, X|D)$ with the set of $\boldsymbol{\theta}$ and X that are accepted.

At this stage, ABC distinguishes itself by postulating that $p(D|X)$, the probability of observing data D given simulation X (*not* the likelihood), is proportional to the probability of the distance between the data and the simulation X being less than an arbitrarily small threshold ϵ

$$p(D|X) \propto p(\rho(D, X) < \epsilon), \tag{7}$$

where $\rho(D, X)$ is the distance between the data D and the simulation X. Equation (7) along with the rejection sampling acceptance criteria (equation 6) leads to the acceptance criteria for ABC: $\boldsymbol{\theta}$ is accepted if $\rho(D, X) < \epsilon$.

The distance function is a positive definite function that measures the closeness of the data and the simulation. The distance can be a vector with multiple components where each component is a distance between a single summary statistic of the data and that of the simulation. In that case, the threshold ϵ in equation (7) will also be a vector with the same dimensions. $\boldsymbol{\theta}$ is accepted if the distance vector is less than the threshold vector for every component.

The ABC procedure begins, in the same fashion as rejection sampling, by drawing $\boldsymbol{\theta}$ from the prior distribution $\pi(\boldsymbol{\theta})$. The simulation is generated from $\boldsymbol{\theta}$ using the forward model, $X = f(\boldsymbol{\theta})$. Then the distance between the data and simulation, $\boldsymbol{\rho}(D, X)$, is calculated and compared to $\boldsymbol{\epsilon}$. If $\boldsymbol{\rho}(D, X) < \boldsymbol{\epsilon}$, $\boldsymbol{\theta}$ is accepted. This process is repeated until we are left with a sample of $\boldsymbol{\theta}$ that all satisfy the distance criteria. This final ensemble approximates the posterior probability distribution $p(\boldsymbol{\theta}, X|D)$.

As it is stated, the ABC method poses some practical challenges. If the threshold ϵ is arbitrarily large, the algorithm essentially samples from the prior $\pi(\boldsymbol{\theta})$. Therefore, a sufficiently small threshold is necessary to sample from the posterior probability distribution. However, an appropriate value for the threshold is not known a priori. Yet, even if an appropriate threshold is selected, a small threshold requires the entire process to be repeated for many draws of $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta})$ until a sufficient sample is acquired. This often presents computation challenges.

We overcome some of the challenges posed by the above ABC method by using a population Monte Carlo (PMC) algorithm as our sampling technique. PMC is an iterative method that performs rejection sampling over a sequence of $\boldsymbol{\theta}$ distributions ($\{p_1(\boldsymbol{\theta}), ..., p_T(\boldsymbol{\theta})\}$ for T iterations), with a distance threshold that decreases at each iteration of the sequence.

Algorithm 1 The procedure for ABC-PMC

if $t = 1$: **then**
  for $i = 1, ..., N$ **do**
    // This loop can now be done in parallel for all i
    while $\rho(X, D) > \epsilon_t$ **do**
      $\theta_t^* \leftarrow \pi(\theta)$
      $X = f(\theta_t^*)$
    end while
    $\theta_t^{(i)} \leftarrow \theta_t^*$
    $w_t^{(i)} \leftarrow 1/N$
  end for
end if
if $t = 2, ..., T$: **then**
  for $i = 1, ..., N$ **do**
    // This loop can now be done in parallel for all i
    while $\rho(X, D) > \epsilon_t$ **do**
      Draw θ_t^* from $\{\theta_{t-1}\}$ with probabilities $\{w_{t-1}\}$
      $\theta_t^* \leftarrow K(\theta_t^*, .)$
      $X = f(\theta_t^*)$
    end while
    $\theta_t^{(i)} \leftarrow \theta_t^*$
    $w_t^{(i)} \leftarrow \pi(\theta_t^{(i)}) / \left(\sum_{j=1}^{N} w_{t-1}^{(i)} K(\theta_{t-1}^{(j)}, \theta_t^{(i)}) \right)$
  end for
end if

As illustrated in Algorithm 1, for the first iteration $t = 1$, we begin with an arbitrarily large distance threshold ϵ_1. We draw $\boldsymbol{\theta}$ (hereafter referred to as particles) from the prior distribution $\pi(\boldsymbol{\theta})$. We forward model the simulation $X = f(\boldsymbol{\theta})$, calculate the distance $\rho(D, X)$, compare this distance to ϵ_1, and then accept or reject the $\boldsymbol{\theta}$ draw. Because we set ϵ_1 arbitrarily large, the particles essentially sample the prior distribution. This process is repeated until we accept N particles. We then assign equal weights to the N particles: $w_1^i = 1/N$.

For subsequent iterations ($t > 1$), the distance threshold is set such that $\epsilon_{i,t} < \epsilon_{i,t-1}$ for all components i. Although there is no general prescription, the distance threshold $\epsilon_{i,t}$ can be assigned based on the empirical distribution of the accepted distances of the previous iteration, $t - 1$. In Weyant et al. (2013), for instance, the threshold of the second iteration is set to the 25th percentile of the distances in the first iterations; afterwards in the subsequent iterations, t, ϵ_t is set to the 50th percentile of the distances in the previous $t - 1$ iteration. Alternatively, Lin & Kilbinger (2015) set ϵ_t to the median of the distances from the previous iteration. In Section 3, we describe our prescription for the distance threshold, which follows Lin & Kilbinger (2015).

Once ϵ_t is set, we draw a particle from the previous weighted set of particles $\boldsymbol{\theta}_{t-1}$. This particle is perturbed by a kernel, set to the covariance of $\boldsymbol{\theta}_{t-1}$. Then once again, we generate a simulation by forward modelling $X = f(\boldsymbol{\theta}^i)$, calculate the distance $\rho(X, D)$ and compare the distance to the new distance threshold (ϵ_t) in order to accept or reject the particle. This process is repeated until we assemble a new set of N particles $\boldsymbol{\theta}_t$. We then update the particle weights according to the kernel, the prior distribution, and the previous set of weights, as described in Algorithm 1. The entire procedure is then repeated for the next iteration, $t + 1$.

There are a number of ways to specify the perturbation kernel in the ABC-PMC algorithm. A widely used technique is to define the perturbation kernel as a multivariate Gaussian centred on the weighted mean of the particle population with a covariance matrix set to the covariance of the particle population. This perturbation kernel is often called the global multivariate Gaussian kernel. For a thorough discussion of various schemes for specifying the perturbation kernel, we refer the reader to Filippi et al. (2011).

The iterations continue in the ABC-PMC algorithm until convergence is confirmed. One way to ensure convergence is to impose a threshold for the acceptance ratio, which is measured in each iteration. The acceptance ratio is the ratio of the number of proposals accepted by the distance threshold, to the full number of proposed particles at every step. Once the acceptance ratio for an iteration falls below the imposed threshold, the algorithm has converged and is suspended. Another way to ensure convergence is by monitoring the fractional change in the distance threshold ($\epsilon_t/\epsilon_{t-1} - 1$) after each iteration. When the fractional change becomes smaller than some specified tolerance level, the algorithm has reached convergence. Another convergence criterion is through the derived uncertainties of the inferred parameters measured after each iteration. When the uncertainties stabilize and show negligible variations, convergence is ensured. In Section 3.2, we detail the specific convergence criteria used in our analysis.

2.2 Forward model

2.2.1 Halo occupation modelling

ABC requires a forward generative model. In LSS studies, this implies a model that is able to generate a galaxy catalogue. We then calculate and compare summary statistics of the data and model catalogue in an identical fashion In this section, we describe the forward generative model we use within the framework of the HOD.

The assumption that galaxies reside in dark matter haloes is the bedrock underlying all contemporary theoretical predictions for galaxy clustering. The HOD is one of the most widely used approaches to characterizing this galaxy–halo connection. The central quantity in the HOD is $p(N_{\rm g}|M_{\rm h})$, the probability that a halo of mass $M_{\rm h}$ hosts $N_{\rm g}$ galaxies.

The most common technical methods for estimating the theoretical galaxy 2PCF utilize the first two moments of P, which contain the necessary information to calculate the one- and two-halo terms of the galaxy correlation function:

$$1+\xi_{\rm gg}^{1h}(r) \simeq \frac{1}{4\pi r^2 \bar{n}_{\rm g}^2} \int {\rm d}M_{\rm h} \frac{{\rm d}n}{{\rm d}M_{\rm h}} \Xi_{\rm gg}(r|M_{\rm h}) \times \left\langle N_{\rm g}(N_{\rm g}-1)|M_{\rm h}\right\rangle, \tag{8}$$

and

$$\xi_{\rm gg}^{2h}(r) \simeq \xi_{\rm mm}(r) \left[\frac{1}{\bar{n}_{\rm g}} \int {\rm d}M_{\rm h} \frac{{\rm d}n}{{\rm d}M_{\rm h}} \left\langle N_{\rm g}|M_{\rm h}\right\rangle b_{\rm h}(M_{\rm h})\right]^2 \tag{9}$$

In equations (8) and (9), $\bar{n}_{\rm g}$ is the galaxy number density, ${\rm d}n/{\rm d}M_{\rm h}$ is the halo mass function, the spatial bias of dark matter haloes is $b_{\rm h}(M_{\rm h})$ and $\xi_{\rm mm}$ is the correlation function of dark matter. If we represent the spherically symmetric intra-halo distribution of galaxies by a unit-normalized $n_{\rm g}(r)$, then the quantity $\Xi_{\rm gg}(r)$ appearing in the above two equations is the convolution of $n_{\rm g}(r)$ with itself. These fitting functions are calibrated using N-body simulations.

Fitting function techniques, however, require many simplifying assumptions. For example, equations (8) and (9) assume that the galaxy distribution within a halo is spherically symmetric. These equations also face well-known difficulties of properly treating halo exclusion and scale-dependent bias, which results in additional inaccuracies commonly exceeding the 10 per cent level (van den Bosch et al. 2013). Direct emulation methods have made significant improvements in precision and accuracy in recent years (Heitmann et al. 2009, 2010); however, a labour- and computation-intensive interpolation exercise must be carried out each time any alternative statistic is explored, which is one of the goals of this work.

To address these problems, throughout this paper we make no appeal to fitting functions or emulators. Instead, we use the HALOTOOLS package to populate dark matter haloes with mock galaxies and then calculate our summary statistics directly on the resulting galaxy catalogue with the same estimators that are used on observational data (Hearin et al. 2016a). Additionally, through our forward modelling approaching, we are able to explore observables beyond the 2PCF, such as the group multiplicity function, for which there is no available fitting function. This framework allows us to use group multiplicity function for providing quantitative constraints on the galaxy–halo connection. In the following section, we will show that using this observable, we can obtain constraints on the HOD parameters comparable to those found from the 2PCF measurements.

For the fiducial HOD used throughout this paper, we use the model described in Zheng et al. (2007). The occupation statistics of central galaxies follow a nearest integer distribution with first moment given by

$$\langle N_{\rm cen}\rangle = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\log M}}\right)\right]. \tag{10}$$

Satellite occupation is governed by a Poisson distribution with the mean given by

$$\langle N_{\rm sat}\rangle = \langle N_{\rm cen}\rangle \left(\frac{M - M_0}{M_1}\right)^{\alpha}. \tag{11}$$

We assume that central galaxies are seated at the exact centre of the host dark matter halo and are at rest with respect to the halo velocity, defined according to `Rockstar` halo finder (Behroozi, Wechsler & Wu 2013a) as the mean velocity of the inner 10 per cent of particles in the halo. Satellite galaxies are confined to reside within the virial radius following an NFW spatial profile (Navarro et al. 2004) with a concentration parameter given by the $c(M)$ relation (Dutton & Macciò 2014). The peculiar velocity of satellites with respect to their host halo is calculated according to the solution of the Jeans equation of an NFW profile (More, van den Bosch & Cacciato 2009). We refer the reader to Hearin et al. (2016b), Hearin et al. (2016a) and http://halotools.readthedocs.io for further details.

For the halo catalogue of our forward model, we use the publicly available `Rockstar` (Behroozi et al. 2013a) halo catalogues of the `MultiDark` cosmological N-body simulation (Riebe et al. 2013).[2]

[2] In particular, we use the `halotools_alpha_version2` version of this catalogue, made publicly available as part of `Halotools`.

MultiDark is a collisionless dark-matter only N-body simulation. The ΛCDM cosmological parameters of MultiDark are $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$, $\Omega_b = 0.042$, $n_s = 0.95$, $\sigma_8 = 0.82$ and $h = 0.7$. The gravity solver used in the N-body simulation is the Adaptive Refinement Tree code (ART; Kravtsov, Klypin & Khokhlov 1997) run on 2048^3 particles in a $1\,h^{-1}$Gpc periodic box. MultiDark particles have a mass of $m_p \simeq 8.72 \times 10^8\,h^{-1}\,\mathrm{M}_\odot$; the force resolution of the simulation is $\epsilon \simeq 7h^{-1}$ kpc.

One key detail of our forward generative model is that when we populate the MultiDark haloes with galaxies, we do not populate the entire simulation volume. Rather, we divide the volume into a grid of 125 cubic subvolumes, each with side lengths of $200\,h^{-1}$Mpc. We refer to these subvolumes as {BOX1, ..., BOX125}. The first subvolume is reserved to generate the mock observations that we describe in Section 3.1. When we simulate a galaxy catalogue for a given $\boldsymbol{\theta}$ in parameter space, we randomly select one of the subvolumes from {BOX2, ..., BOX125} and then populate the haloes within this subvolume with galaxies. We implement this procedure to account for sample variance within our forward generative model.

2.3 Summary statistics

One of the key ingredients for parameter inference using ABC is the distance metric between the data and the simulations. In essence, it quantifies how close the simulation is to reproducing the data. The data and simulation in our scenario (the HOD framework) are galaxy populations and their positions. A direct comparison, which would involve comparing the actual galaxy positions of the populations, proves to be difficult. Instead, a set of statistical summaries are used to encapsulate the information of the data and simulations. These quantities should sufficiently describe the information of the data and simulations while providing the convenience for comparison. For the positions of galaxies, sensible summary statistics, which we later use in our analysis, include the following:

(i) Galaxy number density, $\bar{n}_\mathrm{g}$: the comoving number density of galaxies computed by dividing the comoving volume of the sample from the total number of galaxies. $\bar{n}_\mathrm{g}$ is measured in units of $(\mathrm{Mpc}/h)^{-3}$.

(ii) Galaxy two-point correlation function, $\xi_{\mathrm{gg}}(r)$: a measurement of the excess probability of finding a galaxy pair with separation r over a random distribution. To compute $\xi_{\mathrm{gg}}(rr)$ in our analysis, for computational reasons, we use the Natural estimator (Peebles 1980):

$$\xi(r) = \frac{DD}{RR} - 1, \tag{12}$$

where DD and RR refer to counts of data–data and random–random pairs.

(iii) Galaxy group multiplicity function, $\zeta_\mathrm{g}(N)$: the number density of galaxy groups in bins of group richness N where group richness is the number of galaxies within a galaxy group. We rely on a Friends-of-Friends (hereafter FoF) group-finder algorithm (Davis et al. 1985) to identify galaxy groups in our galaxy samples. That is, if the separation of a galaxy pair is smaller than a specified linking length, the two galaxies are assigned to the same group. The FoF group-finder has been used to identify and analyse the galaxy groups in the SDSS main galaxy sample (Berlind et al. 2006). For details regarding the group finding algorithm, we refer readers to Davis et al. (1985).

In this study, we set the linking length to be 0.25 times the mean separation of galaxies that is given by $\bar{n}_\mathrm{g}^{-1/3}$. Once the galaxy groups are identified, we bin them into bins of group richness. The total number of groups in each bin is divided by the comoving volume to get $\zeta_\mathrm{g}(N)$ – in units of $(\mathrm{Mpc}/h)^{-3}$.

3 ABC AT WORK

With the methodology and the key components of ABC explained above, here we set out to demonstrate how ABC can be used to constrain HOD parameters. We start, in Section 3.1, by creating our 'observation'. We select a set of HOD parameters that we deem as the 'true' parameters and run it through our forward model producing a catalogue of galaxy positions that we treat as our observation. Then, in Section 3.2, we explain the distance metric and other specific choices we make for the ABC-PMC algorithm. Ultimately, we demonstrate the use of ABC in LSS, in Section 3.3, where we present the parameter constraints we get from our ABC analyses. Lastly, in order to both assess the quality of the ABC-PMC parameter inference and also discuss the assumptions of the standard Gaussian likelihood approach, we compare the ABC-PMC results to parameter constraints using the standard approach in Section 3.4.

3.1 Mock observations

In generating our 'observations', and more generally for our forward model, we adopt the HOD model from Zheng et al. (2007) where the expected number of galaxies populating a dark matter halo is governed by equations (10) and (11). For the parameters of the model used to generate the fiducial mock observations, we choose the Zheng et al. (2007) best-fitting HOD parameters for the SDSS main galaxy sample with a luminosity threshold $M_r = -21$:

$\log M_{\mathrm{min}}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	α
12.79	0.39	11.92	13.94	1.15

Since these parameters are used to generate the mock observation, they are the parameters that we ultimately want to recover from our parameter inference. We refer to them as the true HOD parameters. Plugging them into our forward model (Section 2.2), we generate a catalogue of galaxy positions.

For our summary statistics of the catalogues, we use the mean number density $\bar{n}_\mathrm{g}$, the galaxy two-point correlation function $\xi_{\mathrm{gg}}(r)$ and the group multiplicity function $\zeta_\mathrm{g}(N)$. Our mock observation catalogue has $\bar{n}_\mathrm{g} = 9.288\,75 \times 10^{-4}\,h^{-3}\,\mathrm{Mpc}^3$, and in Fig. 1, we plot $\xi_{\mathrm{gg}}(r)$ (left-hand panel) and $\zeta_\mathrm{g}(N)$ (right-hand panel). The width of the shaded region represents the square root of the diagonal elements of the summary statistic covariance matrix, which is computed as we describe below.

We calculate ξ_{gg} using the natural estimator (Section 2.3) with 15 radial bins. The edges of the first radial bin are 0.15 and $0.5\,h^{-1}$ Mpc. The bin edges for the next 14 bins are logarithmically spaced between 0.5 and $20\,h^{-1}$ Mpc. We compute the $\zeta_\mathrm{g}(N)$ as described in Section 2.3 with nine richness bins, where the bin edges are logarithmically spaced between 3 and 20. To calculate the covariance matrix, we first run the forward model using the true HOD parameters for all 125 halo catalogue subvolumes: {BOX1, ..., BOX125}. We compute the summary statistics of each subvolume galaxy sample k:

$$\boldsymbol{x}^{(k)} = [\bar{n}_\mathrm{g},\ \xi_{\mathrm{gg}},\ \zeta_\mathrm{g}], \tag{13}$$

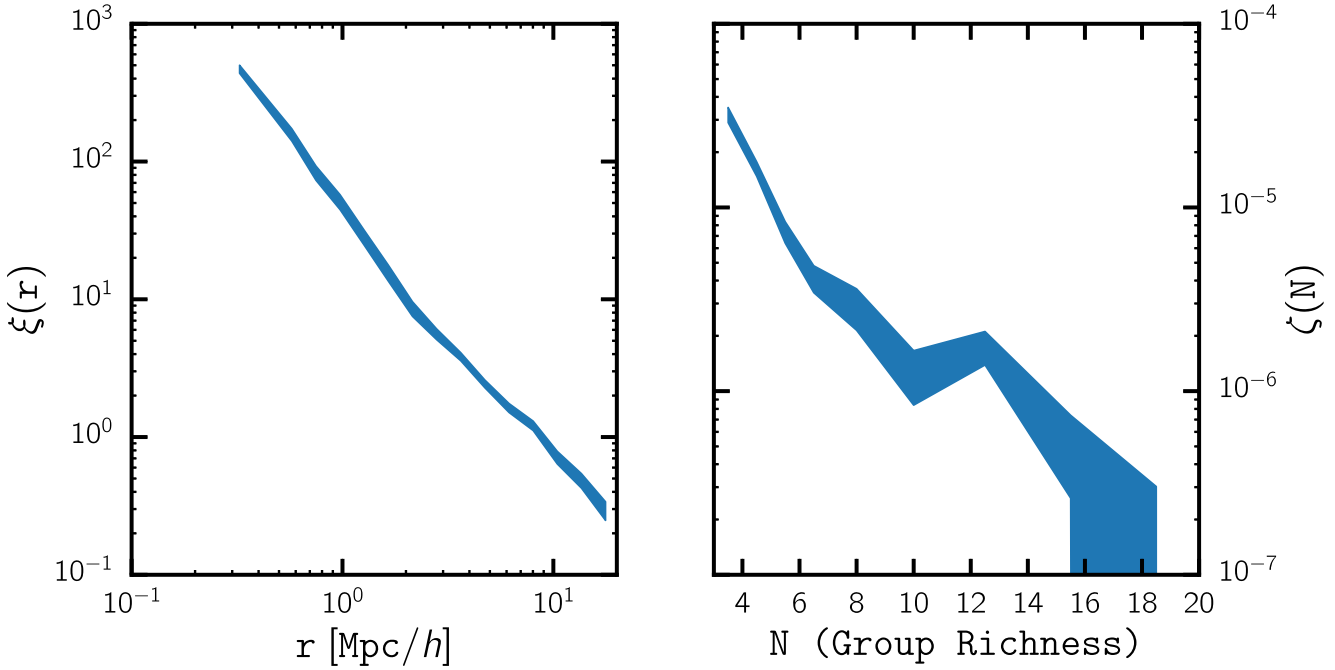

Figure 1. The two-point correlation function $\xi_{\rm gg}(r)$ (left) and group multiplicity function $\zeta_{\rm g}(N)$ (right) summary statistics of the mock observations generated from the 'true' HOD parameters described in Section 3.1. The width of the shaded region corresponds to the square root of the covariance matrix diagonal elements (equation 14). In our ABC analysis, we treat the $\xi_{\rm gg}(r)$ and $\zeta_{\rm g}(N)$ above as the summary statistics of the observation.

Then we compute the covariance matrix as

$$\mathrm{C}_{i,j}^{\rm sample} = \frac{1}{N_{\rm mocks}-1}\sum_{k=1}^{N_{\rm mocks}}\left[\boldsymbol{x}_i^{(k)}-\overline{\boldsymbol{x}}_i\right]\left[\boldsymbol{x}_j^{(k)}-\overline{\boldsymbol{x}}_j\right], \tag{14}$$

$$\text{where } \overline{\boldsymbol{x}}_i = \frac{1}{N_{\rm mocks}}\sum_{k=1}^{N_{\rm mocks}}\boldsymbol{x}_i^{(k)}. \tag{15}$$

Throughout our ABC-PMC analysis, we treat the $\bar{n}_{\rm g}$, $\xi_{\rm gg}(r)$ and $\zeta_{\rm g}(N)$ we describe in this section as if they were the summary statistics of actual observations. However, we benefit from the fact that these observables are generated from mock observations using the true HOD parameters of our choice: we can use the true HOD parameters to assess the quality of the parameter constraints we obtain from ABC-PMC.

3.2 ABC-PMC design

In Section 2.1, we describe the key components of the ABC algorithm we use in our analysis. Now, we describe the more specific choices we make within the algorithm: the distance metric, the choice of priors, the distance threshold and the convergence criteria. So far we have described three summary statistics: $\bar{n}_{\rm g}$, $\xi_{\rm gg}(r)$ and $\zeta_{\rm g}(N)$. In order to explore the detailed differences in the ABC-PMC parameter constraints based on our choice of summary statistics, we run our analysis for two sets of observables: ($\bar{n}_{\rm g}$, $\xi_{\rm gg}$) and ($\bar{n}_{\rm g}$, $\zeta_{\rm g}$).

For both analyses, we use a multicomponent distance (Silk, Filippi & Stumpf 2012, Cisewsky et al., in preparation). Each summary statistic has a distance associated with it: ρ_n, ρ_ξ and ρ_ζ. We calculate each of these distance components as

$$\rho_n = \frac{\left(\bar{n}_{\rm g}^{\rm d}-\bar{n}_{\rm g}^{\rm m}\right)^2}{\sigma_n^2}, \tag{16}$$

$$\rho_\xi = \sum_k \frac{\left[\xi_{\rm gg}^{\rm d}(r_k)-\xi_{\rm gg}^{\rm m}(r_k)\right]^2}{\sigma_{\xi,k}^2}, \tag{17}$$

$$\rho_\zeta = \sum_k \frac{\left[\zeta_{\rm g}^{\rm d}(N_k)-\zeta_{\rm g}^{\rm m}(N_k)\right]^2}{\sigma_{\zeta,k}^2}. \tag{18}$$

The superscripts d and m denote the data and model, respectively. The data are the observables calculated from the mock observation (Section 3.1). σ_n^2, $\sigma_{\xi,k}^2$ and $\sigma_{\zeta,k}^2$ are not the diagonal elements of the covariance matrix (14). Instead, they are diagonal elements of the covariance matrix $C^{\rm ABC}$.

We construct $C^{\rm ABC}$ by populating the entire `MultiDark` halo catalogues 125 times repeatedly, calculating $\bar{n}_{\rm g}$, $\xi_{\rm gg}$ and $\zeta_{\rm g}$ for each realization, and then computing the covariance associated with these observables across all realizations. We highlight that $C^{\rm ABC}$ differs from equation (14), in that it does not populate the 125 subvolumes but the entire `MultiDark` simulation and therefore does not incorporate sample variance. The ABC-PMC analysis instead accounts for the sample variance through the forward generative model, which populates the subvolumes in the same manner as the observations. We use σ_n^2, $\sigma_{\xi,k}^2$ and $\sigma_{\zeta,k}^2$ to ensure that the distance is not biased to variations of observables on specific radial or richness bin.

For our ABC-PMC analysis using the observables $\bar{n}_{\rm g}$ and $\xi_{\rm gg}$, our distance metric is $\boldsymbol{\rho} = [\rho_n, \rho_\xi]$, while the distance metric for the ABC-PMC analysis using the observables $\bar{n}_{\rm g}$ and $\zeta_{\rm g}$ is $\boldsymbol{\rho} = [\rho_n, \rho_\zeta]$. To avoid any complications from the choice for our prior, we select uniform priors over all parameters aside from the scatter parameter $\sigma_{\log M}$, for which we choose a log-uniform prior. We list the range of our prior distributions in Table 1.

With the distances and priors specified, we now describe the distance thresholds and the convergence criteria we impose in our analyses. For the initial iteration, we set distance thresholds for each

Table 1. Prior specifications. The prior probability distribution and its range for each of the Zheng et al. (2007) HOD parameters. All mass parameters are in unit of $h^{-1}\,\mathrm{M}_{\odot}$.

HOD parameter	Prior	Range
α	Uniform	[0.8, 1.3]
$\sigma_{\log \mathrm{M}}$	Log-uniform	[0.1, 0.7]
$\log M_0$	Uniform	[10.0, 13.0]
$\log M_{\rm min}$	Uniform	[11.02, 13.02]
$\log M_1$	Uniform	[13.0, 14.0]

distance component to ∞. This means that the initial pool $\vec{\theta}_1$ is simply sampled from the prior distribution we specify above. After the initial iteration, the distance threshold is adaptively lowered in subsequent iterations. More specifically, we follow the choice of Lin & Kilbinger (2015) and set the distance threshold $\boldsymbol{\epsilon}_t$ to the median of $\boldsymbol{\rho}_{t-1}$, the multicomponent distance of the previous iteration of particles ($\boldsymbol{\theta}_{t-1}$).

The distance threshold $\boldsymbol{\epsilon}_t$ will progressively decrease. Eventually after a sufficient number of iterations, the region of parameter space occupied by $\boldsymbol{\theta}_t$ will remain unchanged. As this happens, the acceptance ratio begins to fall significantly. When the acceptance ratio drops below 0.001, our acceptance ratio threshold of choice, we deem the ABC-PMC algorithm as converged. In addition to the acceptance ratio threshold we impose, we also ensure that distribution of the parameters converges – another sign that the algorithm has converged. Next, we present the results of our ABC-PMC analyses using the sets of observables ($\bar{n}_{\rm g}$, $\xi_{\rm gg}$) and ($\bar{n}_{\rm g}$, $\zeta_{\rm g}$).

3.3 Results: ABC

We describe the ABC algorithm in Section 2.1 and list the particular choices we make in the implementation in the previous section. Finally, we demonstrate how the ABC algorithm produces parameter constraints and present the results of our ABC analysis – the parameter constraints for the Zheng et al. (2007) HOD model.

We begin with a qualitative demonstration of the ABC algorithm in Fig. 2, where we plot the evolution of the ABC $\boldsymbol{\theta}_t$ over the iterations $t = 1$–9, in the parameter space of [$\log \mathcal{M}_1$, $\log \mathcal{M}_{\rm min}$]. The ABC procedure we plot in Fig. 2 uses $\bar{n}$ and $\zeta_{\rm g}(N)$ for observables, but the overall evolution is the same when we use $\bar{n}$ and $\xi_{\rm gg}(r)$. The darker and lighter contours represent the 68 per cent and 95 per cent confident regions of the posterior distribution over $\boldsymbol{\theta}_t$. For reference, we also plot the 'true' HOD parameter $\boldsymbol{\theta}_{\rm true}$ (black star) in each of the panels. The parameter ranges of the panels are equivalent to the ranges of the prior probabilities we specify in Table 1.

For $t = 1$, the initial pool (top left), the distance threshold $\boldsymbol{\epsilon}_1 = [\infty, \infty]$, so $\boldsymbol{\theta}_1$ uniformly samples the prior probability over the parameters. At each subsequent iteration, the threshold is lowered (Section 3), so for $t < 6$ panels, we note that the parameter spaced occupied by $\boldsymbol{\theta}_t$ dramatically shrinks. Eventually when the algorithm begins to converge, $t > 7$, the contours enclosing the 68 per cent and 95 per cent confidence interval stabilize. At the final iteration $t = 9$ (bottom right), the algorithm has converged and we find that $\boldsymbol{\theta}_{\rm true}$ lies within the 68 per cent confidence interval of the $\boldsymbol{\theta}_{t=9}$ particle distribution. This $\boldsymbol{\theta}_t$ distribution at the final iteration represents the posterior distribution of the parameters.

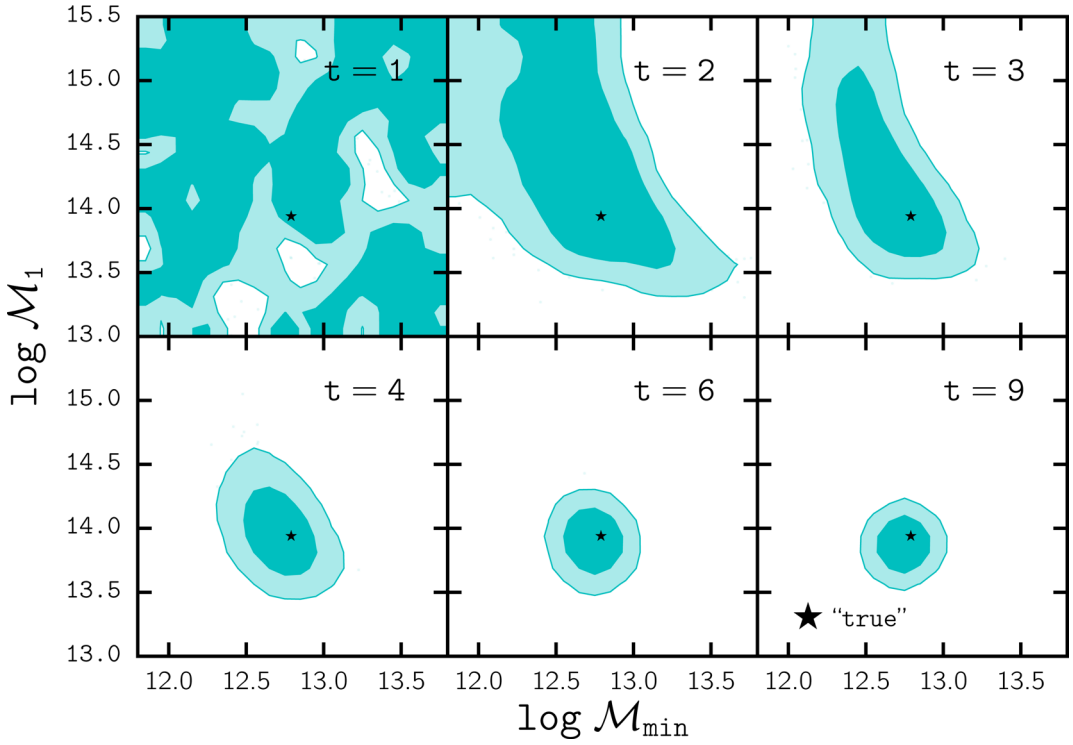

Figure 2. We demonstrate the evolution of the ABC particles, $\boldsymbol{\theta}_t$, over iterations $t = 1$–9 in the $\log \mathcal{M}_{\rm min}$ and $\log \mathcal{M}_1$ parameter space. $\bar{n}$ and $\zeta_{\rm g}(N)$ are used as observables for the above results. For reference, in each panel, we include the 'true' HOD parameters (black star) listed in Section 3.1. The initial distance threshold, $\boldsymbol{\epsilon}_1 = [\infty, \infty]$ at $t = 1$ (top left), so the $\boldsymbol{\theta}_1$ spans the entire range of the prior distribution, which is also the range of the panels. We see for $t < 5$, the parameter space occupied by the ABC $\boldsymbol{\theta}_t$ shrinks dramatically. Eventually when the algorithm converges, $t > 7$, the parameter space occupied by $\boldsymbol{\theta}_t$ no longer shrinks and their distributions represent the posterior distribution of the parameters. At $t = 9$, the final iteration, the ABC algorithm, has converged and we find that $\boldsymbol{\theta}_{\rm true}$ lies safely within the 68 per cent confidence region.

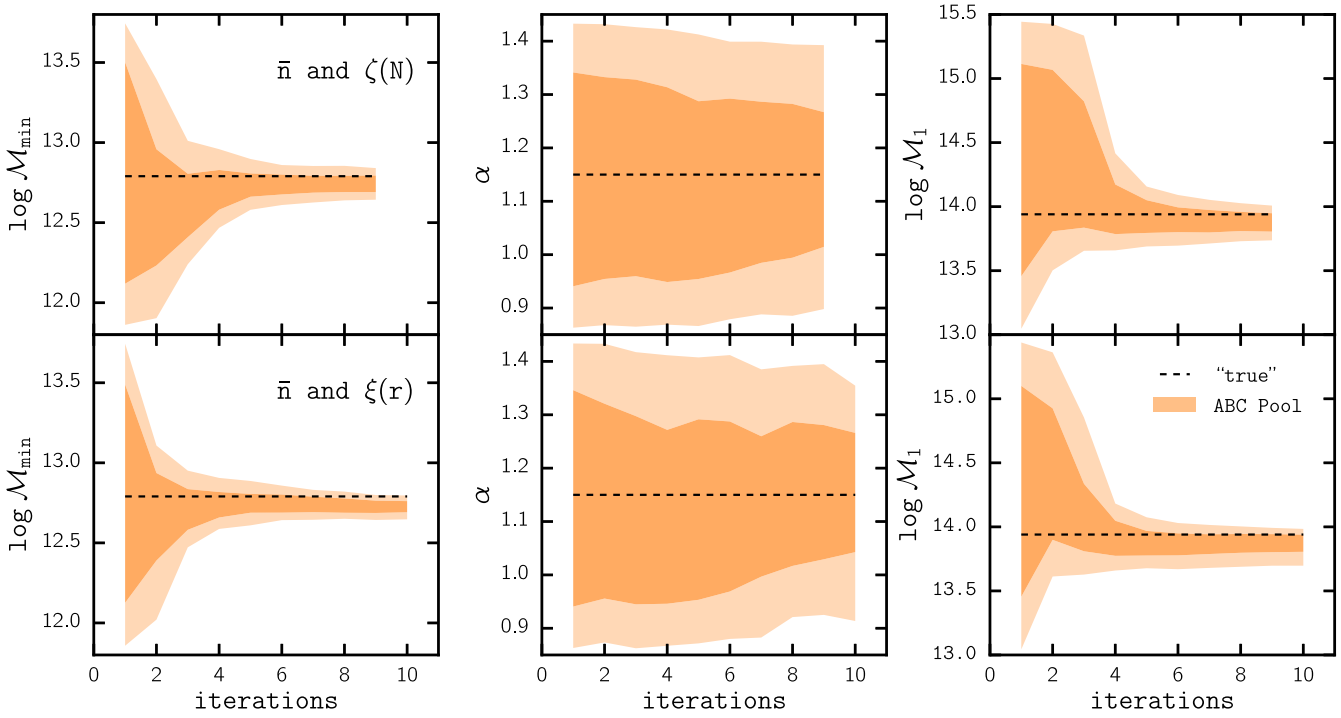

Figure 3. We illustrate the convergence of the ABC algorithm through the evolution of the ABC particle distribution as a function of iteration for parameters $\log \mathcal{M}_{\rm min}$ (left), α (centre) and $\log \mathcal{M}_1$ (right). The top panel corresponds our ABC results using the observables $(\bar{n}, \zeta_{\rm g}(N))$, while the lower panel plots corresponds to the ABC results using $(\bar{n}, \xi_{\rm gg}(r))$. The distributions of parameters show no significant change after $t > 7$, which suggests that the ABC algorithm has converged.

To better illustrate the criteria for convergence, in Fig. 3, we plot the evolution of the $\boldsymbol{\theta}_t$ distribution as a function of iteration for parameters $\log \mathcal{M}_{\rm min}$ (left), α (centre) and $\log \mathcal{M}_1$ (right). The darker and lighter shaded regions correspond to the 68 per cent and 95 per cent confidence levels of the $\boldsymbol{\theta}_t$ distributions. The top panels correspond to our ABC results using $(\bar{n}, \zeta_{\rm g})$ as observables and the bottom panels correspond to our results using $(\bar{n}, \xi_{\rm gg})$. For each of the parameters in both top and bottom panels, we find that the distribution does not evolve significantly for $t > 7$. At this point, additional iterations in our ABC algorithm will neither impact the distance threshold $\boldsymbol{\epsilon}_t$ nor the posterior distribution of $\boldsymbol{\theta}_t$. We also emphasize that the convergence of the parameter distributions coincides with when the acceptance ratio, discussed in Section 3.2, crosses the predetermined shut-off value of 0.001. Based on these criteria, our ABC results for both $(\bar{n}, \zeta_{\rm g})$ and $(\bar{n}, \xi_{\rm gg})$ observables have converged.

We present the parameter constraints from the converged ABC analysis in Figs 4 and 5. Fig. 4 shows the parameter constraints using $\bar{n}$ and $\xi_{\rm gg}(r)$, while Fig. 5 plots the constraints using $\bar{n}$ and $\zeta_{\rm g}(N)$. For both figures, the diagonal panels plot the posterior distribution of the HOD parameters with vertical dashed lines marking the 50 per cent (median) and 68 per cent confidence intervals. The off-diagonal panels plot the degeneracy between parameter pairs. To determine the accuracy of our ABC parameter constraints, we plot the 'true' HOD parameters (black) in each of the panels. For both sets of observables, our ABC constraints are consistent with the 'true' HOD parameters. For $\log \mathcal{M}_0$, $\log \sigma_{\log M}$ and α, the true parameter values lie near the centre of the 68 per cent confidence interval. For the other parameter, which have much tighter constraints, the true parameters lie within the 68 per cent confidence interval.

To further test the ABC results, in Fig. 6, we compare $\xi_{\rm gg}(r)$ (left) and $\zeta_{\rm g}(N)$ (right) of the mock observations from Section 3.1 to the predictions of the ABC posterior distribution (shaded). The error bars of the mock observations represent the square root of the diagonal elements of the covariance matrix (equation 14), while the darker and lighter shaded regions represent the 68 per cent and 95 per cent confidence regions of the ABC posterior predictions. In the lower panels, we plot the ratio of the ABC posterior prediction $\xi_{\rm gg}(r)$ and $\zeta_{\rm g}(N)$ over the mock observation $\xi_{\rm gg}^{\rm obvs}(r)$ and $\zeta_{\rm g}^{\rm obvs}(N)$. Overall, the ratio of the 68 per cent confidence region of ABC posterior predictions is consistent with unity throughout the r and N range. We observe slight deviations in the $\xi_{\rm gg}$ ratio for $r > 5\,{\rm Mpc}/h$; however, any deviation is within the uncertainties of the mock observations. Therefore, the observables drawn from the ABC posterior distributions are in good agreement with the observables of the mock observation.

The ABC results we obtain using the algorithm of Section 2.1 with the choices of Section 3.2 produce parameter constraints that are consistent with the 'true' HOD parameters (Figs 4 and 5). They also produce observables $\xi_{\rm gg}(r)$ and $\zeta_{\rm g}(N)$ that are consistent with $\xi_{\rm gg}^{\rm obvs}$ and $\zeta_{\rm g}^{\rm obvs}$. Thus, through ABC we are able to produce consistent parameter constraints. More importantly, we demonstrate that ABC is feasible for parameter inference in LSS.

3.4 Comparison to the Gaussian pseudo-likelihood MCMC analysis

In order to assess the quality of the parameter inference described in the previous section, we compare the ABC-PMC results with the HOD parameter constraints from assuming a Gaussian likelihood function. The model used for the Gaussian likelihood analysis is different than the forward generative model adopted for the ABC-PMC algorithm to be consistent with the standard approach.

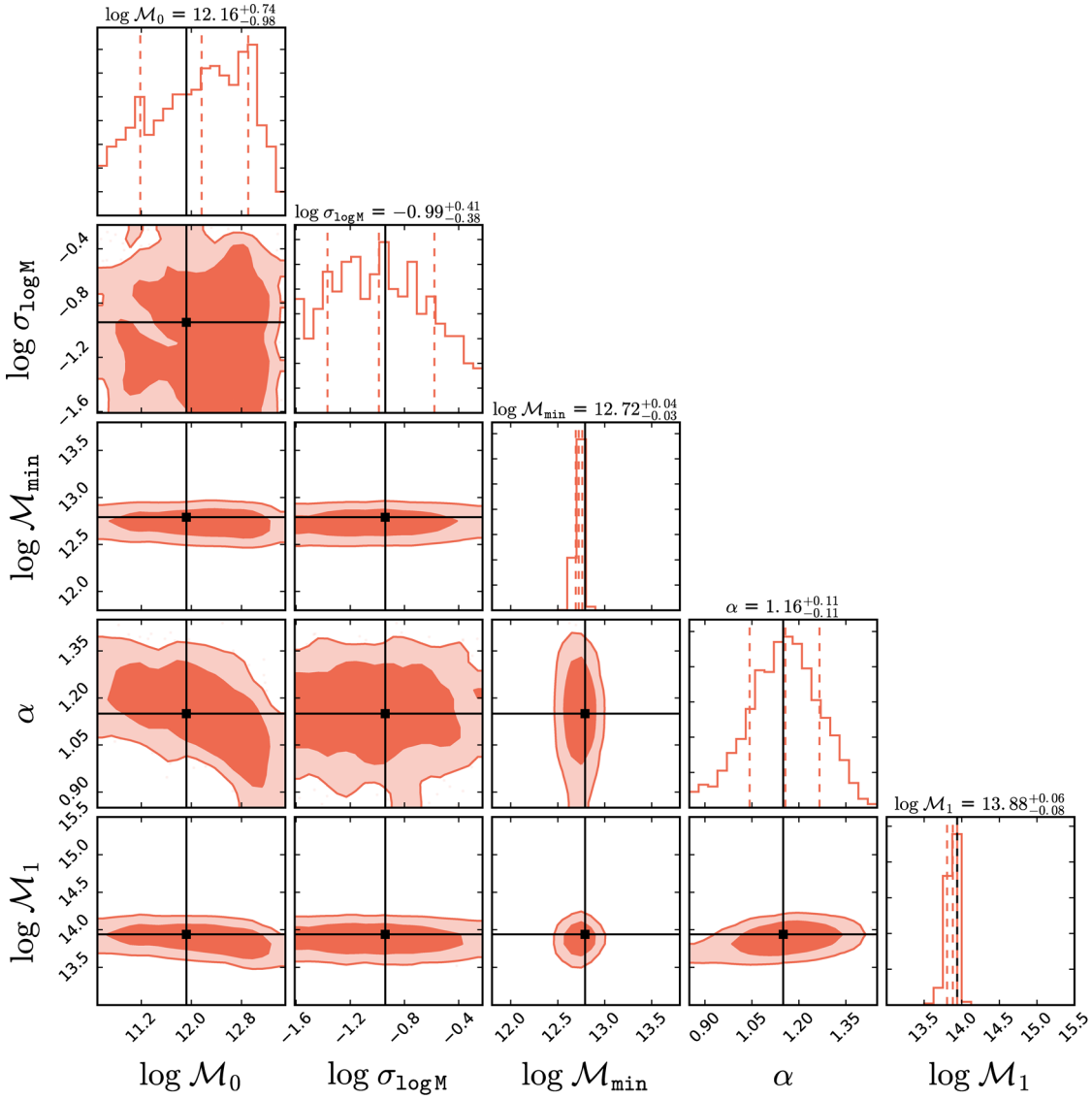

Figure 4. We present the constraints on the Zheng et al. (2007) HOD model parameters obtained from our ABC-PMC analysis using $\bar{n}$ and $\xi_{\rm gg}(r)$ as observables. The diagonal panels plot the posterior distribution of each HOD parameter with vertical dashed lines marking the 50 per cent quantile and 68 per cent confidence intervals of the distribution. The off-diagonal panels plot the degeneracies between parameter pairs. The range of each panel corresponds to the range of our prior choice. The 'true' HOD parameters, listed in Section 3.1, are also plotted in each of the panels (black). For $\log \mathcal{M}_0$, α and $\sigma_{\log M}$, the 'true' parameter values lie near the centre of the 68 per cent confidence interval of the posterior distribution. For $\log \mathcal{M}_1$ and $\log \mathcal{M}_{\rm min}$, which have tight constraints, the 'true' values lie within the 68 per cent confidence interval. Ultimately, the ABC parameter constraints, we obtain in our analysis are consistent with the 'true' HOD parameters.

In the ABC analysis, the model accounts for sample variance by randomly sampling a subvolume to be populated with galaxies. Instead, in the Gaussian pseudo-likelihood analysis, the covariance matrix is assumed to capture the uncertainties from sample variance. Hence, in the model for the Gaussian pseudo-likelihood analysis, we populate haloes of the *entire* `MultiDark` simulation rather than a subvolume. We describe the Gaussian pseudo-likelihood analysis below.

To write down the Gaussian pseudo-likelihood, we first introduce the vector $\boldsymbol{x}$: a combination of the summary statistics (observables) for a galaxy catalogue. When we use $\bar{n}_{\rm g}$ and $\xi_{\rm gg}(r)$ as observables in the analysis: $\boldsymbol{x} = [\bar{n}_{\rm g}, \xi_{\rm gg}]$; when we use $\bar{n}_{\rm g}$ and $\zeta_{\rm g}(N)$ as observables in the analysis: $\boldsymbol{x} = [\bar{n}_{\rm g}, \zeta_{\rm g}]$. Based on this notation, we can write pseudo-likelihood function as

$$-2 \ln \mathcal{L}(\theta|d) = \Delta \boldsymbol{x}^T \widehat{C^{-1}} \Delta \boldsymbol{x} + \ln[(2\pi)^d \det(C)], \quad (19)$$

where

$$\Delta \boldsymbol{x} = [\boldsymbol{x}_{\rm obs} - \boldsymbol{x}_{\rm mod}], \quad (20)$$

the difference between $\boldsymbol{x}_{\rm obs}$, measured from the mock observation and $\boldsymbol{x}_{\rm mod}(\theta)$ measured from the mock catalogue generated from the model with parameters θ. d here is the dimension of $\boldsymbol{x}$ (for $\boldsymbol{x} = [\bar{n}_{\rm g}, \xi_{\rm gg}]$, $d = 13$; for $\boldsymbol{x} = [\bar{n}_{\rm g}, \zeta_{\rm g}]$, $d = 10$). $\widehat{C^{-1}}$ is the inverse covariance matrix, which we estimate following Hartlap, Simon &

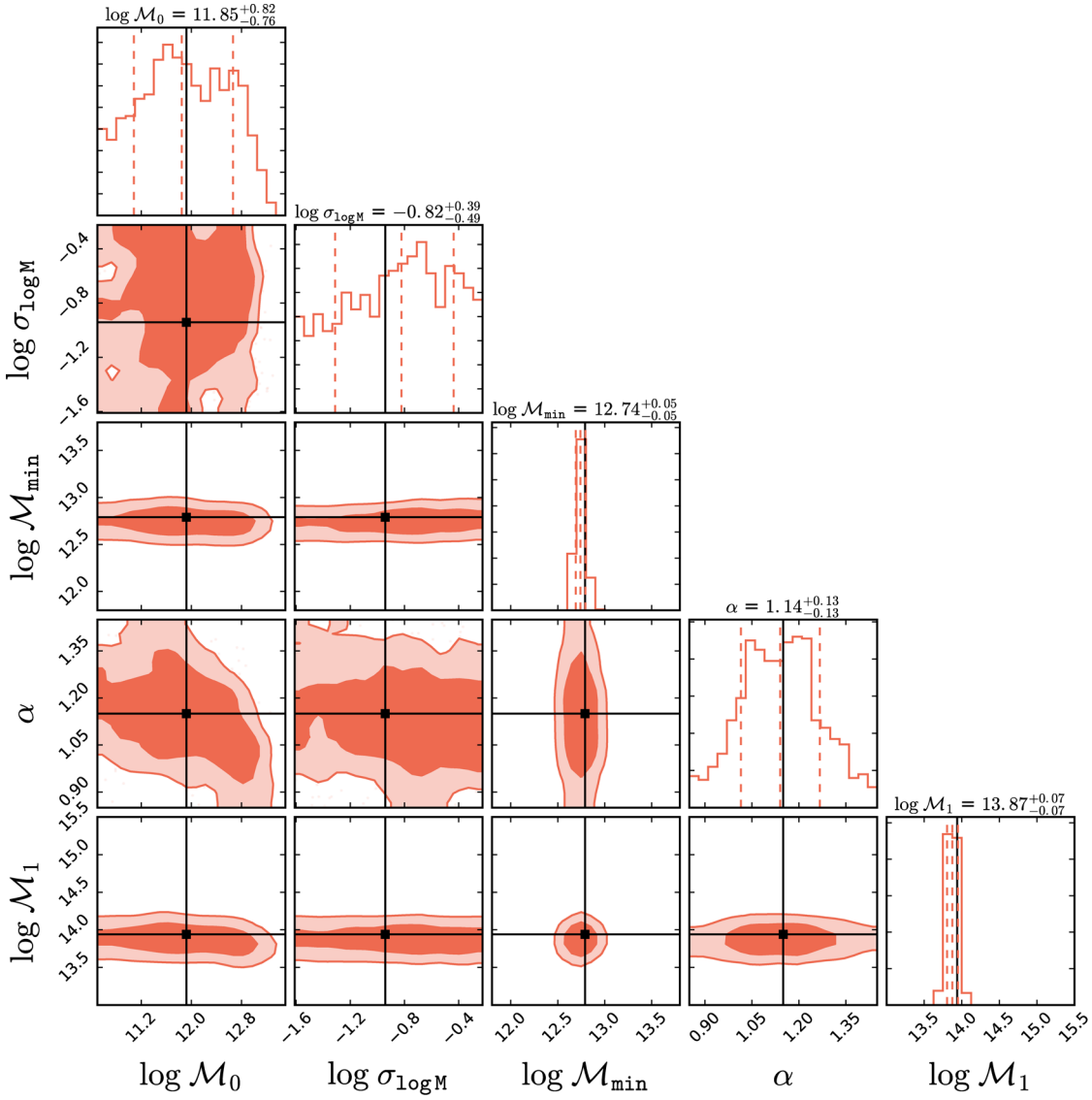

Figure 5. Same as Fig. 4 but for our ABC analysis using $\bar{n}$ and $\zeta_{\rm g}(N)$ as observables. The ABC parameter constraints we obtain are consistent with the 'true' HOD parameters.

Schneider (2007):

$$\widehat{C^{-1}} = \frac{N_{\rm mocks} - d - 1}{N_{\rm mocks} - 1}\,\widehat{C}^{-1}. \quad (21)$$

$\widehat{C}$ is the estimated covariance matrix, calculated using the corresponding $\boldsymbol{x}$ block of the covariance matrix from equation (14) and $N_{\rm mock}$ is the number of mocks used for the estimation ($N_{\rm mock} = 124$; see Section 3.1). We note that in $\widehat{C}$ the dependence on the HOD parameters is neglected, so the second term in the expression of equation (19) can be neglected. Finally, using this pseudo-likelihood, we sample from the posterior distribution given the prior distribution using the MCMC sampler `emcee` (Foreman-Mackey et al. 2013).

In Figs 7 and 8, we compare the results from ABC-PMC and Gaussian pseudo-likelihood MCMC analyses using $[\bar{n}_{\rm g}, \xi_{\rm gg}]$ and $[\bar{n}_{\rm g}, \zeta_{\rm g}]$ as observables, respectively. The top panels in each figure compares the marginalized posterior PDFs for three parameters of the HOD model: $\{\log \mathcal{M}_{\rm min}, \alpha, \log \mathcal{M}_1\}$. The lower panels in each figure compares the 68 per cent and 95 per cent confidence intervals of the constraints derived from the two inference methods as a box plot. The 'true' HOD parameters are marked by vertical dashed lines in each panel.

In both Figs 7 and 8, the marginalized posteriors for each of the parameters from both inference methods are comparable and consistent with the 'true' HOD parameters. However, we note that there are minor discrepancies between the marginalized posterior distributions. In particular, the distribution for α derived from ABC-PMC is less biased than the α constraints from the Gaussian pseudo-likelihood approach.

In Figs 9 and 10, we plot the contours enclosing the 68 per cent and 95 per cent confidence regions of the posterior probabilities of the two methods using $[\bar{n}_{\rm g}, \xi_{\rm gg}]$ and $[\bar{n}_{\rm g}, \zeta_{\rm g}]$ as observables, respectively. In both figures, we mark the 'true' HOD parameters (black star). The overall shape of the contours is in agreement with each other. However, we note that the contours for the ABC-PMC method are more extended along α.

Overall, the HOD parameter constraints from ABC-PMC are consistent with those from the Gaussian pseudo-likelihood MCMC method; however, using ABC-PMC has a number of advantages.

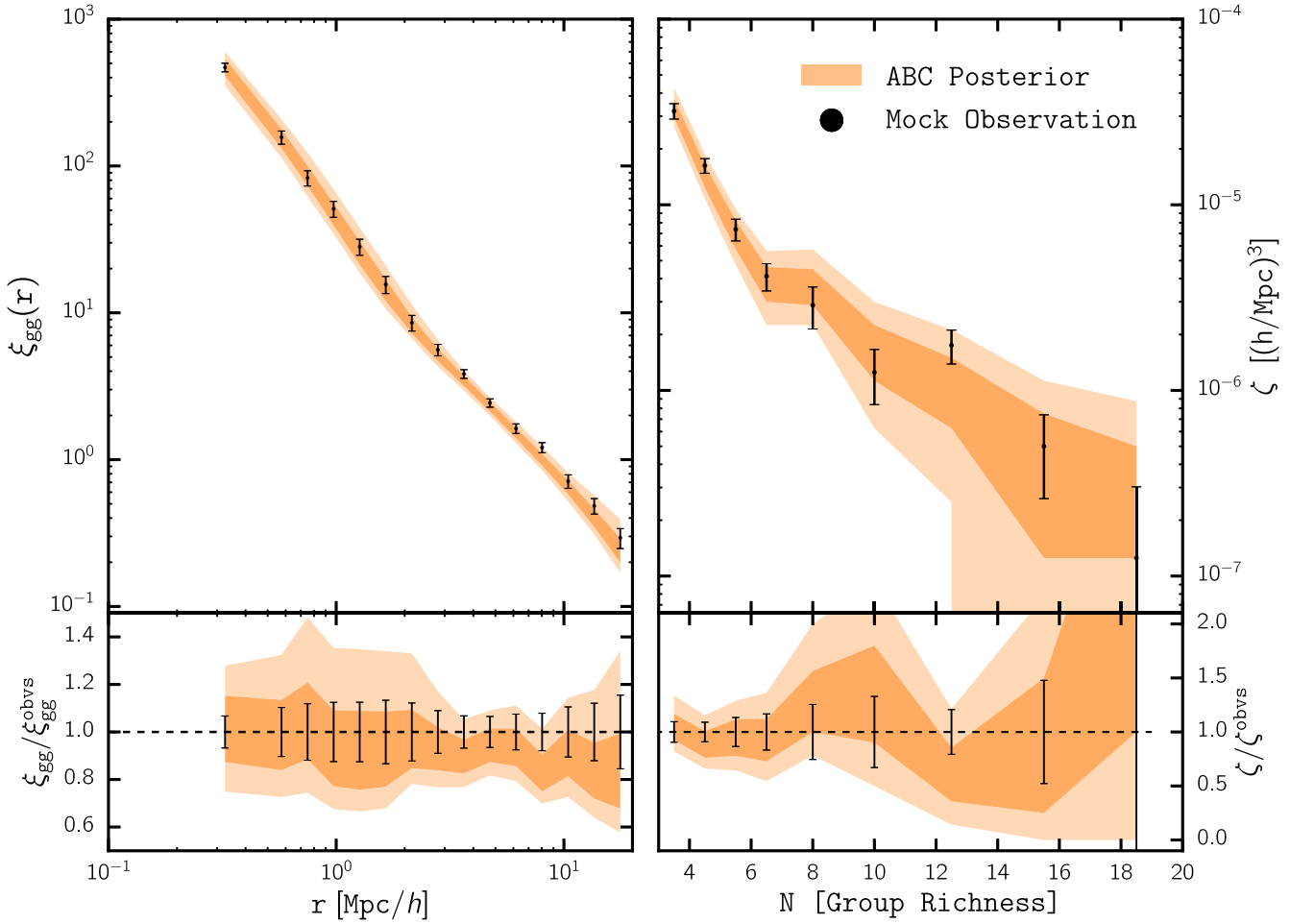

Figure 6. We compare the ABC-PMC posterior prediction for the observables $\xi_{gg}(r)$ (left) and $\zeta_g(N)$ (right) (orange; Section 3.3) to $\xi_{gg}(r)$ and $\zeta_g(N)$ of the mock observation (black) in the top panels. In the lower panels, we plot the ratio between the ABC-PMC posterior predictions for ξ_{gg} and ζ_g to the mock observation ξ_{gg}^{obvs} and ζ_g^{obvs}. The darker and lighter shaded regions represent the 68 per cent and 95 per cent confidence regions of the posterior predictions, respectively. The errorbars represent the square root of the diagonal elements of the error covariance matrix (equation 14) of the mock observations. Overall, the observables drawn from the ABC-PMC posteriors are in good agreement with ξ_{gg} and ζ_g of the mock observations. The lower panels demonstrate that for both observables, the error-bars of the mock observations lie within the 68 per cent confidence interval of the ABC-PMC posterior predictions.

For instance, ABC-PMC utilizes a forward generative model. Our forward generative model accounts for sample variance. On the other hand, the Gaussian pseudo-likelihood approach, as mentioned earlier this section, does not account for sample variance in the model and relies on the covariance matrix estimate to capture the sample variance of the data.

Accurate estimation of the covariance matrix in LSS, however, faces a number of challenges. It is both labour and computationally expensive and dependent on the accuracy of simulated mock catalogues, known to be unreliable on small scales (see Heitmann et al. 2008; Chuang et al. 2015 and references therein). In fact, as Sellentin & Heavens (2016) points out, using estimates of the covariance matrix in the Gaussian psuedo-likelihood approach become further problematic. Even when inferring parameters from a Gaussian-distributed data set, using covariance matrix estimates rather than the *true* covariance matrix leads to a likelihood function that is *no longer* Gaussian. ABC-PMC does not depend on a covariance matrix estimate; hence, it does not face these problems.

In addition to not requiring accurate covariance matrix estimates, forward models of the ABC-PMC method, in principle, also have the advantage that they can account for sources of systematic uncertainties that affect observations. All observations suffer from significant systematic effects that are often difficult to correct. For instance, in SDSS-III BOSS (Dawson et al. 2013), fibre collisions and redshift failures significantly bias measurements and analysis of observables such as ξ_{gg} or the galaxy power spectrum (Guo, Zehavi & Zheng 2012; Ross et al. 2012; Hahn et al. 2017). In parameter inference, these systematics can affect the likelihood, and thus any analysis that requires writing down the likelihood, in unknown ways. With a forward generative model of the ABC-PMC method, the systematics can be simulated and marginalized out to achieve unbiased constraints.

Furthermore, *ABC-PMC – unlike the Gaussian pseudo-likelihood approach – is agnostic about the functional form of the underlying distribution of the summary statistics* (e.g. ξ_{gg} and ζ_g). As we explain throughout the paper, the likelihood function in LSS *cannot* be Gaussian. For ξ_{gg}, the correlation function must satisfy non-trivial positive-definiteness requirements and hence the Gaussian pseudo-likelihood function assumption is not correct in detail. In the case of $\zeta_g(N)$, assuming a Gaussian functional form for the likelihood, which in reality is more likely Poisson, misrepresents the true likelihood function. In fact, this incorrect likelihood may explain why

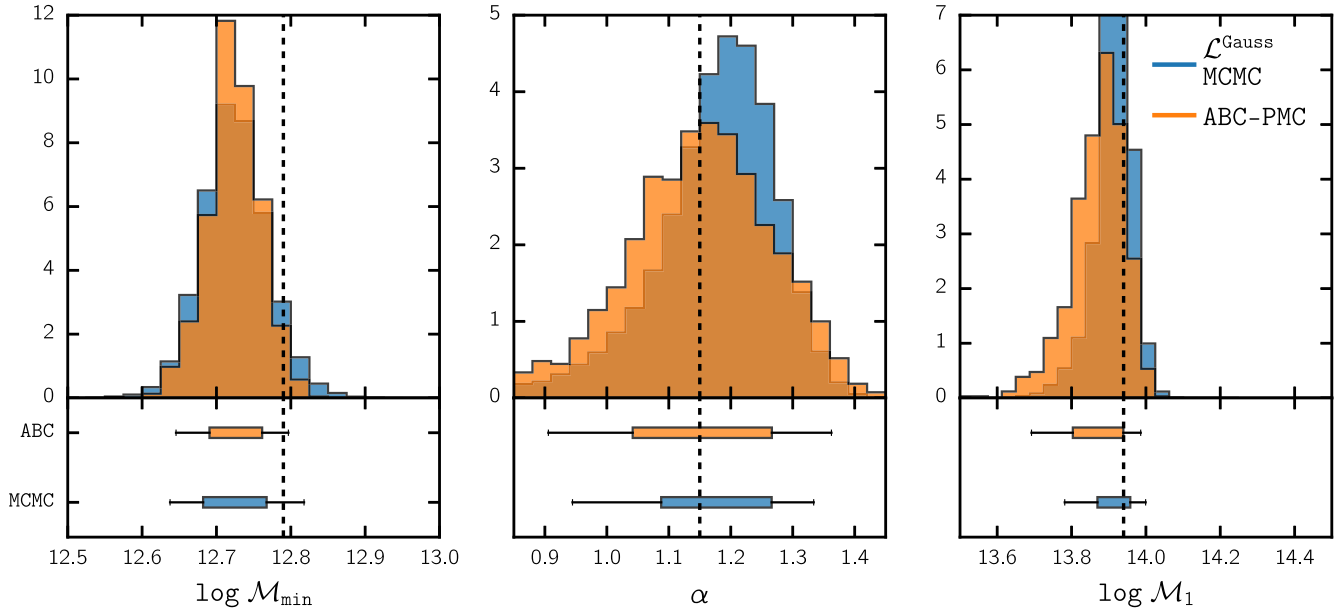

Figure 7. We compare the $\log \mathcal{M}_{\rm min}$, α and $\log \mathcal{M}_1$ parameter constraints from ABC-PMC (orange) to constraints from the Gaussian pseudo-likelihood MCMC (blue) using $\bar{n}_{\rm g}$ and $\xi_{\rm gg}(r)$ as observables. The top panels compares the two methods' marginalized posterior PDFs over the parameters. In the bottom panels, we include box plots marking the confidence intervals of the posterior distributions. The boxes represent the 68 per cent confidence interval, while the 'whiskers' represent the 95 per cent confidence interval. We mark the 'true' HOD parameters with vertical black dashed line. The marginalized posterior PDFs obtained from the two methods are consistent with each other. The ABC-PMC and Gaussian pseudo-likelihood constraints are generally consistent for $\log \mathcal{M}_{\rm min}$ and $\log \mathcal{M}_1$. The ABC-PMC constraint for α is slightly less biased and has slightly larger uncertainty then the constraint from Gaussian pseudo-likelihood analysis.

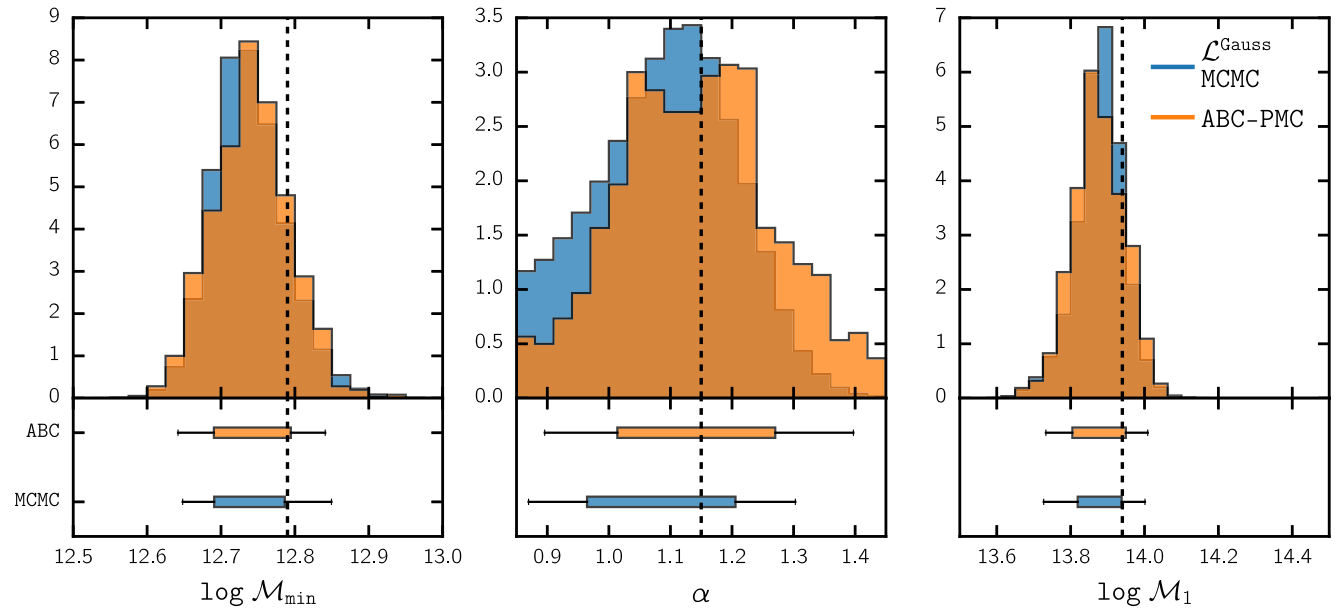

Figure 8. Same as Fig. 7, but both the ABC-PMC analysis and the Gaussian pseudo-likelihood MCMC analysis use $\bar{n}_{\rm g}$ and $\zeta_{\rm g}(N)$ as observables. Both methods derive constraints consistent with the 'true' HOD parameters and infer the region of allowed values to similar precision. We note that the MCMC constraint on α is slightly more biased compared to ABC-PMC estimate. This discrepancy may stem from the fact that the use of Gaussian pseudo-likelihood and its associated assumptions is more spurious when modelling the group multiplicity function.

the constraints on α are less biased for the ABC-PMC analysis than the Gaussian-likelihood analysis in Fig. 10.

Although in our comparison using simple mock observations, we find generally consistent parameter constraints from both the ABC-PMC analysis and the standard Gaussian pseudo-likelihood analysis, more realistic scenarios present many factors that can generate inconsistencies. Consider a typical galaxy catalogue from LSS observations. These catalogues consist of objects with different data qualities, signal-to-noise ratios and systematic effects. For example, catalogues are often incomplete beyond some luminosity/redshift or have some threshold signal-to-noise ratio cut imposed on them.

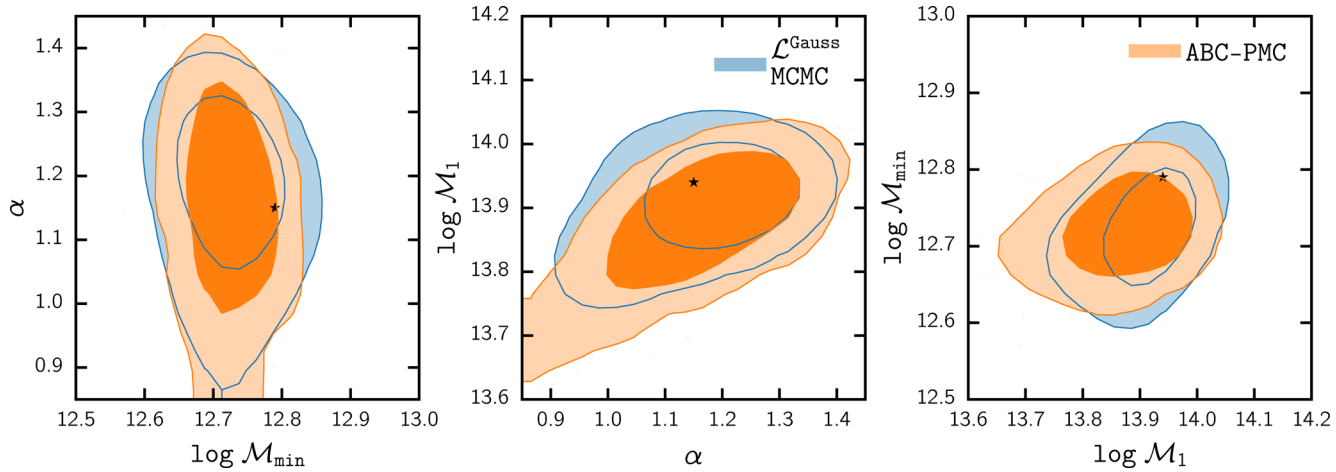

Figure 9. We compare the ABC-PMC (orange) and the Gaussian pseudo-likelihood MCMC (blue) predictions of the 68 per cent and 95 per cent posterior confidence regions over the HOD parameters ($\log \mathcal{M}_{\rm min}$, α and $\log \mathcal{M}_1$) using $\bar{n}_{\rm g}$ and $\xi_{\rm gg}(r)$ as observables. In each panel, the black star represents the 'true' HOD parameters used to generate the mock observations. Both inference methods derive confidence regions consistent with the 'true' HOD parameters.

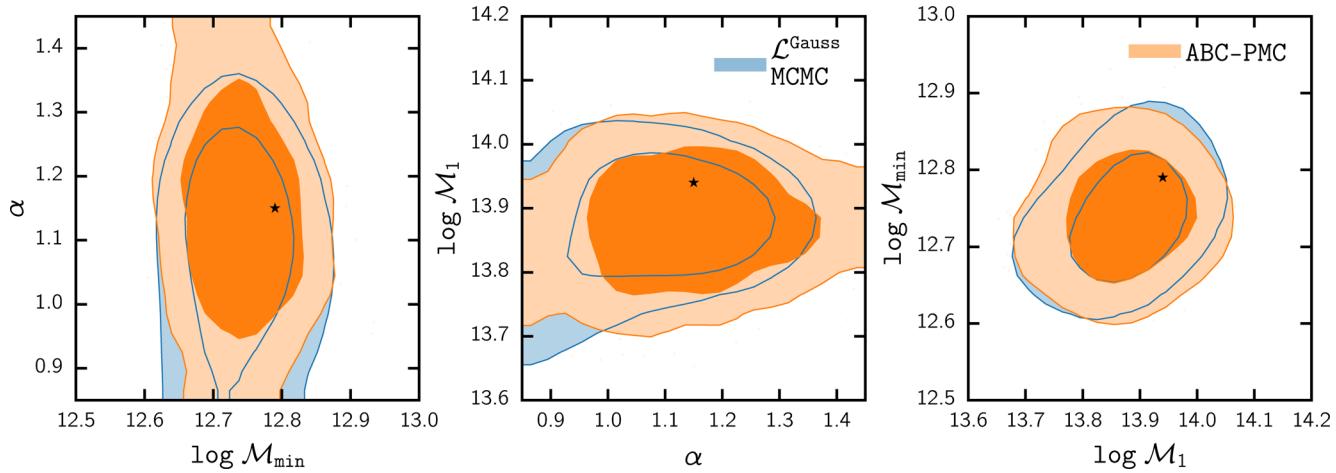

Figure 10. Same as Fig. 9, but using $\bar{n}_{\rm g}$ and $\zeta_{\rm g}(N)$ as observables. Again, the confidence regions derived from both methods are consistent with the 'true' HOD parameters used to generate the mock observations. The confidence region of α from the Gaussian pseudo-likelihood method is biased compared to the ABC-PMC contours. This may be due to the fact that the true likelihood function that describes $\zeta_{\rm g}(N)$ deviates significantly from the assumed Gaussian functional form.

These selection effects, coupled with the systematic effects earlier this section, make correctly predicting the likelihood intractable. In the standard Gaussian pseudo-likelihood analysis, and other analysis that require writing down a likelihood function, these effects can significantly bias the inferred parameter constraints. In these situations, employing ABC equipped with a generative forward model that incorporates selection and systematic effects may produce less biased parameter constraints.

Despite the advantages of ABC, one obstacle for adopting it to parameter inference has been the computational costs of generative forward models, a key element of ABC. By combining ABC with the PMC sampling method, however, ABC-PMC efficiently converges to give reliable posterior parameter constraints. In fact, in our analysis, the total computational resources required for the ABC-PMC analysis were *comparable* to the computational resources used for the Gaussian pseudo-likelihood analysis with MCMC sampling.

Applying ABC-PMC beyond the analysis in this work, to broader LSS analyses, imposes some caveats. In this work, we focus on the galaxy–halo connection, so our generative forward model populates haloes with galaxies. The LSS analyses for inferring cosmological parameters would require generating haloes by running cosmological simulations. The forward models also need to accurately model the observation systematic effects of the latest observations. Hence, accurate generative forward models in LSS analyses demand improvements in simulations and significant computational resources in order to infer unbiased parameter constraints. Recent cosmology simulations show promising improvements in both accuracy and speed (e.g. Feng et al. 2016). Such developments will be crucial for applying ABC-PMC to broader LSS analyses and exploiting the significant advantages that ABC-PMC offers.

4 SUMMARY AND CONCLUSION

Approximate Bayesian Computation, ABC, is a generative, simulation-based inference that can deliver correct parameter estimation with appropriate choices for its design. It has the advantage over the standard approach in that it does not require explicit knowledge of the likelihood function. It only relies on the ability to simulate the observed data, accounting for the uncertainties associated with observation and on specifying a metric for the distance

between the observed data and simulation. When the specification of the likelihood function proves to be challenging or when the true underlying distribution of the observable is unknown, ABC provides a promising alternative for inference.

The standard approach to LSS studies relies on the assumption that the likelihood function for the observables – often two-point correlation function – given the model has a Gaussian functional form. In other words, it assumes that the statistical summaries are Gaussian distributed. In principle to rigorously test such an assumption, a large number of realistic simulations would need to be generated in order to examine the actual distribution of the observables. This process, however, is prohibitively – both labour and computationally – expensive. Therefore, our assumption of a Gaussian likelihood function remains largely unconfirmed and so unknown. Fortunately, the framework of ABC permits us to bypass any assumptions regarding the distribution of observables. Through ABC, we can provide constraints for our models without making the unexamined assumption of Gaussianity.

With the ultimate goal of demonstrating that ABC is feasible for the LSS studies, we use it to constrain parameters of the halo occupation distribution, which dictates the galaxy–halo connection. We begin by constructing a mock observation of galaxy distribution with a chosen set of 'true' HOD model parameters. Then, we attempt to constrain these parameters using ABC. More specifically, in this paper,

(i) we provide an explanation of the ABC algorithm and present how Population Monte Carlo can be utilized to efficiently reach convergence and estimate the posterior distributions of model parameters. We use this ABC-PMC algorithm with a generative forward model built with HALOTOOLS, a software package for creating catalogues of galaxy positions based on models of the galaxy–halo connection such as the HOD;

(ii) we choose $\bar{n}_{\rm g}$, $\xi_{\rm gg}$ and $\zeta_{\rm g}$ as observables and summary statistics of the galaxy position catalogues. And for our ABC-PMC algorithm, we specify a multicomponent distance metric, uniform priors, a median threshold implementation and an acceptance rate-based convergence criterion;

(iii) from our specific ABC-PMC method, we obtain parameter constraints that are consistent with the 'true' HOD parameters of our mock observations. Hence, we demonstrate that ABC-PMC can be used for parameter inference in the LSS studies;

(iv) we compare our ABC-PMC parameter constraints to constraints using the standard Gaussian-likelihood MCMC analysis. The constraints we get from both methods are comparable in accuracy and precision. However, for our analysis using $\bar{n}_{\rm g}$ and $\zeta_{\rm g}$ in particular, we obtain less biased posterior distributions when comparing to the 'true' HOD parameters.

Based on our results, we conclude that ABC-PMC is able to consistently infer parameters in the context of LSS. We also find that the computation required for our ABC-PMC and standard Gaussian-likelihood analyses are comparable. Therefore, with the statistical advantages that ABC offers, we present ABC-PMC as an improved alternative for parameter inference.

ACKNOWLEDGEMENTS

We thank Jessie Cisewsky for reading and making valuable comments on the draft. We would also like to thank Michael R. Blanton, Jeremy R. Tinker, Uros Seljak, Layne Price, Boris Leidstadt, Alex Malz, Patrick McDonald and Dan Foreman-Mackey for productive and insightful discussions. MV was supported by NSF grant AST-1517237. DWH was supported by NSF (grants IIS-1124794 and AST-1517237), NASA (grant NNX12AI50G) and the Moore-Sloan Data Science Environment at NYU. KW was supported by NSF grant AST-1211889. Computations were performed using computational resources at NYU-HPC. We thank Shenglong Wang, the administrator of NYU-HPC computational facility, for his consistent and continuous support throughout the development of this project. We would like to thank the organizers of the AstroHackWeek 2015 workshop (http://astrohackweek.org/2015/), since the direction and the scope of this investigation was – to some degree – initiated through discussions in this workshop. Throughout this investigation, we have made use of publicly available software packages EMCEE and ABCPMC. We have also used the publicly available PYTHON implementation of the FoF algorithm `pyfof` (https://github.com/simongibbons/pyfof).

REFERENCES

Akeret J., Refregier A., Amara A., Seehars S., Hasner C., 2015, J. Cosmol. Astropart. Phys., 8, 043
Ata M., Kitaura F.-S., Müller V., 2015, MNRAS, 446, 4250
Beaumont M. A., Cornuet J.-M., Marin J.-M., Robert C. P., 2009, Biometrika, 96, 983
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013a, ApJ, 762, 109
Behroozi P. S., Wechsler R. H., Conroy C., 2013b, ApJ, 770, 57
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Berlind A. A. et al., 2006, ApJS, 167, 1
Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
Bishop C., 2007, Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. Springer-Verlag, New York
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Cacciato M., van den Bosch F. C., More S., Mo H., Yang X., 2013, MNRAS, 430, 767
Cameron E., Pettitt A. N., 2012, MNRAS, 425, 44
Casas-Miranda R., Mo H. J., Sheth R. K., Boerner G., 2002, MNRAS, 333, 730
Chuang C.-H. et al., 2015, MNRAS, 452, 686
Conroy C., Wechsler R. H., 2009, ApJ, 696, 620
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Dawson K. S. et al., 2013, AJ, 145, 10
Del Moral P., Doucet A., Jasra A., 2006, J. R. Stat. Soc. B, 68, 411
Dressler A., 1980, ApJ, 236, 351
Dutton A. A., Macciò A. V., 2014, MNRAS, 441, 3359
Eriksen H. K. et al., 2004, ApJS, 155, 227
Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, MNRAS, 463, 2273
Filippi S., Barnes C., Stumpf M., 2011, preprint (arXiv:1106.6280)
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Guo H., Zehavi I., Zheng Z., 2012, ApJ, 756, 127
Hahn C., Scoccimarro R., Blanton M. R., Tinker J. L., Rodríguez-Torres S., 2017, MNRAS, 467, 1940
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hearin A. et al., 2016a, preprint (arXiv:1606.04106)
Hearin A. P., Zentner A. R., van den Bosch F. C., Campbell D., Tollerud E., 2016b, MNRAS, 460, 2552
Heitmann K. et al., 2008, Comput. Sci. Discovery, 1, 015003
Heitmann K., Higdon D., White M., Habib S., Williams B. J., Lawrence E., Wagner C., 2009, ApJ, 705, 156
Heitmann K., White M., Wagner C., Habib S., Higdon D., 2010, ApJ, 715, 104
Ishida E. E. O., Vitenti S. D. P., Penna-Lima M., Cisewski J., de Souza R. S., Trindade A. M. M., Cameron E., Busti V. C., 2015, Astron. Comput., 13, 1
Kaiser N., 1984, ApJ, 284, L9

Klypin A. A., Trujillo-Gomez S., Primack J., 2011, ApJ, 740, 102
Knox L., 1995, Phys. Rev. D, 52, 4307
Kravtsov A. V., Klypin A. A., Khokhlov A. M., 1997, ApJS, 111, 73
Leach S. M. et al., 2008, A&A, 491, 597
Leauthaud A. et al., 2012, ApJ, 744, 159
Lemson G., Kauffmann G., 1999, MNRAS, 302, 111
Lin C.-A., Kilbinger M., 2015, A&A, 583, A70
Lin C.-A., Kilbinger M., Pires S., 2016, A&A, 593, A88
Miyatake H. et al., 2015, ApJ, 806, 1
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
More S., van den Bosch F. C., Cacciato M., 2009, MNRAS, 392, 917
More S., van den Bosch F. C., Cacciato M., More A., Mo H., Yang X., 2013, MNRAS, 430, 747
Navarro J. F. et al., 2004, MNRAS, 349, 1039
Oh S. P., Spergel D. N., Hinshaw G., 1999, ApJ, 510, 551
Peebles P. J. E., 1980, The Large-scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Planck Collaboration XVI, 2014, A&A, 571, A16
Planck Collaboration XIII, 2016, A&A, 594, A13
Planck Collaboration XVII, 2016, A&A, 594, A17
Planck Collaboration XX, 2016, A&A, 594, A20
Press W. H., Schechter P., 1974, ApJ, 187, 425
Pritchard J. K., Seielstad M. T., Perez-Lezaun A., Feldman M. W., 1999, Mol. Biol. Evol., 16, 1791
Riebe K. et al., 2013, Astron. Nachr., 334, 691
Riess A. G. et al., 1998, AJ, 116, 1009
Rodríguez-Torres S. A. et al., 2016, MNRAS, 460, 1173
Ross A. J. et al., 2012, MNRAS, 424, 564
Santiago B. X., Strauss M. A., 1992, ApJ, 387, 9
Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, 20
Seljak U., 2000, MNRAS, 318, 203
Sellentin E., Heavens A. F., 2016, MNRAS, 456, L132
Silk D., Filippi S., Stumpf M. P. H., 2012, preprint (arXiv:1210.3296)
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, MNRAS, 320, 289
Steidel C. C., Adelberger K. L., Dickinson M., Giavalisco M., Pettini M., Kellogg M., 1998, ApJ, 492, 428
Tinker J. L., Weinberg D. H., Zheng Z., Zehavi I., 2005, ApJ, 631, 41
Tinker J., Wetzel A., Conroy C., 2011, preprint (arXiv:1107.5046)
Tinker J. L., Leauthaud A., Bundy K., George M. R., Behroozi P., Massey R., Rhodes J., Wechsler R. H., 2013, ApJ, 778, 93
van den Bosch F. C., Mo H. J., Yang X., 2003, MNRAS, 345, 923
van den Bosch F. C., More S., Cacciato M., Mo H., Yang X., 2013, MNRAS, 430, 725
Wandelt B. D., Larson D. L., Lakshminarayanan A., 2004, Phys. Rev. D, 70, 083511
Weyant A., Schafer C., Wood-Vasey W. M., 2013, ApJ, 764, 116
White M., Scott D., 1996, Comm. Astrophys., 18, 289
Zheng Z. et al., 2005, ApJ, 633, 791
Zheng Z., Coil A. L., Zehavi I., 2007, ApJ, 667, 760

This paper has been typeset from a TEX/LATEX file prepared by the author.