

Likelihood free inference of the halo model parameters with simulations of galaxy mock catalogs

ABSTRACT

The Halo Occupation Distribution (HOD) model is a powerful prescription of how galaxies populate dark matter halos that has had much success in reproducing large scale galaxy clustering and understanding galaxy evolution within the context of the halo model. Constraints on the parameters of the HOD model have so far been inferred using a likelihood function that assumes, *First* a functional (often times Gaussian) form of the likelihood, *Second* analytical model for the galaxy clustering statistics. We use Approximate Bayesian Computation (ABC) in order to infer the HOD parameters by assuming a prior over the parameters and a distance function between statistical summaries of the data and statistical summaries of the simulations. Using a mock galaxy catalog created with known HOD parameters, we obtain precise constraints on the HOD parameters that reproduces the summary statistics of the mock galaxy catalog accurately. Since ABC does not require computing the likelihood, we include observables beyond galaxy clustering such as the galaxy number density and the galaxy group richness, for which accurately constructing a likelihood is intractable. We demonstrate that adding these measurements to our set of observables leads to stronger constraints on the HOD parameters, specifically the ones that govern the satellite galaxy population. Moreover, the constraints we infer that reproduce the known HOD parameters illustrate the success of ABC in the context of constraining HOD parameters.

1. Introduction

form of the likelihood. It is often assumed that the likelihood is Gaussian distribution with a given covariance matrix that needs to be evaluated from simulating a large set of mock catalogs, or other techniques such as bootstrap. In order to compute the likelihood function, one needs to invert a high dimensional covariance matrix which makes the likelihood evaluation and therefore sampling from the posterior computationally intractable.

Approximate Bayesian Computation, ABC, is a method that eliminates the need for likelihood. In recent years, this technique has been used in a number of cosmological studies, such as the inference of cosmological parameters from modelling of the distances of

supernovae type Ia, and modelling of weak lensing peak maps. Exploring the parameter space with ABC algorithm requires three essential ingredients: (1) a prior pdf on the parameters that we wish to estimate, (2) a forward model (simulator) of the observations, and (3) a distance metric for comparing the model and the observed data. In problems where likelihood evaluation is not analytically or computationally tractable, this algorithm bypasses the need for likelihood evaluation.

To date, different variations of ABC algorithm have been introduced and implemented in the literature. One of the earliest methods is likelihood free Markov Chain Monte Carlo. In this algorithm, an initial draw from the prior is kept, if the distance between the simulated model corresponding to that parameter and the observed data is less than a threshold. In the consequent steps, the proposed parameters are accepted if *first* they satisfy the distance requirement, and *second* the Metropolis-Hasting (MH) acceptance ratio evaluated for the proposal is less than one. In the MH acceptance ratio, the likelihood is replaced with some transition kernel.

Based on importance sampling, several modifications to this algorithm have been proposed. These methods are called ABC Sequential Monte Carlo (ABC-SMC), or ABC Population Monte Carlo (ABC-PMC). These algorithms rely on sampling a large set of parameters from the prior pdf, and assigning equal weights to them as an initial step. In the subsequent steps, sampling is modified by proposing intermediate distribution built from the previous sample of parameters and their corresponding weights. Parameters are updated by sampling from the intermediate distribution and are accepted if they pass the distance requirements. Furthermore, weights are updated according to the intermediate distribution.

Talk about importance sampling here ... Sampling from the parameter space in ABC algorithm can become more efficient by employing adaptive importance sampling...

2. Method

2.1. Approximate Bayesian Computation

ABC is a likelihood free method for inference of a set of parameters that are required for the simulation of a physical system, through a forward model, a distance function, and a prior probability distribution over the parameters. Let us denote the the data by \mathcal{D} . We assume that there exist a forward model \mathcal{M} that produces a simulation of the data $\bar{\mathcal{D}}$ for a given set of parameters θ

$$\bar{\mathcal{D}}_{\theta} = \mathcal{M}(\theta) \tag{1}$$

Furthermore, ABC requires adopting a distance function between the data \mathcal{D} and the simulation $\bar{\mathcal{D}}_\theta$. The distance function, in general, can be a vector $\mathbf{d}(\mathcal{D}, \bar{\mathcal{D}}_\theta)$

$$\mathbf{d}(\mathcal{D}, \bar{\mathcal{D}}_\theta) = [d_{1,\theta}, \dots, d_{M,\theta}], \quad (2)$$

$$d_{i,\theta} = d_i(\mathcal{S}_i(\mathcal{D}), \mathcal{S}_i(\bar{\mathcal{D}}_\theta)) \text{ for } i = \{1, \dots, M\}, \quad (3)$$

where $d_i(\cdot)$ is a scalar function that defines a distance between i -th summary statistics of the data $\mathcal{S}_i(\mathcal{D})$ and the same summary statistics of the simulation $\mathcal{S}_i(\bar{\mathcal{D}}_\theta)$, and M is the total number of statistical summaries used in the analysis.

ABC algorithm draws a large number of samples from the prior distribution over parameter space, and rejects every sample for which the distance between the data and the simulation corresponding to that sample is not less than a certain threshold ϵ . This requirement is satisfied *if and only if* every component of the distance vector is less than its corresponding threshold. Note that, each summary statistics has a separate threshold. This procedure is done iteratively until, an investigator is left with a certain number of samples that satisfy the distance criteria. The final ensemble approximates the posterior pdf over the parameters of the forward model of the data. This sampling method is called *rejected sampling*.

In practice however, rejected sampling is not desirable because (i) it requires fine tuning of the threshold ϵ whose elements are small, and (ii) finding a large number of samples that pass the threshold criteria is computationally expensive. Alternatively, we draw a large number of proposals from the prior pdf over the HOD parameters with the only requirement that they pass the distance criterion. But since we select an arbitrarily large initial threshold, all the proposals pass the test. Moreover, we assign similar weights to all the proposals in the initial pool. In the next steps, we draw new proposals from the previous pool through weighted sampling, and we perturb the particles by a kernel that is twice the covariance matrix of the proposals in the previous pool.

3. Summary statistics

4. Forward modeling of the galaxy population with HALOTOOLS

5. Mock data

We use HALOTOOLS to populate halos with galaxies above a certain luminosity threshold $M_r > -20$.

6. Results

Hello World!

REFERENCES