Methods in computational cosmology

by

Mohammadjavad Vakili

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Physics

New York University

May 2017

Professor David W. Hogg

Acknowledgements

I am grateful to David Hogg for ...

Abstract

Contents

List of Figures

List of Tables

Introduction

Over the past few years, the field of extrasolar planet (exoplanet) research has really taken off thanks, in large part, to the exquisite time series photometry measured by the Mission (**?**). The Mission enabled the discovery of thousands of planets and planet candidates outside the Solar System (**?**). The zoo of planetary systems is extremely diverse—with sizes, masses, and orbital periods spanning orders of magnitude—and the statistics are now sufficient to test theories of planet formation and evolution.

The Mission has changed the face of exoplanet research because of its photometric precision and the sheer volume of the dataset. In order to discover small planets that serendipitously transit their host stars, the spacecraft was designed to monitor the brightness of about 150,000 stars in one $10^\circ \times 10^\circ$ patch of the sky nearly continuously—at a half-hour cadence—for more than three years with a relative precision of a few parts-per-million for the brightest stars. The Mission surpassed its fiducial goals and took data for over 4 years before two of the reaction wheels used to stabilize the pointing failed in the Spring of 2013.

Despite the fact that most planets never transit their host star—based on geometric effects alone—and the fact that transit surveys are most sensitive to large planets on short orbits, the discoveries made in the dataset and careful characterization of the selection effects and search completeness have enabled detailed studies of the true underlying distribution of planets over a wide range in parameter space (examples include **????**, and Chapter **??**).

These observational studies of the population of exoplanets are arguably the ultimate goal of the Mission because they open the door to direct comparison with theories of planet formation and evolution.

The different constraints on the intrinsic rate and distribution of planets differ in detail but several overarching results are solid. The evidence suggests that every cool M-star has at least one planet in orbit (**??**) and more than half of the other main sequence stars should have planetary systems (**????**). Of these planets, the most intrinsically common are in the "super-Earth" or "mini-Neptune" range from about twice to four-times the radius of Earth. A combination of radial velocity follow-up and hierarchical inference methods indicate that most of these mini-Neptunes gaseous instead of rocky (**??**) but since there are no planets like this in the Solar System, understanding them is crucial to our theories of planetary system formation.

One shortcoming of the Mission was that it only targeted one field and in that frame, the main focus was on relatively faint F, G, and K dwarf stars. This target selection was chosen to enable the study of long-period planets and the discovery of Earth-sized planets orbiting Sun-like stars. Unfortunately many of these stars and their planetary systems are not amenable to radial velocity follow-up because the star is too faint to achieve the required velocity precision or the expected velocity amplitude is too small to detect. In the Summer of 2014, the instrument was re-purposed and it began taking data in a mode called with substantial degraded pointing accuracy (**?**). Because of technical constraints, targets a different field in the ecliptic plane every three months. This means that it can target stars in different environments and focus on gathering the census of planets orbiting bright, nearby stars. It has been demonstrated that the data from can reach precisions comparable to the original Mission and that it can be used to discover transiting exoplanets (**????**, and Chapter **??**). The discoveries made using —and the upcoming Mission—improve

our knowledge of the population of exoplanets, especially those planets that orbit the cool M-stars that were not prioritized by the Mission. These discoveries also present excellent targets for radial velocity follow-up and even spectroscopic observations of their atmospheres using the planned James Webb Space Telescope.

The technical problem of searching for transits in the massive datasets produced by a time series mission like is a hard one. The relative change in brightness caused by the transit of planet in front of its host star is given by the area ratio between the planet and star (**?**). Therefore, when an Earth-sized planet transits a Sun-like star, the amplitude of the signal is smaller than 100 parts-per-million. What's more, in the case of the Earth's orbit, this signal would only last for a little over half a day, once every year. Add to this the fact that most light curves are fraught with signals induced by stellar variability (**?**), spot activity (**?**), and instrumental effects (**??**) with amplitudes far exceeding most transits. In order to find transits, we must, therefore, develop methods for efficiently and robustly mining large sets of light curves for tiny, sparse signals. Nearly all transit search algorithms rely on some sort of matched filter that is made insensitive to noise by pre-processing the light curves to remove the trends or by designing an estimator that is insensitive to these effects (**?????**).

Despite the attempts made to develop search algorithms that are robust to systematics and variability, all automated search results are completely dominated by false signals induced by poorly characterized noise in the light curves. In practice, automatic removal of these events has not been demonstrated to be sufficient—although the results are starting to look promising (**?**)—and all published catalogs of planet candidates are manually vetted. This means that the published list of candidates is *chosen by a person—or group of people—going through the data by hand*. This method is not efficient or scalable so a substantial set of heuristic filtering is applied to the candidate list even before anyone looks at the light curves. One of the standard filters is to only consider candidates with at least three observed

transits (for example **???**). This greatly restricts the range of parameter space that can be search. In particular, these methods will miss any planets on orbits longer than a fraction of the survey baseline.

While transit surveys present the most effective means for systematic exoplanet characterization, their use is limited by existing transit search methodologies for planets with long orbital periods. In many cases, massive long-period planets dominate the dynamics of the planetary systems—like Jupiter in the Solar System—but their existence is completely missed by . This shortcoming becomes even more severe for and , transit surveys with shorter baselines. The final Chapter of this dissertation presents a novel method for transit search designed specifically to discover and quantify these important planets.

The study of exoplanets and their population has been driven by the public dataset and, in particular, by methods and software solutions developed by graduate students and young researchers around the world to squeeze all the available information out of the existing dataset. This dissertation presents methods developed with exactly these goals in mind. Each Chapter is accompanied by an open source implementation of the method and code to reproduce the results and figures. Of these projects, the most popular is the Markov Chain Monte Carlo implementation emcee (**?**, and Chapter **??**). With nearly 300 citations at the time of writing[1] and an active community on GitHub[2], emcee has enabled many modest and ambitious probabilistic inferences across astrophysics.

Chapters **??** and **??** have both been refereed and published in the astronomical literature. Chapter **??** has been submitted to *The Astrophysical Journal* and updated in response to the referee's comments. Chapter **??** is in preparation for submission. All of these Chapters were co-authored with collaborators but the majority of the work and writing in each Chapter is

[1] `http://adsabs.harvard.edu/cgi-bin/nph-ref_query?bibcode=2013PASP..125..306F&refs=CITATIONS&db_key=AST`

[2] `https://github.com/dfm/emcee`

mine. Here, I describe my specific contributions to each Chapter:

1. For Chapter **??**, I generalized the algorithm proposed by **?** through discussions with Jonathan Goodman and David Hogg. I implemented the algorithm with contributions from Dustin Lang and wrote the paper with some additions by David Hogg.

2. For Chapter **??**, I developed the project idea in collaboration with David Hogg and Timothy Morton. I then implemented the project and wrote the paper with contributions from David Hogg.

3. Of the published Chapters, Chapter **??** was the most collaborative. I developed the idea for the algorithm building on previous work with David Hogg, Dun Wang, and Bernhard Schölkopf. Using this algorithm, I wrote the code to search for transits in the K2 Campagin 1 dataset and deployed it on the NYU HPC Butinah cluster[3]. I wrote the majority of the paper with Sections contributed by Ben Montet and Timothy Morton.

4. The fundamental ideas underlying Chapter **??** were developed through discussions with Bernhard Schölkopf and David Hogg. The implementation and text are mine.

*arg max

[3] http://nyuad.nyu.edu/en/research/infrastructure-and-support/nyuad-hpc.html

Chapter 1

Fast stellar centroiding and saturation of the Cramer-Rao lower bound

This Chapter is joint work with David W. Hogg (NYU), submitted to the *Astronomical Astronomical Journal* as **?**.

1.1 Chapter abstract

One of the most demanding tasks in astronomical image processing—in terms of precision—is the centroiding of stars. Upcoming large surveys are going to take images of billions of point sources, including many faint stars, with short exposure times. Real-time estimation of the centroids of stars is crucial for real-time PSF estimation, and maximal precision is required for measurements of proper motion.

The fundamental Cramér-Rao lower bound sets a limit on the root-mean-squared-error achievable by optimal estimators. In this work, we aim to compare the performance of various centroiding methods, in terms of saturating the bound, when they are applied to relatively

low signal-to-noise ratio unsaturated stars assuming zero-mean constant Gaussian noise. In order to make this comparison, we present the ratio of the root-mean-squared-errors of these estimators to their corresponding Cramér-Rao bound as a function of the signal-to-noise ratio and the full-width at half-maximum of faint stars.

We discuss two general circumstances in centroiding of faint stars: (i) when we have a good estimate of the PSF, (ii) when we do not know the PSF. In the case that we know the PSF, we show that a fast polynomial centroiding after smoothing the image by the PSF can be as efficient as the maximum-likelihood estimator at saturating the bound. In the case that we do not know the PSF, we demonstrate that although polynomial centroiding is not as optimal as PSF profile fitting, it comes very close to saturating the Cramér-Rao lower bound in a wide range of conditions. We also show that the moment-based method of center-of-light never comes close to saturating the bound, and thus it does not deliver reliable estimates of centroids.

1.2 Introduction

Accuarate estimates of the centers of point sources, which are convolved with telescope point spread function (PSF), and atmospheric PSF in case of ground based telescopes, and the pixel response function, are crucial to further steps of astronomical image processing. For instance, proper measurement of the shapes of galaxies requires interpolation of the PSF estimates from the positions of stars across the image to the positions of galaxies. At the position of each star, the PSF is estimated by sub-pixel shifting of the star so that the PSF is centered on its centroid. If the sub-pixel shifts are wrong, then the PSF estimates will be biased. Moreover, measurements of the parallaxes and the proper motions of stars depend on how well we can measure their centroids.

Ideally, we want a centroiding procedure that provides measurements as precise as possible without putting a huge computational burden on the photometric pipeline. Reducing the computational cost becomes even more important in large surveys, where we want to estimate the centroids of thousands of point sources detected on the telescope's focal plane, for various real-time applications.

The Cramér-Rao lower bound (CRLB) sets a lower limit on the root-mean-squared error of estimators. When the root-mean-squared error arsing from an estimator approaches the bound, the bound is saturated by that estimator. In this paper, we study the optimality of various techniques for centroiding faint, unsaturated stars. Our requirement for optimality is saturation of the theoretically-set lower bound, known as the Cramér-Rao lower bound, by the centroiding methods considered in this study.

We apply a number of centroiding methods to a large number of simulated faint stars, assuming uncorrelated Gaussian noise, with different signal-to-noise ratio and size realizations. The Cramér-Rao lower bound has an inverse relation with the signal-to-noise-ratio of stars. In the context of astrometry, the Cramér-Rao lower bound saturation for least-squares estimators has been tested in specific limits in which the centroiding bias is negligible (**?**).

Saturating the Cramér-Rao lower bound in estimating the centroids of stars however, is limited by the lack of knowledge about the exact shape of the PSF and presence of noise. There are many sources of noise such as the CCD readout noise, sky noise, errors resulting from incorrect flatfield corrections, and photon noise from the astronomical object itself. In this study, we limit our investigation to the simulated images that contain non-overlapping faint sources that are sky-limited.

We focus the scope of this investigation to sky-limited images for which the sky level has been subtracted. Furthermore, we assume that any instrument gain has been calibrated out and that the simulated images are free of any contamination by cosmic rays, stray light from

neighboring fields, or any other type of defect in real images. We expect these defects to move the centroiding errors further from the fundamental bound. We intend to investigate whether fast centroiding estimates can saturate the bound in a realistic range of low signal-to-noise ratio images that are sky-limited.

Given an analytic expression for the PSF model adopted in this study, we derive an expression for the fundamental lower bound on the centroiding error as a function of the parameters of the PSF model (e.g. PSF size), and signal-to-noise-ratio of stars. We create two sets of simulations for which we can compute the CRLB, one with variable signal-to-noise ratio and constant full width at half maximum (FWHM), and one with variable FWHM and constant signal-to-noise ratio. After applying different centroiding methods to the simulations, we investigate how close these methods can get to saturating the CRLB for various ranges of background Gaussian noise level and PSF FWHM.

In this work, we focus on four centroiding methods. The first method is the maximum-likelihood estimator which involves fitting a PSF profile, assuming that we have a good PSF estimate, to the star. The second method estimates the centroid of a star by fitting a 2d second-order polynomial to the 3×3 patch around the brightest pixel of the image after convolution with the PSF. The third method centroids stars by smoothing the image of stars by a Gaussian kernel of a fixed size, and then applying the same 3×3 polynomial trick to the smooth image. This method is fast and does not require any knowledge of the PSF. The last method we consider, is a center-of-light centroiding (measurement of a first moment), applied to the 7×7 patch around the brightest pixel of the image.

This paper is structured as follows. In Section 1.3, we discuss the Cramér-Rao lower bound and derive an analytic expression for the lower bound on the centroiding error of the simulated data. In Section 5.3 we give a brief overview of centroiding methods used in our investigation. In Section 4.4 we discuss the Cramér-Rao lower bound satuaration tests and

their corresponding simulated data. In Section 4.6, we compare the performances of the methods discussed in Section 5.3 with the CRLB derived in Section 1.3. Finally, we discuss and conclude in Section 5.5.

1.3 The Cramér-Rao lower bound

The Cramér-Rao lower bound sets a limit, in some sense, on how well a measurement can be made in noisy data. The bound can only be computed in the context of a generative model, or a probabilistic forward model of the data. That is, we can only compute the CRLB in the context of assumptions about the properties of the data. However, it makes sense for us to use centroiding methods that saturate the CRLB under some reasonable assumptions, even if we find that those assumptions are not strictly correct in real situations.

The closer an estimator is to saturating the CRLB, the more information about the quantity that we need to estimate is preserved. The closer the root-mean-squared-error (RMSE) of a given estimator is to the bound, the more optimal—in terms of preseving the information—the estimator is.

The Cramér-Rao inequality (**?**) sets a lower bound on the root-mean-squared error of unbiased estimators. The CRLB is given by the square-root of the inverse of the Fisher information matrix $\mathcal{F}$. Thus, in order to find the CRLB, it is sufficient to compute the Fisher matrix. This computation relies on a set of assumptions:

- Known PSF model. In this work the presumed model is the Moffat PSF profile.
- Known, stationary noise process. In the context of centroiding stars, this is equivalent to having background limited noise from sky background and CCD readout noise.
- Images are calibrated correctly. Flat-field is correctly calibrated.

- Uncorrelated Gaussian noise with no outliers.

Note that in this study, we explicitly focus on sky-limited images. In the sky-limited images, the contribution to the Poisson pixel noise is largely dominated by the sky rather than the objects. In sky-limited images, when the number of photons per pixel is large, the Poisson noise can be approximated by a Gaussian distribution. Therefore, the Gaussian noise assumption is only an approximation to the Poisson noise. This is a good approximation for a large set of astronomical images.

A number of factors can produce correlation between pixels. These include detector imperfection, saturation, and post-processing of images such as smoothing, rotating, and shifting the images. In raw unsaturated images, pixel noise is close to uncorrelated. Instrument gain can introduce heteroscedasticity. In that case, the noise variance varies between pixels. In an upcoming publication on the inference of the HST WFC3-IR channel PSF (Vakili *et al.*, in preparation), we discuss proper treatment of centroiding in the presence of gain. For simplicity, we assume that per-pixel uncertainty remains constant across all pixels.

Let us assume that there are M observables $\mathbf{f} = (f_1, ..., f_M)$, each related to B model parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_B)$

$$f_m = f_m(\theta_1, ..., \theta_B). \tag{1.1}$$

Assuming uncorrelated Gaussian error with variance σ_m^2 for each observable f_m, elements of the $B \times B$ Fisher matrix $\mathcal{F}_{ij}$ are given by

$$\mathcal{F}_{ij} = \sum_{m=1}^{M} \frac{1}{\sigma_m^2} \frac{\partial f_m}{\partial \theta_i} \frac{\partial f_m}{\partial \theta_j} \tag{1.2}$$

Let us assume that we have computed the root-mean-squared error on the parameter θ_i arising from applying an estimator to a large number of data. The Cramér-Rao inequality states that this root-mean-squared error is greater than or equal to the i-th diagonal element

of the inverse of the Fisher information matrix:

$$\text{RMSE} \geq \sqrt{[\mathcal{F}^{-1}]_{ii}}, \tag{1.3}$$

where the left hand side of the inequality is called the Cramér-Rao bound on the root-mean-squared error of estimating the parameter θ_i. Note that the bound is computed assuming that the model (equation 1.1) generating the data is known, and that uncertainties are given by additive uncorrelated Gaussian noise.

Based on Cramér-Rao inequality (1.3), **?** defines efficiency of optimal estimators as the ratio of the CRLB and the root-mean-squared-error such that the maximum efficiency achievable by an estimator is unity. The closer the RMSE to the CRLB, the more information about the parameter of interest is preserved, and thus the more efficient the estimator is.

Let us consider the case of a maximum likelihood estimate $\boldsymbol{\theta}_{\text{ML}}$, where the likelihood function corresponds to the same generative assumptions that we used to compute the CRLB.

$$\begin{aligned} \boldsymbol{\theta}_{\text{ML}} &= \mathcal{L}, & (1.4)\\ -2\ln\mathcal{L} &= \sum_m \frac{1}{\sigma_m^2}(y_m - f_m(\boldsymbol{\theta}))^2, & (1.5)\\ & & (1.6) \end{aligned}$$

where y_m is the mth component of the observed data $\mathbf{y}$

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}_{\text{true}}) + \mathbf{n}. \tag{1.7}$$

Maximum likelihood estimators can achieve maximum efficiency. That is, when a maximum likelihood estimator is applied to a large number of data and RMSE is computed,

the RMSE approaches the CRLB (see **?**; **?** for proof) in which case the CRLB is saturated. Therefore, we want to investigate the conditions under which the RMSE arising from a given fast centroiding method is close to the CRLB, or whether it can saturate the CRLB.

In this investigation, the model observables for the noisy data are the pixel-convolved PSF (PSF profile evaluated at different pixel locations), and the model parameters under consideration are the centroid coordinates. Therefore, $\mathcal{F}$ is a 2×2 matrix whose elements are given by

$$\mathcal{F}_{ij} = \sum_m \frac{1}{\sigma^2} \frac{\partial f_m}{\partial \theta_i} \frac{\partial f_m}{\partial \theta_j}, \tag{1.8}$$

where the summation is over pixels, f_m is the value of the PSF at pixel location m, $\theta = \{x_c, y_c\}$, and σ^2 is variance of the uncorrelated Gaussian noise map $n(\mathbf{x}_m)$

$$\mathbb{E}[n(\mathbf{x}_m)] = 0, \tag{1.9}$$

$$\mathbb{E}[n(\mathbf{x}_m)n(\mathbf{x}_{m'})] = \sigma^2 \delta_{m,m'}. \tag{1.10}$$

Derivation of an explicit expression for the Fisher matrix $\mathcal{F}$ requires specifying a presumed correct PSF model. We use the Moffat profile (**?**) for our PSF simulations. The Moffat profile is an analytic model for stellar PSFs. It has broader wings than a simple Gaussian profile. The surface brightness of the Moffat profile is given by

$$I(r) = \frac{F(\beta - 1)}{\pi \alpha^2} [1 + (r/\alpha)^2]^{-\beta}, \tag{1.11}$$

where F is the total flux, β is a dimensionless parameter, and α is the scale radius of the Moffat profile, with FWHM (hereafter denoted by γ) being $2\alpha\sqrt{2^{1/\beta} - 1}$. The Moffat PSF profile has been used in the PSF modeling required for weak lensing galaxy shape

measurements (see **??**). It has also been used as one of the methods for generation of the PSF in simulation of images needed for weak lensing systematic studies (**?**). At a fixed γ, Moffat profiles with lower values of β have broader tails. It is also important to note that for sufficiently large values of the parameter β, the Moffat PSF becomes arbitrarily close to a simple Gaussian PSF.

Note that in our data generation, we simulate images (in the pixel space) that are Nyquist-sampled or close to Nyquist-sampled. All pixels in the images are identical, and the stars are simulated by sampling from the pixel-convolved PSF. In well-sampled images, the center of the pixel-convolved PSF must be very close to the center of the optical PSF.

In order to investigate the performance of centroiding methods for different background noise levels and different values of the parameter γ, simulation of a large number of images of stars—for which the exact positions of centroids and their corresponding lower bounds are known—is required.

Given the PSF model (1.11), an expression for the CRLB as a function of the size, and SNR of stars can be derived. For further simplicity, the flux of all stars in our simulations are set to unity and per-pixel uncertainties are assumed to be uncorrelated Gaussian.

Moreover, it is more convenient to work with the signal-to-noise ratio (SNR) instead of the variance of the Gaussian noise. We use the definition of SNR according to which SNR is given by the ratio of the mean and variance of the distribution which the flux estimator is drawn from. Assuming that the total flux from the point source is F, and that the sub-pixel shifted PSF at the i-th pixel is given by P_i. Therefore the brightness of the i-th pixel y_i is drawn from a Gaussian distribution

$$p(y_i) = \mathcal{N}(FP_i, \sigma^2). \tag{1.12}$$

The optimal estimator of flux is the matched-filter flux estimator $\tilde{F} = \sum_i y_i P_i$. It can be shown that

$$p(\tilde{F}) = \mathcal{N}(F, \frac{\sigma^2}{\sum_i P_i^2}), \tag{1.13}$$

which leads us to

$$\text{SNR} = \frac{F\sqrt{\sum_i P_i^2}}{\sigma}. \tag{1.14}$$

In the case of Moffat profiles (1.11) with total flux of stars set to unity, the SNR given in (1.14) can be analytically expressed in terms of the per pixel uncertainty σ, FWHM γ, and also β, the dimensionless parameter of (1.11)

$$\text{SNR} = \frac{2(\beta-1)(2^{1/\beta}-1)^{1/2}}{\pi^{1/2}(2\beta-1)^{1/2}} \frac{1}{\sigma\gamma}. \tag{1.15}$$

Equation (1.15) implies that at a fixed γ and background Gaussian noise with variance σ^2, stars with broader tails (lower β) have a lower SNR. On the other hand, stars with higher β have higher SNR. For sufficiently large β—where the PSF can be approximated by Gaussian profile—SNR is approximately given by $0.664/(\sigma\gamma)$. Furthermore, at a fixed β and variance of the background noise σ^2, observed stars with higher γ have lower SNR.

Throughout this investigation, β is held fixed at the fiducial value of $\beta = 2.5$, where SNR is given by the following expression

$$\text{SNR} \simeq \frac{0.478}{\sigma\gamma} \quad \text{for} \quad \beta = 2.5. \tag{1.16}$$

Given the analytic expression for the Moffat PSF model (1.11), and choice of $\beta = 2.5$,

the inverse of the Fisher matrix is given by

$$\mathcal{F}^{-1} \simeq \left(0.685\frac{\gamma}{\mathrm{SNR}}\right)^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{1.17}$$

Equation (1.17) implies that at given SNR and γ, CRLB for each component of centroid is approximately given by $0.685\gamma/\mathrm{SNR}$, and that a good centroiding technique delivers centroids with root-mean-squared-error close to this.

It is worth noting that for any PSF model whose radial light profile is some function of r/γ, CRLB has the same functional form, in that it is proportional to the ratio between γ and the SNR. For PSF profiles with shorter tails (e.g., Gaussian), the prefactor of 0.685 in (1.17) becomes smaller. In the particular case of Gaussian PSF, the prefactor is approximately 0.6.

1.4 Centroiding methods

In this section, we briefly discuss the approximate and the non-approximate centroiding methods considered in this study. The first two methods require knowledge of the PSF at the position of star. That is, the shape and the size of the PSF is known and the only unknown variables are the coordinates of the centroids of stars. Note that in practice however, size and shape of the PSF are also estimated along with the centroid. In the following, we assume that the size and shape of the PSF are known. For the last two methods, we do not use any information about the PSF.

1.4.1 Centroiding by fitting a correct PSF profile

We examine fitting an exact PSF profile to the stars. That is, in our Cramér-Rao bound saturation tests, we find the best estimates of flux and centroid by maximizing the likelihood

using the correct PSF model. In the model, the size of the Moffat PSF is assumed to be correct. We expect this method to perform best in determining the centroids of stars, and deliver RMSE equal to Cramér-Rao bound.

1.4.2 Matched-filter polynomial centroiding

Let us consider the case in which we have a good estimate of the pixel-convolved PSF at the position of the faint star under consideration. We can smooth the image of the star, by correlating it with the full PSF $\mathcal{P}$ at the position of the star.

$$Y^{(s)} = Y \star \mathcal{P}, \tag{1.18}$$

$$Y^{(s)}_{[i,j]} = \sum_{k,l} Y_{[i-k,j-l]} \mathcal{P}_{[k,l]}, \tag{1.19}$$

where Y is the image of the star, and $Y^{(s)}$ is sometimes called a matched filter. A matched filter is a method in which the data Y is correlated (convolved in the case of symmetrical PSF) with the PSF $\mathcal{P}$. It is equivalent to optimizing the likelihood and therefore provides an optimal map where the peak of the map is the likely position of the point source (Lang *et al.*, in preparation).

Then, we fit a simple 2d second-order polynomial $P(x, y) = a + bx + cy + dx^2 + exy + fy^2$ to the 3×3 patch centered on the brightest pixel of the matched-filter image Y^s. Upon

constructing a universal 9×6 design matrix

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & y_1 & x_1^2 & x_1 y_1 & y_1^2 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_9 & y_9 & x_9^2 & x_9 y_9 & y_9^2 \end{bmatrix}, \tag{1.20}$$

the free parameters $\{a, b, c, d, e, f\}$ (hereafter compactly denoted by $\mathbf{X}$) can be determined by

$$\mathbf{X} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Z}, \tag{1.21}$$

where $\mathbf{Z}$ is given by $(z_1, ..., z_9)^T$, with z_i, being the brightness of the $i-$th pixel of the 3×3 patch centered on the brightest pixel of $Y^{(s)}$. Afterwards, the best fit parameters can be used to compute the centroid coordinate

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \begin{bmatrix} 2d & e \\ e & 2f \end{bmatrix}^{-1} \begin{bmatrix} -b \\ -c \end{bmatrix}. \tag{1.22}$$

It is important to note that the algebraic operation in (1.22) involves inverting a 2×2 curvature matrix

$$D = \begin{bmatrix} 2d & e \\ e & 2f \end{bmatrix}. \tag{1.23}$$

When the curvature matrix D has a zero (or very close to zero) deteminant, centroid estimates obtained from equation (1.22) can become arbitrarily large, which leads to catastrophic outliers. In order to tackle this issue, we add a soft regularization term proportional to σ to the diagonals of D prior to inversion.

The procedure of convolving the image of star with the PSF results in a smoother image. Therefore, a simple second-order polynomial will provide a better fit since convolution with the PSF makes the variation of the brightness of the image across the 3×3 patch very smooth.

1.4.3 Fixed-Gaussian polynomial centroiding

In the case that we do not know the PSF at the position of star, we change the smoothing step in the following way. Instead of smoothing the image by convolving it with the PSF, smoothing is done by convolving the image with a fixed Gaussian kernel with a fixed size

$$k(\mathbf{x}) = \frac{1}{2\pi w^2} \exp(-\mathbf{x}^2/2w^2), \tag{1.24}$$

where throughout this study, the full-width at half-maximum of the Gaussian kernel is held at a fixed value of 2.8 pixels (corresponding to $w \simeq 1.2$ pixels). The smoothing step is done as follows

$$Y^{(s)} = Y \star \mathcal{K}, \tag{1.25}$$

$$Y^{(s)}_{[i,j]} = \sum_{k,l} Y_{[i-k,j-l]} \mathcal{K}_{[k,l]}, \tag{1.26}$$

where Y is the image of the star, $Y^{(s)}$ is the smooth image, and $\mathcal{K}$ is an array whose elements are given by the Gaussian kernel

$$\mathcal{K}_{[k,l]} = k(x_k, y_l). \tag{1.27}$$

Note that the size of the kernel $\mathcal{K}$ is equal to the size of the kernel $\mathcal{P}$ used in the matched-filter polynomial centroiding. Then we apply the same 2d second-order polynomial method (see (1.21), (1.22)) to the 3×3 patch centered on the brightest pixel of the smooth image

$Y^{(s)}$. Therefore, for a given star and a smoothing kernel, the outcome of equation (1.21) can be plugged into equation (1.22) to find the centroid estimate of the star. This is inspired by the 3×3 quartic approximation used in the *Sloan Digital Sky Surveys* photometric pipeline (**?**).

1.4.4 Center-of-light centroiding

In addition to the fitting methods mentioned so far, we examine centroiding stars by computing their first moments in a 7×7 patch around the brightest pixel of the image.

$$x_c = \frac{\sum_m x_m Y_m}{\sum_m Y_m}, \tag{1.28}$$

$$y_c = \frac{\sum_m y_m Y_m}{\sum_m Y_m}, \tag{1.29}$$

where the summation is done over all the pixels of the 7×7 patch, and x_m, y_m, and Y_m, are the x coordinate, y coordinate, and the brightness of pixel m respectively.

In terms of saturating the Cramér-Rao lower bound, we expect this simple center-of-light centroiding to perform worse than all other methods mentioned in this section. Hereafter, we call this method 7×7 moment centroiding.

1.5 Tests

We perform two sets of simulations. In the first set, we choose four values of 2, 2.8, 4, and 5.6 pixels for γ. For each γ, we generate 100,000 17×17 postage-stamps of Moffat profiles with centroids randomly drawn within the central pixel of the 17×17 postage-stamps. Moreover, zero-mean uncorrelated Gaussian noise is added to each postage-stamp such that

the simulated stars are uniformly distributed in log-SNR between SNR = 5 to SNR = 100.

In the second set, we generate 100,000 17×17 postage-stamps of Moffat profile, with values of γ uniformly distributed between 2 and 6 pixels, and with centroids drawn randomly within the central pixel. We choose four values for SNR: 5, 10, 20, and 40. For each SNR, and for each postage-stamp with a given γ, zero-mean uncorrelated Gaussian noise, with standard deviation corresponding to SNR and γ through equation (1.15), is added to each postage-stamp.

In the first experiment, we study how the centroiding error behaves with changing SNR, while γ is held constant. In the second experiment, we study how the centroiding error behaves with changing γ while SNR is held constant.

1.6 Results

1.6.1 Experiment 1 : variable SNR; constant γ

In this experiment, after finding the centroiding errors for each method, we compute the RMSE in bins of SNR in order to compare it to the CRLB. Results of the first experiment are shown in Figures 1.1, 1.2, 1.3, 1.4. Note that the centroid errors, the CRLB, and the RMSE values shown in these figures are computed for only one component. As we expected, the RMSE from centroiding by fitting the exact PSF model (Figure 1.1) lies on the CRLB.

Figure 1.2 demonstrates that even the matched filter polynomial centroiding is able deliver centroiding estimates as efficient as the PSF fitting method in terms of saturating the bound for the simulated stars with $\gamma = 2.8, 4, 5.6$ pixels. For stars with $\gamma = 2$ pixels, although this method gets very close to saturating the CRLB, the RMSE arising from this method shows slight deviations from the CRLB since the images of stars are not sufficiently smooth even after correlation of these images with the PSF. For simulated images with higher γ,

convolving the data with the PSF results in images that are smooth around the brightest pixel. This enables the polynomial centroiding to deliver estimates that can saturate the CRLB.

The RMSE from the fixed-Gaussian polynomial centroiding (Figure 1.3), is very close to the CRLB. As we increase γ from 2 pixels to 2.8 pixels, RMSE approaches the CRLB. For stars with $\gamma = 2$ pixels, the rate at which the RMSE from this method drops eventually becomes smaller than the constant rate at which the CRLB decreases with increasing SNR. The reason for this is that even after smoothing the data with a Gaussian kernel, the images are not smooth enough for a second-order polynomial fitting to deliver estimates with RMSE close to the bound. For stars with $\gamma = 2.8$ pixels, a significant fraction of information is in the 3×3 patch of the smooth image and this method is able to saturate the bound. When we increase γ to 4 and 5.6 pixels, the mismatch between the width of the Gaussian kernel and the PSF increases and the RMSE deviates from the CRLB. The deviation is largest for the simulated stars with $\gamma = 5.6$ pixels.

On the other hand, Figure 1.4 shows that in case of 7×7 moment method, the RMSE becomes quite large as we move toward fainter stars in our simulation. For stars with larger γ, centroid estimates from the naive center-of-light centroiding do not even come close to saturating the CRLB. As γ increases, the RMSE deviates further from the CRLB.

1.6.2 Experiment 2 : constant SNR; variable γ

In this experiment, after finding the centroiding errors for each method, we compute the RMSE in bins of γ in order to compare it to the CRLB. Behavior of error as a function of γ for different values of SNR, is shown in Figures 1.5, 1.6, 1.7, and 1.8. Note that the centroid errors, the CRLB, and the RMSE values shown in these figures are computed for only one component.

Once again, the RMSE from centroiding by fitting the exact PSF model as a function of FWHM lies on the CRLB (see Figure 1.5). Thus, centroid estimates from fitting the exact PSF model always saturate the CRLB. Once again, we observe that the centroid estimates found by the matched filter polynomial centroiding saturate the CRLB with the exception of simulated stars with γ very close to 2 pixels (see Figure 1.6).

Figure 1.7 illustrates that the fixed-Gaussian polynomial method results in RMSE very close to the CRLB. For all four values of SNR, as we increase γ from 2 pixels to 3 pixels, the RMSE gets closer to the CRLB since the method starts to perform slightly better as we move away from undersampled stars and as the FWHM of the smoothing kernel gets closer to that of the simulated images. After approximately 3 pixels, increasing γ results in deviation of the RMSE of the method from the CRLB. This is a characteristic of the fixed-Gaussian polynomial method as we apply it to a smooth image in which some fraction of the available information is lost in the 3×3 patch around the brightest pixel. Furthermore, increasing the SNR from 5 to 40 makes the RMSE (as a function of γ) closer to the CRLB. In the case of extremely faint stars (SNR = 5), the fixed-Gaussian polynomial centroiding fails to saturate the bound.

The centroid estimates obtained from the naive 7×7 moment method (see Figure 1.8) result in RMSE much larger than the CRLB in all ranges of FWHM and for all four values of SNR in this experiment.

1.7 Discussion

An efficient stellar centroiding algorithm must saturate—or come close to saturating—the fundamental Cramér-Rao lower bound. That is, in all ranges of background noise level, size, radial light profile, and shape, it must preserve information about the centroids of stars.

In practice however, this is only achievable when we have a reasonably good estimate of the PSF. Since we do not always know the exact PSF profile, we must make use of approximate centroiding algorithms. In this work, we studied how close we get to saturating the CRLB with approximate methods acting on relatively low signal-to-noise ratio unsaturated stars.

We focused on examples from two classes of centroiding algorithms. The first class contains fast and approximate methods that do not require any knowledge of the PSF at the positions of stars. Of methods that belong to this class, we consider centroiding stars based on fitting a second-order polynomial to a 3×3 patch of star images smoothed by a Gaussian kernel of fixed width, and finding the center of light of a 7×7 patch around the brightest pixel of the star.

The second class of centroiding algorithms make use of the PSF (or some good estimate of the PSF) at the positions of stars. In our investigation, it is assumed that the size and the shape of the PSF are known prior to applying these algorithms to the images of stars. We considered two examples from this class. The first example is the matched filter polynomial centroiding, and the second example is the PSF fitting. In the PSF fitting method, we find the maximum likelihood estimates of the flux and centroids of stars by fitting a PSF model that has the correct shape and size.

In terms of saturating the Cramér-Rao bound, we compared the performances of these methods against each other. Our results suggest that in all ranges of FWHM and SNR, the PSF fitting method returns centroid estimates that saturate the CRLB. This confirms our expectation that maximum-likelihood estimators saturate the Cramér-Rao lower bound.

We note that the estimates found by the 7×7 moment method, except in the case of very high SNR values and small values of γ, do not come close to saturating the CRLB. In a considerable range of PSF sizes and background noise levels, this method fails to deliver any centroiding estimate close to saturating the bound. When applied to stars with $\gamma = 2.8,\ 4$

pixels, we find deviation of RMSE from the CRLB as large as $600\% - 800\%$ below signal-to-noise ratio of 10. For the simulated stars with $\gamma = 5.6$ pixels, we find deviations as large as 500% for SNR below 10 and as large as 200% for SNR ~ 100. It can be noted in Figure 1.8 that in the simulations with the lowest SNR (SNR ~ 5), the errors arising from the 7×7 moment method are suppressed by the fact that 17×17 postage-stamps are used to simulate images. Therefore in the case of SNR ~ 5, we expect the deviation of the RMSE from the CRLB to be larger for this method.

On the other hand, the RMSE of centroid estimates of the fixed-Gaussian polynomial centroiding are much closer to saturating the CRLB in all ranges of signal-to-noise ratio even though this method does not require knowledge of the PSF at the positions of stars. We note that when the FWHM of the stars are close to 2.8 pixels (the FWHM of the Gaussian kernel), the fixed-Gaussian polynomial method saturates the CRLB. Deviation of the RMSE of this method from the CRLB is larger for the simulated stars with larger values of FWHM ($\gamma \simeq 5$ pixels). Presence of noise is another limiting factor. Although this method is able to get very close to saturating the bound in a wide range of signal-to-noise ratios, it is not reliable in the case of centroiding extremely faint stars ($5 < \mathrm{S/N} < 10$).

In matched filter polynomial centroiding, the fixed-Gaussian polynomial method is modified by convolving the image with the correct PSF. Our results on the simulated stars show that the matched filter estimator saturates the CRLB for all PSF sizes and noise levels. This is due to the fact that once the images of stars are convolved with the correct PSF, they become smooth that fitting a second-order polynomial to the 3×3 patch centered on the brightest pixel of the smooth image is sufficient for us to obtain results as accurate as those from fitting a PSF profile.

The Gaussian kernel (see equation 1.24) in the fixed-Gaussian polynomial centroiding is separable, and correlation of the kernel with an image of star can be performed *exactly* in

no time. Therefore in terms of computational cost, this method is more efficient than the matched filter method in which the image of star is correlated with a PSF of arbitrary shape.

In the case that we have a good estimate of the PSF, the matched filter polynomial method can be faster than PSF fitting method for *centroiding purposes*. Additionally, this method is able to saturate the CRLB in a wide range of conditions. It is however important to note that in many cases, reliable estimation of the flux requires a technique as accurate as PSF-fitting. However, in cases in which an investigator only needs an empirical estimates of the centroid offsets, fixed-Gaussian polynomial or matched filter polynomial centroiding methods can be employed with negligible loss of information. For instance in modeling the stellar light curves in the *K2* mission, **?** uses a simple polynomial centroiding to marginalize out the systematic trends caused by centroid offsets.

Moreover, we note that the PSF fitting method can be made faster by only keeping the term proportional to the dot product of the PSF model and the image in χ^2:

$$\chi^2 = \big(\mathbf{y}\cdot\mathbf{y} - 2F\mathbf{y}\cdot\mathbf{m} + F^2\mathbf{m}\cdot\mathbf{m}\big)/\sigma^2, \tag{1.30}$$

where F is the flux, σ is the per-pixel uncertainty, the dot product between two vectors is denoted by $(\cdot)$, and the image of star and the normalized shifted PSF model are denoted by $\mathbf{y}$ and $\mathbf{m}$ respectively. Upon varying only the centroid, the terms $\mathbf{y}.\mathbf{y}$ and $\mathbf{m}.\mathbf{m}$ remain approximately constant. However, this only allows us to vary the position of centroid, and not the flux, while fitting the PSF model to the star.

Finding a centroid coordinate that maximizes the dot product of the PSF and the star image is equivalent to finding the peak of the correlation of the PSF and the image. Therefore optimizing the modified χ^2 is equivalent to finding the location of the peak of the matched filter.

In the initial smoothing step of the fixed-Gaussian polynomial method, the image of the star is correlated with an approximate Gaussian PSF. When there is mismatch between the widths of the smoothing kernel and the that of the PSF, we loose some information by employing a 3×3 polynomial fitting. When we have the advantage of knowing the PSF, this issue can be resolved by employing the matched filter polynomial method.

In this investigation we showed that PSF fitting always performs better—in terms of saturating the CRLB—at centroiding stars. Having a reasonable PSF model always helps us obtain more reliable centroid estimates, but over a certain range of low signal-to-noise ratios and PSF sizes, one can achieve sensibly accurate results by employing a simple 3×3 method after smoothing the image with a Gaussian kernel of a fixed width, and without making any assumption about the PSF model at the positions of stars.

In this investigation we narrowed our focus on a set of data simulated from a particular PSF profile. Although there are various cases where Moffat profiles provide reasonable representations of the point spread function, these profiles are not generic enough to let us reach a more general conclusion.

This work was partially supported by the NSF (grants IIS-1124794 and AST-1517237), NASA (grant NNX12AI50G), and the Moore-Sloan Data Science Environment at NYU. We thank Jo Bovy and Alex Malz for discussions related to this work. We are also grateful to Dustin Lang and Alex Malz for reading and making comments on draft.

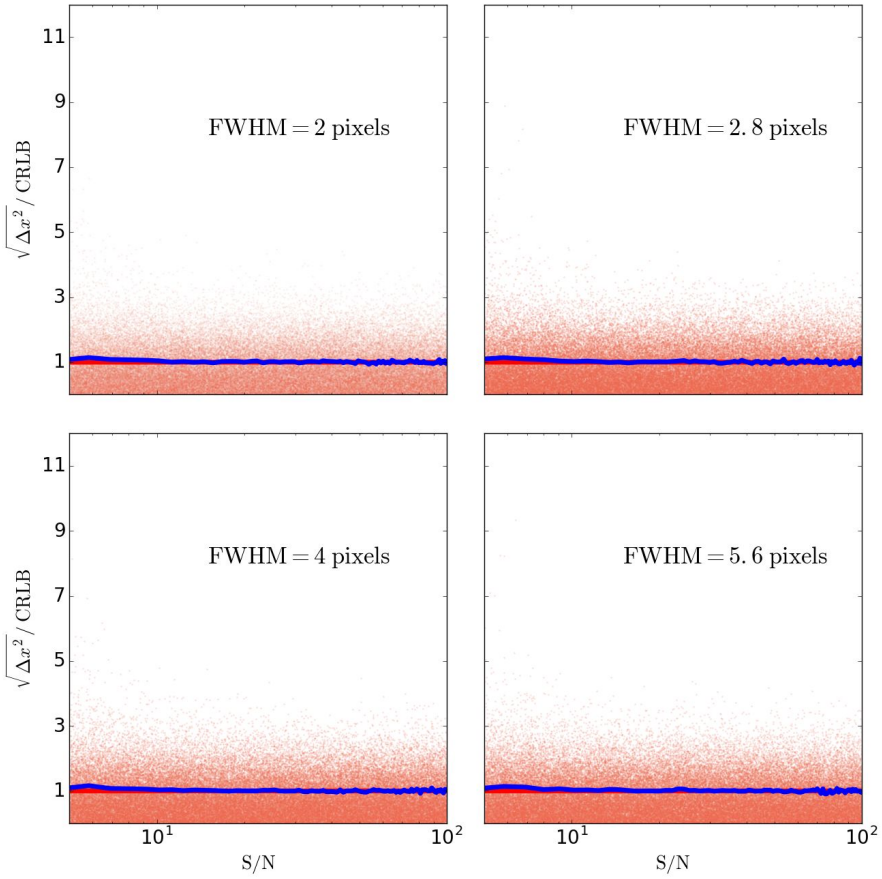

Figure 1.1: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the signal-to-noise ratio of stars. Errors are found from fitting the exact PSF model to the stars, with FWHM of : 2 (upper left), 2.8 (upper right), 4 (lower left), and 5.6 (lower right) pixels. In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

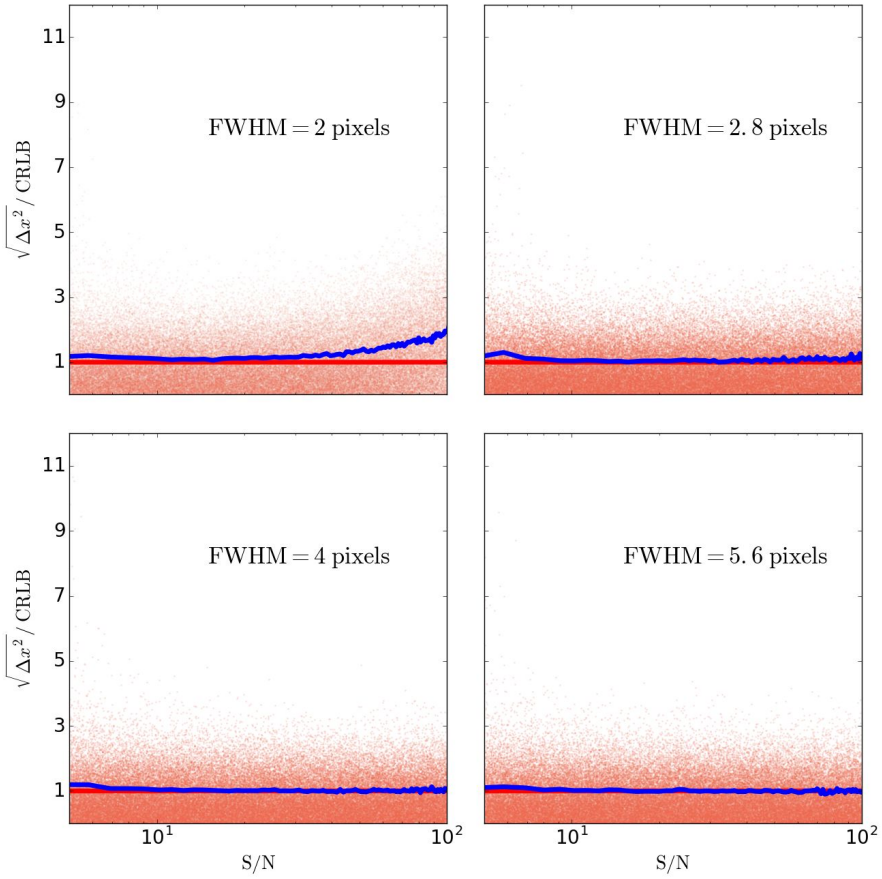

Figure 1.2: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the signal-to-noise ratio of stars. Errors are found from applying the matched filter polynomial centroiding to the stars, with FWHM of : 2 (upper left), 2.8 (upper right), 4 (lower left), and 5.6 (lower right) pixels. In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

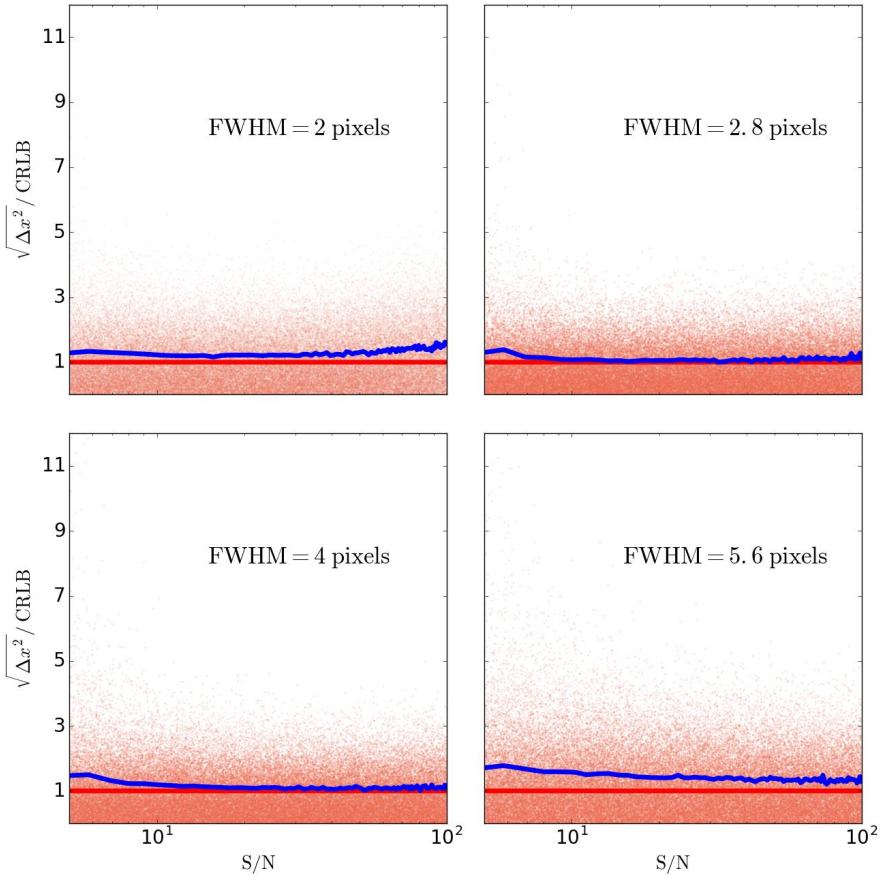

Figure 1.3: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the signal-to-noise ratio of stars. Errors are found from applying the fixed-Gaussian polynomial centroiding to the stars, with FWHM of : 2 (upper left), 2.8 (upper right), 4 (lower left), and 5.6 (lower right) pixels. In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

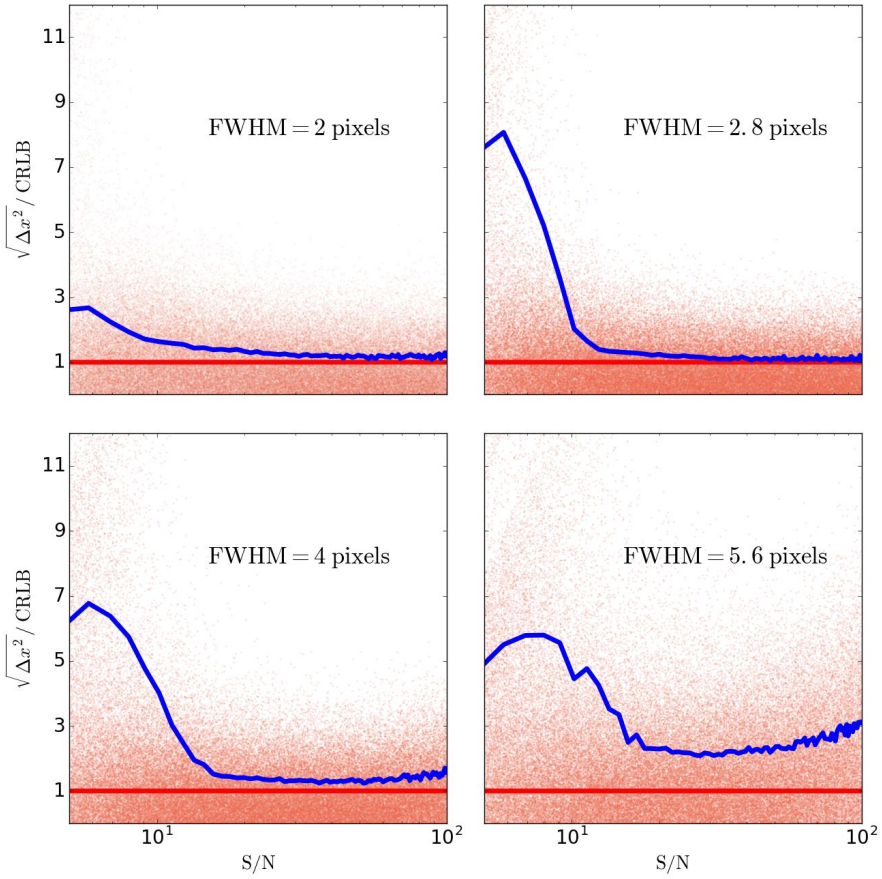

Figure 1.4: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the signal-to-noise ratio of stars. Errors are found from applying the 7×7 moment method to the stars, with FWHM of : 2 (upper left), 2.8 (upper right), 4 (lower left), and 5.6 (lower right) pixels. In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

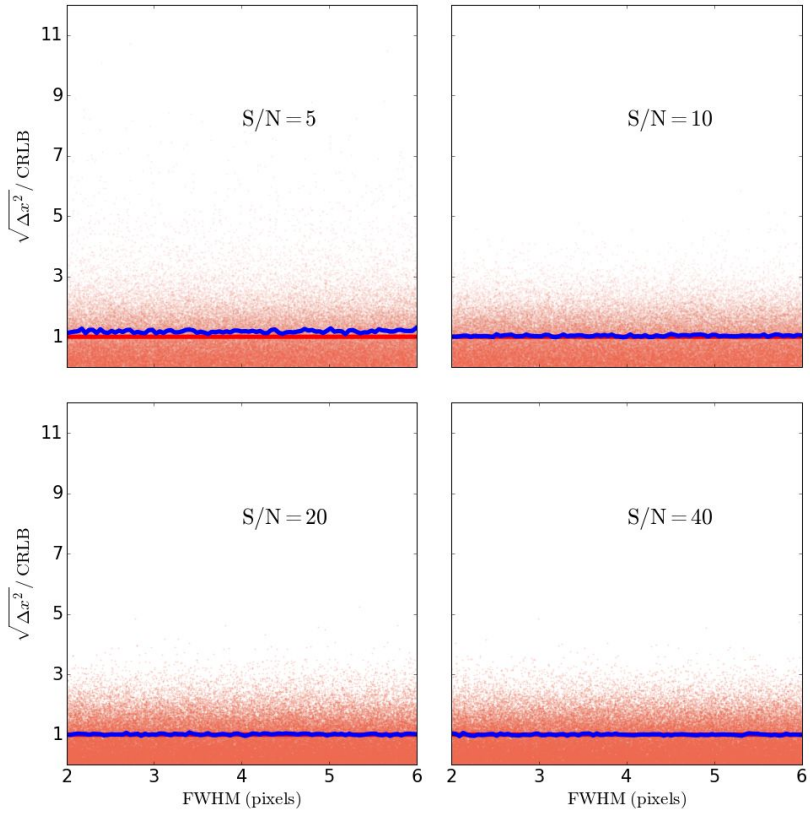

Figure 1.5: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the FWHM of stars. Errors are found from fitting the exact PSF model to the stars, with SNR of : 5 (upper left), 10 (upper right), 20 (lower left), and 40 (lower right). In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

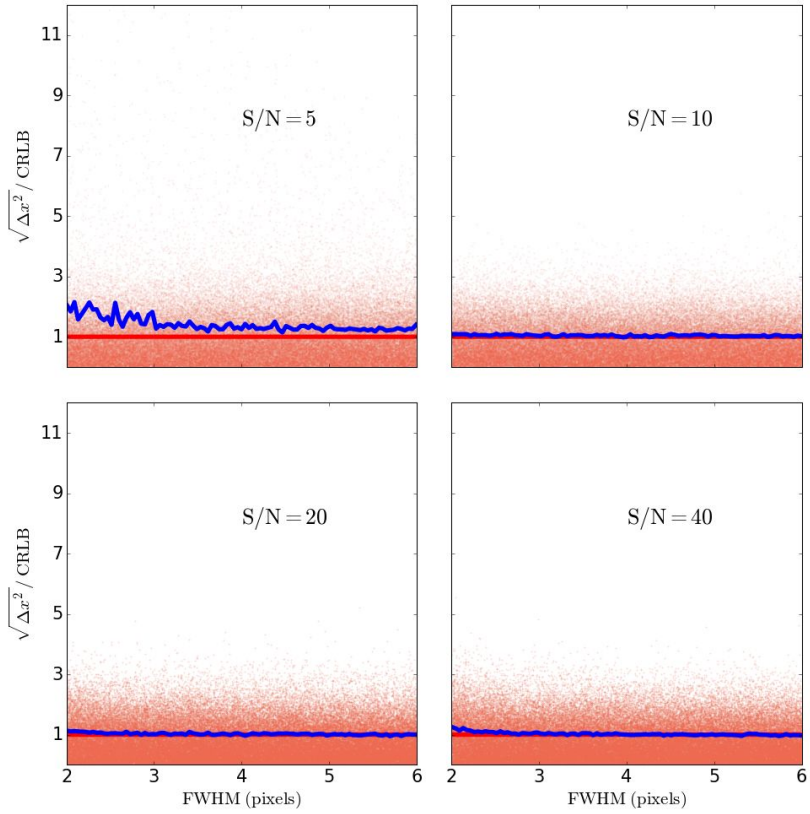

Figure 1.6: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the FWHM of stars. Errors are found from applying the matched filter polynomial centroiding to the stars, with SNR of : 5 (upper left), 10 (upper right), 20 (lower left), and 40 (lower right). In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

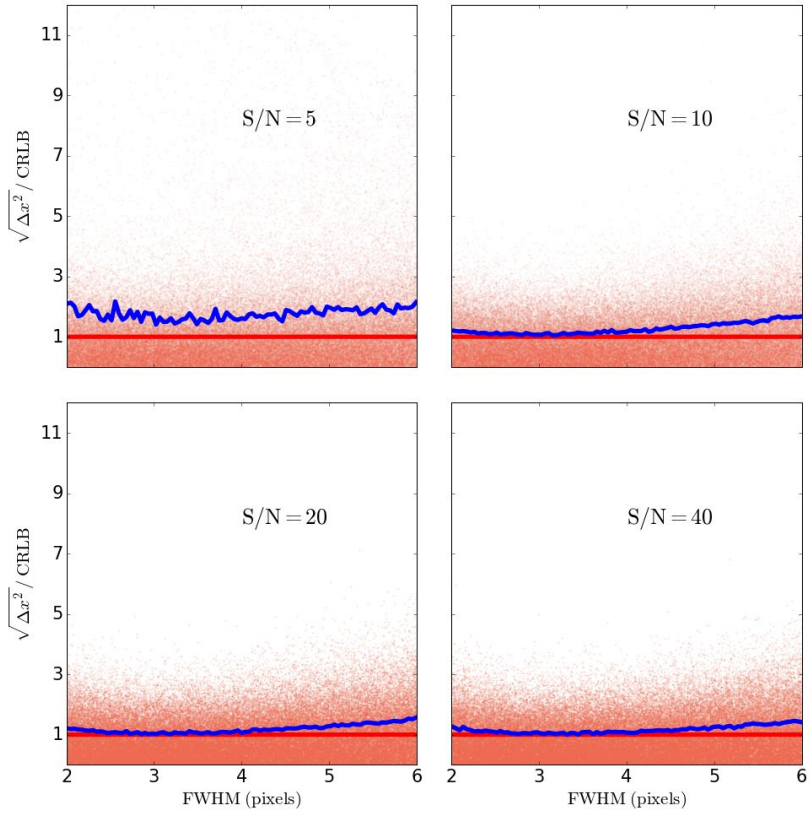

Figure 1.7: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the FWHM of stars. Errors are found from applying the fixed-Gaussian polynomial centroiding to the stars, with SNR of : 5 (upper left), 10 (upper right), 20 (lower left), and 40 (lower right). In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

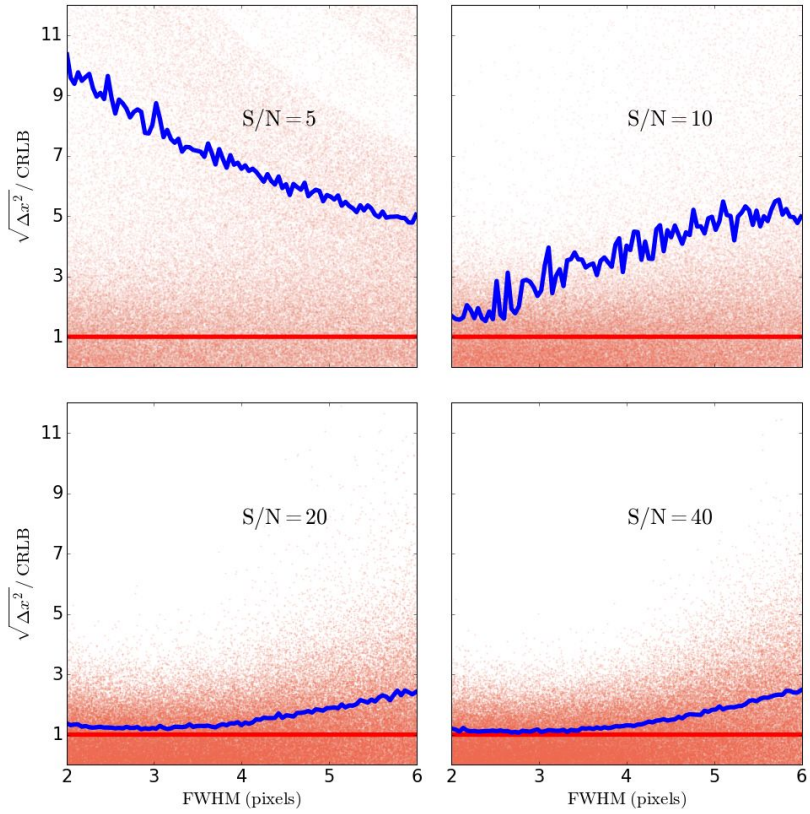

Figure 1.8: Scatter plots showing the relation between the ratio of error (in x-axis of the centroid poistions) to the CRLB and the FWHM of stars. Errors are found from applying the 7×7 moment method to the stars, with SNR of : 5 (upper left), 10 (upper right), 20 (lower left), and 40 (lower right). In each scatter plot, the blue solid line represents the ratio of the root-mean-squared-error to the CRLB, and the red line represents the ratio achievable by an optimal estimator.

Chapter 2

Super-resolution PSF model of HST WFC3-IR

Chapter 3

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

This Chapter is joint work with ChangHoon Hahn (NYU), Mohammadjavad Vakili (NYU), Kilian Walsh (NYU), Andrew P. Hearin (Yale), David W. Hogg (NYU), and Duncan Campbell (Yale) submitted to the *Monthly Royal Astronomical Society notice* as **?**.

3.1 chapter abstract

The standard approaches to Bayesian parameter inference in large scale structure (LSS) assume a Gaussian functional form (chi-squared form) for the likelihood. They are also typically restricted to measurements such as the two point correlation function. Likelihood free inferences such as Approximate Bayesian Computation (ABC) make inference possible without assuming any functional form for the likelihood, thereby relaxing the assumptions

and restrictions of the standard approach. Instead it 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 halos 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.

3.2 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 (**??**). 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 (**?**, **?**, **?**, **??**).

Fundamentally, the tractability of likelihood functions in cosmology flows from the fact

that the initial conditions are exceedingly close to Gaussian in form (**??**), and that many sources of measurement noise are also Gaussian (**??**). 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 nonlinear gravitational evolution and biasing, the likelihood *cannot* be Gaussian. Even if the initial conditions are perfectly Gaussian, the growth of structure creates non-linearities which are non-Gaussian (see **?** for a comprehensive review). Galaxies form within the density field in some complex manner that is modeled only effectively (**????**; see **?** for a recent review). Even if the galaxies were a Poisson sampling of the density field, which they are not (**???**), it would be tremendously difficult to write down even an approximate likelihood function (**?**).

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 the present work, we use Approximate Bayesian Computation—ABC—which provides a *rejection sampling* framework 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; **?**). 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 (**?**). 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 (**?**), the study of the morphological

properties of galaxies at high redshifts (**?**), stellar initial mass function modeling (Cisewski et al. in preparation), and cosmological inference with with weak-lensing peak counts (**??**), Type Ia Supernovae (**?**), and galaxy cluster number counts (**?**).

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 halos are biased tracers of the underlying cosmic density field (**???**). The HOD specifies how the dark matter halos are populated with galaxies by modeling the probability that a given halo hosts N galaxies subject to some observational selection criteria (**?????**). This statistical prescription for connecting galaxies to halos has been remarkably successful in reproducing the galaxy clustering, galaxy–galaxy lensing, and other observational statistics (**??**), and is a useful framework for constraining cosmological parameters (**????**) as well as galaxy evolution models (**?????**, Walsh et al. 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 catalogs obtained from high resolution N-body simulations (**??**) using `Halotools`[1] (**?**), an open-source package for modeling 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 5.3 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 3.4.1, we discuss the mock galaxy catalog,

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

which we treat as observation. With the specific choices of ABC-PMC ingredients, which we describe in Section 3.4.2, in Section 3.4.3 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 3.4.4. Finally, we discuss and conclude in Section 5.5.

3.3 Methods

3.3.1 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 **?** and refernces therein).

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

$$p(\vec{\theta}, X|D) = \frac{p(D|X)p(X|\vec{\theta})\pi(\vec{\theta})}{\mathcal{Z}} \tag{3.1}$$

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

$$\mathcal{Z} = \int d\vec{\theta}\, dX\; p(D|X)p(X|\vec{\theta})\pi(\vec{\theta}), \tag{3.2}$$

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

$$q(\vec{\theta}, X) = p(X|\vec{\theta})\pi(\vec{\theta}) \tag{3.3}$$

and the choice of

$$M = \frac{\max p(D|X)}{\mathcal{Z}} > 1. \tag{3.4}$$

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

$$p(\vec{\theta}, X|D) < Mq(\vec{\theta}, X) \tag{3.5}$$

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

$$\frac{p(\vec{\theta}, X|D)}{Mq(\vec{\theta}, X)} = \frac{p(D|X)}{\max p(D|X)} > u \tag{3.6}$$

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

$$p(D|X) \propto p(\rho(D, X) < \epsilon) \tag{3.7}$$

where $\rho(D, X)$ is the distance between the data D and simulation X. Eq. 3.7 along with the rejection sampling acceptance criteria (Eq. 3.6), leads to the acceptance criteria for ABC: $\vec{\theta}$ 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 Eq. 3.7 will also be a vector with the same dimensions. $\vec{\theta}$ 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 $\vec{\theta}$ from the prior distribution $\pi(\vec{\theta})$. The simulation is generated from $\vec{\theta}$ using the forward model, $X = f(\vec{\theta})$. Then the distance between the data and simulation, $\vec{\rho}(D, X)$, is calculated and compared to $\vec{\epsilon}$. If $\vec{\rho}(D, X) < \vec{\epsilon}$, $\vec{\theta}$ is accepted. This process is repeated until we are left with a sample of $\vec{\theta}$ that all satisfy the distance criteria. This final ensemble approximates the posterior probability distribution $p(\vec{\theta}, X|D)$.

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 $\pi(\vec{\theta})$. 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 $\vec{\theta}$ from $\pi(\vec{\theta})$ 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 Popula-

tion Monte Carlo (PMC) algorithm as our sampling technique. PMC is an iterative method that performs rejection sampling over a sequence of $\vec{\theta}$ distributions ($\{p_1(\vec{\theta}), ..., p_T(\vec{\theta})\}$ 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**
      $\vec{\theta}_t^* \leftarrow \pi(\vec{\theta})$
      $X = f(\vec{\theta}_t^*)$
    end while
    $\vec{\theta}_t^{(i)} \leftarrow \vec{\theta}_t^*$
    $w_t^{(i)} \leftarrow 1/N$
  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 $\vec{\theta}_t^*$ from $\{\vec{\theta}_{t-1}\}$ with probabilities $\{w_{t-1}\}$
      $\vec{\theta}_t^* \leftarrow K(\vec{\theta}_t^*, .)$
      $X = f(\vec{\theta}_t^*)$
    end while
    $\vec{\theta}_t^{(i)} \leftarrow \vec{\theta}_t^*$
    $w_t^{(i)} \leftarrow \pi(\vec{\theta}_t^{(i)}) / \big(\sum_{j=1}^{N} w_{t-1}^{(i)} K(\vec{\theta}_{t-1}^{(j)}, \vec{\theta}_t^{(i)}) \big)$
  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 $\vec{\theta}$ (hereafter referred to as particles) from the prior distribution $\pi(\vec{\theta})$. We forward model the simulation $X = f(\vec{\theta})$, calculate the distance $\rho(D, X)$, compare this distance to ϵ_1, and then accept or reject the $\vec{\theta}$ 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 **?**, for instance, the threshold of the second iteration is set to the 25^{th} percentile of the distances in the first iterations; afterwards in the subsequent iterations, t, ϵ_t is set to the 50^{th} percentile of the distances in the previous $t-1$ iteration. Alternatively, **?** set ϵ_t to the median of the distances from the previous iteration. In Section 3.4, we describe our prescription for the distance threshold, which follows **?**.

Once ϵ_t is set, we draw a particle from the previous weighted set of particles $\vec{\theta}_{t-1}$. This particle is perturbed by a kernel, set to the covariance of $\vec{\theta}_{t-1}$. Then once again, we generate a simulation by forward modeling $X = f(\vec{\theta}^i)$, 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 $\vec{\theta}_t$. 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 centered 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 **?**.

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 criteria, 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 3.4.2 we detail the specific convergence criteria used in our analysis.

3.3.2 Forward model

3.3.2.1 Halo Occupation Modeling

ABC requires a forward generative model. In large scale structure studies, this implies a model that is able to generate a galaxy catalog. We then calculate and compare summary statistics of the data and model catalog in an identical fashion In this section, we describe the forward generative model we use within the framework of the halo occupation distribution.

The assumption that galaxies reside in dark matter halos is the bedrock underlying all contemporary theoretical predictions for galaxy clustering. The Halo Occupation Distribution (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:

$$1 + \xi^{1h}_{\rm gg}(r) \simeq \frac{1}{4\pi r^2 \bar{n}_{\rm g}^2} \int {\rm d}M_{\rm h} \frac{{\rm d}n}{{\rm d}M_{\rm h}} \Xi_{\rm gg}(r|M_{\rm h}) \times \langle N_{\rm g}(N_{\rm g}-1)|M_{\rm h}\rangle \,, \tag{3.8}$$

and

$$\xi^{2h}_{\rm gg}(r) \simeq \xi_{\rm mm}(r) \left[\frac{1}{\bar{n}_{\rm g}} \int {\rm d}M_{\rm h} \frac{{\rm d}n}{{\rm d}M_{\rm h}} \langle N_{\rm g}|M_{\rm h}\rangle \, b_{\rm h}(M_{\rm h})\right]^2 \tag{3.9}$$

In Eqs. (3.8) and (3.9), $\bar{n}_{\rm g}$ is the galaxy number density, ${\rm d}n/\,{\rm d}M_{\rm h}$ is the halo mass function, the spatial bias of dark matter halos 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, Eqs. (3.8) and (3.9) 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% level (**?**). Direct emulation methods have made significant improvements in precision and accuracy in recent years (**??**); however, a labor- and computation-intensive interpolation exercise must be carried out each time any alternative statistic is explored, which is one of the goals of the present 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 halos with mock galaxies and then calculate our summary statistics directly on the resulting galaxy

catalog with the same estimators that are used on observational data (**?**). Additionally, through our forward modeling 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 **?**. The occupation statistics of central galaxies follow a nearest-integer distribution with first moment given by

$$\langle N_{\rm cen} \rangle = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\log M}}\right)\right]. \tag{3.10}$$

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

$$\langle N_{\rm sat} \rangle = \langle N_{\rm cen} \rangle \left(\frac{M - M_0}{M_1}\right)^{\alpha}. \tag{3.11}$$

We assume that central galaxies are seated at the exact center of the host dark matter halo and are at rest with respect to the halo velocity, defined according to `Rockstar` halo finder (**?**) as the mean velocity of the inner 10% of particles in the halo. Satellite galaxies are confined to reside within the virial radius following an NFW spatial profile (**?**) with a concentration parameter given by the $c(M)$ relation (**?**). 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 (**?**). We refer the reader to **?**, **?**, and `http://halotools.readthedocs.io` for further details.

For the halo catalog of our forward model, we use the publicly available `Rockstar` (**?**) halo

catalogs of the MultiDark cosmological N-body simulation (**?**).[2] MultiDark is a collision-less 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; **?**) run on 2048^3 particles in a 1 h^{-1}Gpc periodic box. MultiDark particles have a mass of $m_p \simeq 8.72 \times 10^8\ h^{-1}M_\odot$; 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 halos 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 $\{\texttt{BOX1}, ..., \texttt{BOX125}\}$. The first subvolume is reserved to generate the mock observations which we describe in Section 3.4.1. When we simulate a galaxy catalog for a given $\vec{\theta}$ in parameter space, we randomly select one of the subvolumes from $\{\texttt{BOX2}, ..., \texttt{BOX125}\}$ and then populate the halos within this subvolume with galaxies. We implement this procedure to account for sample variance within our forward generative model.

3.3.3 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

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

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

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

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

$$\xi(r) = \frac{DD}{RR} - 1, \tag{3.12}$$

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

- Galaxy group multiplicity function, $\zeta_{\rm g}(N)$: 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 (**?**) 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 analyze the galaxy groups in the SDSS main galaxy sample (**?**). For details regarding the group finding algorithm, we refer readers to **?**.

  In this study we set the linking length to be 0.25 times the mean separation of galaxies which is given by $\bar{n}_{\rm g}^{-1/3}$. 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 $\zeta_g(N)$ — in units of $(\mathrm{Mpc}/h)^{-3}$.

3.4 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 3.4.1 by creating our "observation". We select a set of HOD parameters which we deem as the "true" parameters and run it through our forward model producing a catalog of galaxy positions which we treat as our observation. Then, in Section 3.4.2, 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 3.4.3, 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 3.4.4.

3.4.1 Mock Observations

In generating our "observations", and more generally for our forward model, we adopt the HOD model from **?** where the expected number of galaxies populating a dark matter halo is governed by Eqs (3.10) and (3.11). For the parameters of the model used to generate the fiducial mock observations, we choose the **?** best-fit HOD parameters for the SDSS main galaxy sample with a luminosity threshold $M_r = -21$:

$\log M_{\min}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	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 3.3.2), we generate a catalog of galaxy positions.

For our summary statistics of the catalogs we use: the mean number density $\bar{n}_{\rm g}$, the galaxy two-point correlation function $\xi_{\rm gg}(r)$, and the group multiplicity function $\zeta_{\rm g}(N)$. Our mock observation catalog has $\bar{n}_{\rm g} = 9.28875 \times 10^{-4}\ h^{-3}{\rm Mpc}^3$ and in Figure 3.1 we plot $\xi_{\rm gg}(r)$ (left panel) and $\zeta_{\rm g}(N)$ (right panel). The width of the shaded region represent the square root of the diagonal elements of the summary statistic covariance matrix, which is computed as we describe below.

We calculate $\xi_{\rm gg}$ using the natural estimator (Section 3.3.3) with fifteen 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 $\zeta_{\rm g}(N)$ as described in Section 3.3.3 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 catalog subvolumes: $\{\texttt{BOX1}, ..., \texttt{BOX125}\}$. We compute the summary statistics of each subvolume galaxy sample k:

$$\mathbf{x}^{(k)} = [\bar{n}_{\rm g},\ \xi_{\rm gg},\ \zeta_{\rm g}], \tag{3.13}$$

Then we compute the covariance matrix as

$$\mathrm{C}_{i,j}^{\rm sample} = \frac{1}{N_{\rm mocks} - 1} \sum_{k=1}^{N_{\rm mocks}} \Big[\mathbf{x}_i^{(k)} - \overline{\mathbf{x}}_i\Big]\Big[\mathbf{x}_j^{(k)} - \overline{\mathbf{x}}_j\Big], \tag{3.14}$$

$$\text{where } \overline{\mathbf{x}}_i = \frac{1}{N_{\rm mocks}} \sum_{k=1}^{N_{\rm mocks}} \mathbf{x}_i^{(k)}. \tag{3.15}$$

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.

3.4.2 ABC-PMC Design

In Section 3.3.1, 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 multi-component distance (**?**, Cisewsky et al in preparation). Each summary statistic has a distance associated to it: ρ_n, ρ_ξ, and ρ_ζ. We calculate each of these distance components as,

$$\rho_n = \frac{\left(\bar{n}_{\rm g}^{\rm d} - \bar{n}_{\rm g}^{\rm m}\right)^2}{\sigma_n^2}, \tag{3.16}$$

$$\rho_\xi = \sum_k \frac{\left[\xi_{\rm gg}^{\rm d}(r_k) - \xi_{\rm gg}^{\rm m}(r_k)\right]^2}{\sigma_{\xi,k}^2}, \tag{3.17}$$

$$\rho_\zeta = \sum_k \frac{\left[\zeta_{\rm g}^{\rm d}(N_k) - \zeta_{\rm g}^{\rm m}(N_k)\right]^2}{\sigma_{\zeta,k}^2}. \tag{3.18}$$

The superscripts d and m denote the data and model respectively. The data, are the observables calculated from the mock observation (Section 3.4.1). σ_n^2, $\sigma_{\xi,k}^2$, and $\sigma_{\zeta,k}^2$ are not the diagonal elements of the covariance matrix (3.14). Instead, they are diagonal elements

of the covariance matrix $C^{\rm ABC}$.

We construct $C^{\rm ABC}$ by populating the entire `MultiDark` halo catalogs 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 Eq. 3.14, 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 $\vec{\rho} = [\rho_n, \rho_\xi]$ while the distance metric for the ABC-PMC analysis using the observables $\bar{n}_{\rm g}$ and $\zeta_{\rm g}$, is $\vec{\rho} = [\rho_n, \rho_\zeta]$. 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 4.5.1.

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 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 **?** and set the distance threshold $\vec{\epsilon}_t$ to the median of $\vec{\rho}_{t-1}$, the multi-component distance of the previous iteration of particles ($\vec{\theta}_{t-1}$).

The distance threshold $\vec{\epsilon}_t$ will progressively decrease. Eventually after a sufficient number of iterations, the region of parameter space occupied by $\vec{\theta}_t$ will remain unchanged. As this happens, the acceptance ratio begins to fall significantly. When the acceptance ratio drops

Table 3.1: **Prior Specifications**: The prior probability distribution and its range for each of the **?** HOD parameters. All mass parameters are in unit of $h^{-1}M_{\odot}$

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]
$\log M_{min}$	Uniform	[11.02, 13.02]
$\log M_1$	Uniform	[13.0, 14.0]

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 ($\bar{n}_{\mathrm{g}}$, ξ_{gg}) and ($\bar{n}_{\mathrm{g}}$, ζ_{g}).

3.4.3 Results: ABC

We describe the ABC algorithm in Section 3.3.1 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 **?** HOD model.

We begin with a qualitative demonstration of the ABC algorithm in Figure 3.2, where we plot the evolution of the ABC $\vec{\theta}_t$ over the iterations $t = 1$ to 9, in the parameter space of $[\log \mathcal{M}_1, \log \mathcal{M}_{min}]$. The ABC procedure we plot in Figure 3.2 uses $\bar{n}$ and $\zeta_{\mathrm{g}}(N)$ for observables, but the overall evolution is the same when we use $\bar{n}$ and $\xi_{\mathrm{gg}}(r)$. The darker and lighter contours represent the 68% and 95% confident regions of the posterior distribution over $\vec{\theta}_t$. For reference, we also plot the "true" HOD parameter $\vec{\theta}_{\mathrm{true}}$ (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 4.5.1.

For $t = 1$, the initial pool (top left), the distance threshold $\vec{\epsilon}_1 = [\infty, \infty]$, so $\vec{\theta}_1$ uniformly samples the prior probability over the parameters. At each subsequent iteration, the threshold is lowered (Section 3.4), so for $t < 6$ panels, we note that the parameter spaced occupied by $\vec{\theta}_t$ dramatically shrinks. Eventually when the algorithm begins to converge, $t > 7$, the contours enclosing the 68% and 95% confidence interval stabilize. At the final iteration $t = 9$ (bottom right), the algorithm has converged and we find that $\vec{\theta}_{\rm true}$ lies within the 68% confidence interval of the $\vec{\theta}_{t=9}$ particle distribution. This $\vec{\theta}_t$ distribution at the final iteration represents the posterior distribution of the parameters.

To better illustrate the criteria for convergence, in Figure 3.3, we plot the evolution of the $\vec{\theta}_t$ distribution as a function of iteration for parameters $\log \mathcal{M}_{\rm min}$ (left), α (center), and $\log \mathcal{M}_1$ (right). The darker and lighter shaded regions correspond to the 68% and 95% confidence levels of the $\vec{\theta}_t$ 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 $\vec{\epsilon}_t$ nor the posterior distribution of $\vec{\theta}_t$. We also emphasize that the convergence of the parameter distributions coincides with when the acceptance ratio, discussed in Section 3.4.2, 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 Figure 3.4 and Figure 3.5. Figure 3.4 shows the parameter constraints using $\bar{n}$ and $\xi_{\rm gg}(r)$ while Figure 3.5 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% (median) and 68% 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 center of the 68% confidence interval. For the other parameter, which have much tighter constraints, the true parameters lie within the 68% confidence interval.

To further test the ABC results, in Figure 3.6, we compare $\xi_{\rm gg}(r)$ (left) and $\zeta_{\rm g}(N)$ (right) of the mock observations from Section 3.4.1 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 (Eq. 3.14) while the darker and lighter shaded regions represent the 68% and 95% 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% 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 $r > 5$ 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 3.3.1 with the choices of Section 3.4.2 produce parameter constraints that are consistent with the "true" HOD parameters (Figures 3.4 and 3.5). 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 large scale structure.*

3.4.4 Comparison to the Gaussian 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 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. In the ABC analysis, the model accounts for sample variance by randomly sampling a subvolume to be populated with galaxies. The Gaussian likelihood analysis assumes a covariance matrix that captures the uncertainties from the sample variance, and therefore in the model, we populate the halos of the entire `MultiDark` simulation rather than a subvolume.

We introduce the observable $\mathbf{x}$ as a combination of the summary statistics of the galaxy mock catalog used in the inference. When we use $\bar{n}_{\rm g}$ and $\xi_{\rm gg}(r)$ as observables, $\mathbf{x} = [\bar{n}_{\rm g}, \xi_{\rm gg}]$, while $\mathbf{x} = [\bar{n}_{\rm g}, \zeta_{\rm g}]$ when we use $\bar{n}_{\rm g}$ and $\zeta_{\rm g}(N)$ Based on this notation, we write likelihood function as

$$-2\ln\mathcal{L}(\theta|d) = \Delta\mathbf{x}^T\widehat{C^{-1}}\Delta\mathbf{x} + \ln\Big[(2\pi)^d\det(C)\Big], \tag{3.19}$$

$$\Delta\mathbf{x} = [\mathbf{x}_{obs} - \mathbf{x}_{mod}], \tag{3.20}$$

where $\Delta\mathbf{x}$ is the difference between the measured $\mathbf{x}_{obs}$ of the mock observations and that of the model $\mathbf{x}_{mod}(\theta)$. d is the dimension of the observable $\mathbf{x}$. When we use $\mathbf{x} = [\bar{n}_{\rm g}, \xi_{\rm gg}]$, $d = 13$. When we use $\mathbf{x} = [\bar{n}_{\rm g}, \zeta_{\rm g}]$, $d = 10$. $\widehat{C^{-1}}$ is the estimate of the inverse covariance matrix, which we calculate following **?**:

$$\widehat{C^{-1}} = \frac{N_{\rm mocks} - d - 1}{N_{\rm mocks} - 1}\widehat{C}^{-1}, \tag{3.21}$$

where $\widehat{C}$ is the estimated covariance matrix and $N_{\rm mock}$ is the number of mocks used to estimate $\widehat{C}$. $N_{\rm mock} = 124$ (Section 3.4.1). Depending on $\mathbf{x}$, $\widehat{C}$ is given by the appropriate block of the covariance matrix Eq. 3.14 corresponding to the observables. We note that in the covariance matrix the dependence on the HOD parameters is neglected. Therefore, the second term in the expression of the log-likelihood (Eq. 3.19) can be neglected. We sample from the posterior distribution of the likelihood given the prior distribution using the MCMC sampler `EMCEE` (**?**).

In Figure 3.7 we compare the ABC-PMC and Gaussian likelihood MCMC results for the marginalized posterior PDFs over three parameters of the HOD model $\{\log \mathcal{M}_{\rm min}, \alpha, \log \mathcal{M}_1\}$ using $\bar{n}_{\rm g}$ and $\xi_{\rm gg}$ as observables. Figure 3.8 makes the same comparison when $\bar{n}_{\rm g}$ and $\zeta_{\rm g}$ are used in the inference.

Figure 3.7 demonstrates in the top panel that constraints obtained from the two methods are consistent with the "true" HOD parameters represented by the vertical dashed lines. The marginalized posteriors over the parameters follow each other very closely. The lower panel of Figure 3.7 plot the 68% and 95% confidence intervals of the constraints derived from the two methods as a box plot. The uncertainty over the parameter $\log \mathcal{M}_{\rm min}$ is slightly smaller while the uncertainties of the parameters α and $\log M_1$ are slightly larger for our ABC-PMC results.

In Figure 3.8, when we use $\bar{n}_{\rm g}$ and $\zeta_{\rm g}$ as observables, both methods produce consistent constraints with each other and "true" HOD parameters. The lower panel of Figure 3.8 shows that the uncertainties from the two methods are comparable. The ABC-PMC constraints over α is slightly less biased and slightly less precise.

In Figures 3.9 and 3.10, we plot the contours enclosing the 68% and 95% confidence regions of the posterior probabilities of the two methods. Figure 3.9 uses $\bar{n}_{\rm g}$ and $\xi_{\rm gg}$ as observables, while Figure 3.10 uses $\bar{n}_{\rm g}$ and $\zeta_{\rm g}$ as observables. In both figures, the "true"

HOD parameters are plotted as black stars. The overall shape of the contours are generally in good agreement with each other. However, the contours for the ABC-PMC method are more elongated along α.

Overall, the constraints from ABC-PMC are consistent with those from the Gaussian pseudo-likelihood MCMC method; however, using ABC-PMC has a number of advantages. ABC-PMC utilizes a forward generative model. Our forward generative model, for instance, accounts for sample variance. Forward generative models also have the advantage that they can account for sources of systematic uncertainties that affect observational data.

CHH: @MJV I'm not a big fan of this entire paragraph. The accuracy of the covariance matrix depends on the accuracy of the mock catalogs (a.k.a. the forward model). While it's definitely true that on small scales they break down, forward models in ABC suffers from the same issues. Furthermore, the Gaussian-likelihood approach relies on constructing an accurate covariance matrix estimate that captures the sample variance of the data. While we are able to do this accurately within the scope of the HOD framework, for more general LSS parameter inference situations, it is both labor and computationally expensive and dependent on the accuracy of simulated mock catalogs, which are known to be unreliable on small scales (see **??** and references therein). Since ABC-PMC utilizes a forward model to account for sample variance, it does not depend on a covariance matrix estimate; hence it does not face these problems.

ABC-PMC – unlike the Gaussian-likelihood approach – is agnostic about the functional form of the underlying distribution of the summary statistics ($\xi_{\rm gg}$ and $\zeta_{\rm g}$). This may explain why we find less biased constraints from the ABC-PMC analysis compared to the Gaussian-likelihood analysis, when using $\zeta_{\rm g}(N)$ as an observable (Figures 3.8 and 3.10).

The ABC method of parameter inference has often been viewed as computationally in-

feasible due to the fact that it necessitates a generative forward model for a large number of realizations. However, by combining ABC with the PMC sampling method, we have a method that efficiently converges to give us reliable posteriors of the model parameters. Furthermore, in our analysis, the computational resources required for the ABC-PMC inference were comparable to those used in achieving convergence in the MCMC Gaussian-Likelihood inference. This is excluding the typically computationally intensive step of calculating a covariance matrix estimate. Therefore, we find that ABC-PMC offers a method for parameter inference in large scale structure studies with a number of advantages we describe above.

3.5 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 large scale structure 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

labor 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 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:

- 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 catalogs of galaxy positions based on models of the galaxy-halo connection such as the HOD.

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

- 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 LSS studies.

- 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`

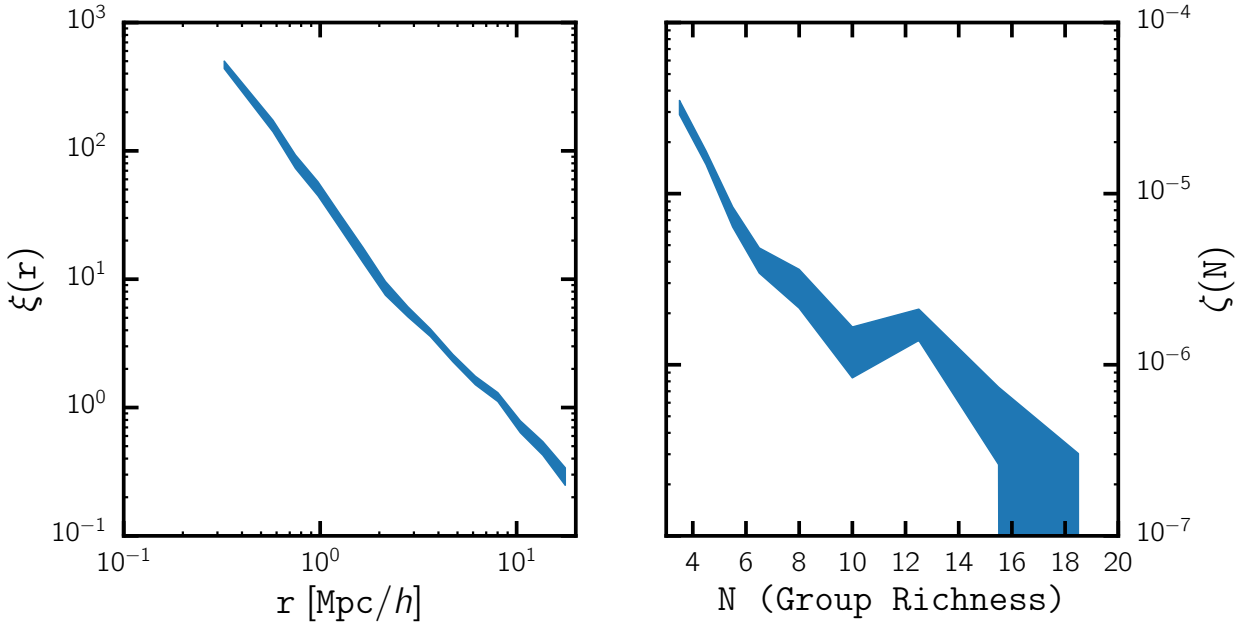

Figure 3.1: The two-point correlation function $\xi_{\rm gg}(r)$ (left) and group multiplicity function $\zeta_{\rm g}(N)$ (right) summary statistics of the mock observations generated from the "true" HOD parameters described in Section 3.4.1. The width of the shaded region corresponds to the square root of the covariance matrix diagonal elements (Eq. 3.14). In our ABC analysis, we treat the $\xi_{\rm gg}(r)$ and $\zeta_{\rm g}(N)$ above as the summary statistics of the observation.

(`https://github.com/simongibbons/pyfof`).

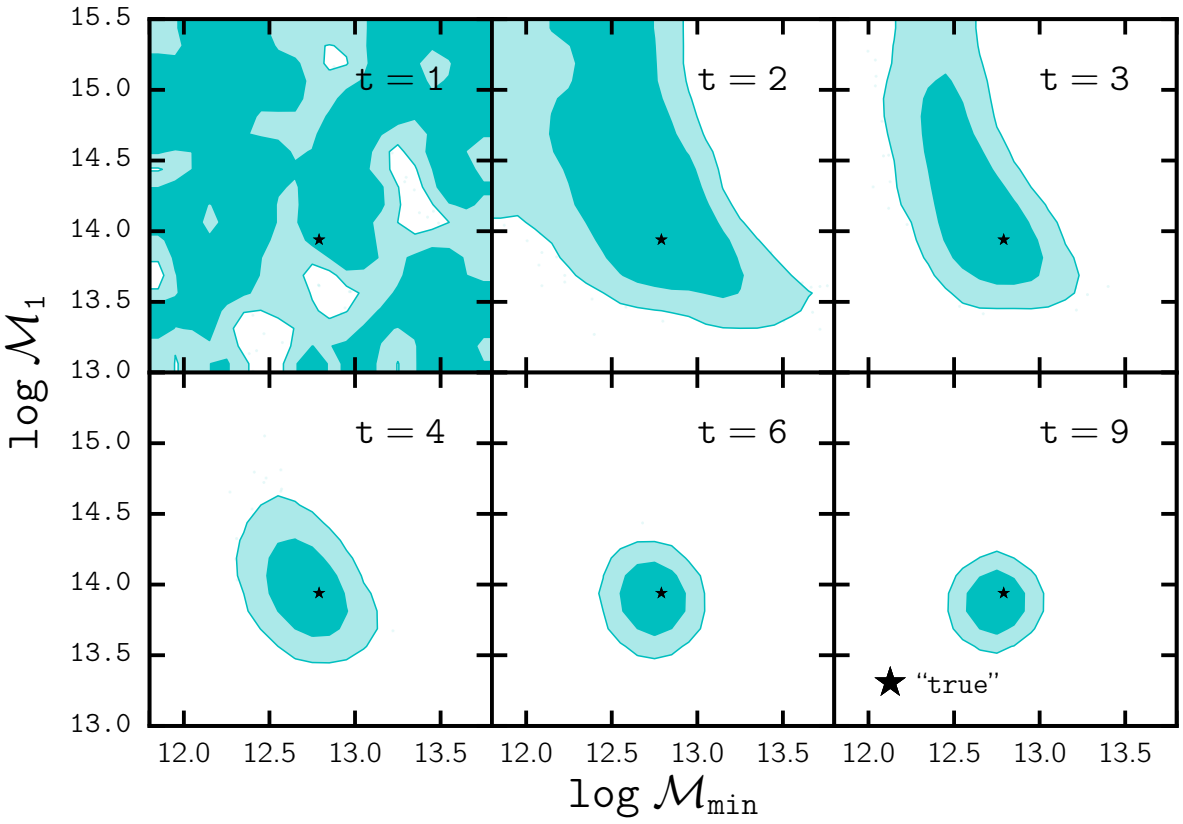

Figure 3.2: We demonstrate the evolution of the ABC particles, $\vec{\theta}_t$, over iterations $t = 1$ to 9 in the $\log \mathcal{M}_{min}$ and $\log \mathcal{M}_1$ parameter space. $\bar{n}$ and $\zeta_{\rm g}(N)$ are used as observables for the above results. For reference, in each panel, we include the "true" HOD parameters (black star) listed in Section 3.4.1. The initial distance threshold, $\vec{\epsilon}_1 = [\infty, \infty]$ at $t = 1$ (top left) so the $\vec{\theta}_1$ 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 $\vec{\theta}_t$ shrinks dramatically. Eventually when the algorithm converges, $t > 7$, the parameter space occupied by $\vec{\theta}_t$ 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 $\vec{\theta}_{\rm true}$ lies safely within the 68% confidence region.

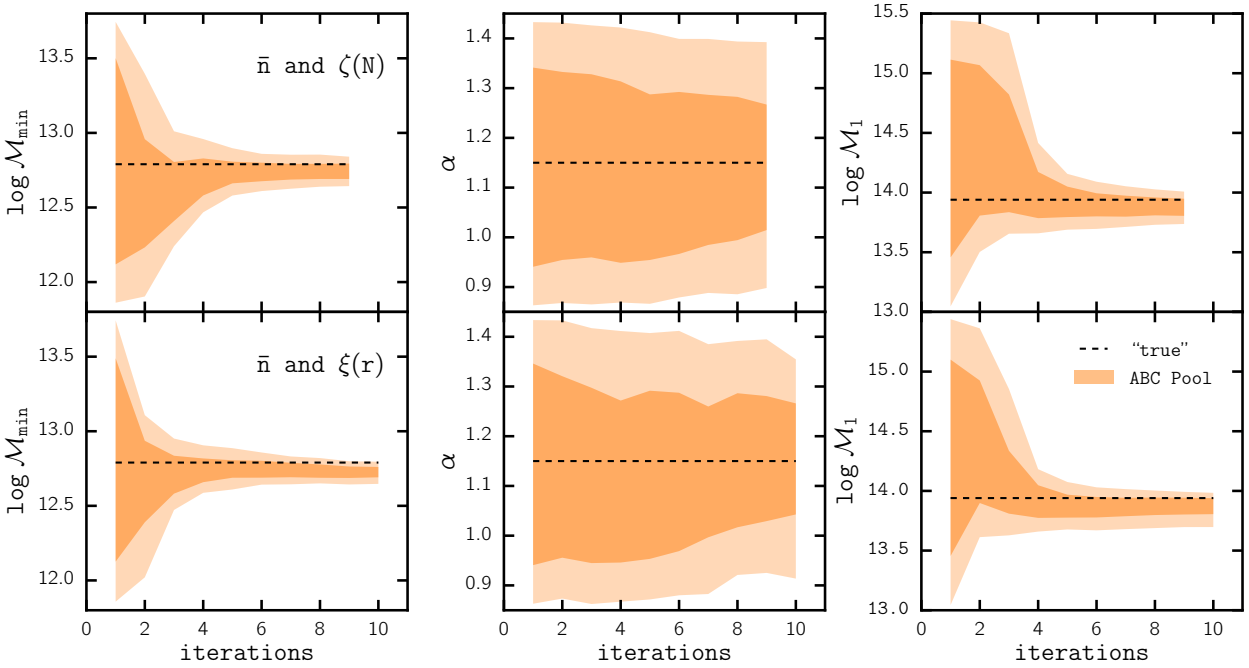

Figure 3.3: We illustrate the convergence of the ABC algorithm through the evolution of the ABC particle distribution as a function of iteration for parameters $\log \mathcal{M}_{\rm min}$ (left), α (center), and $\log \mathcal{M}_1$ (right). The top panel corresponds our ABC results using the observables $(\bar{n}, \zeta_{\rm g}(N))$, 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.

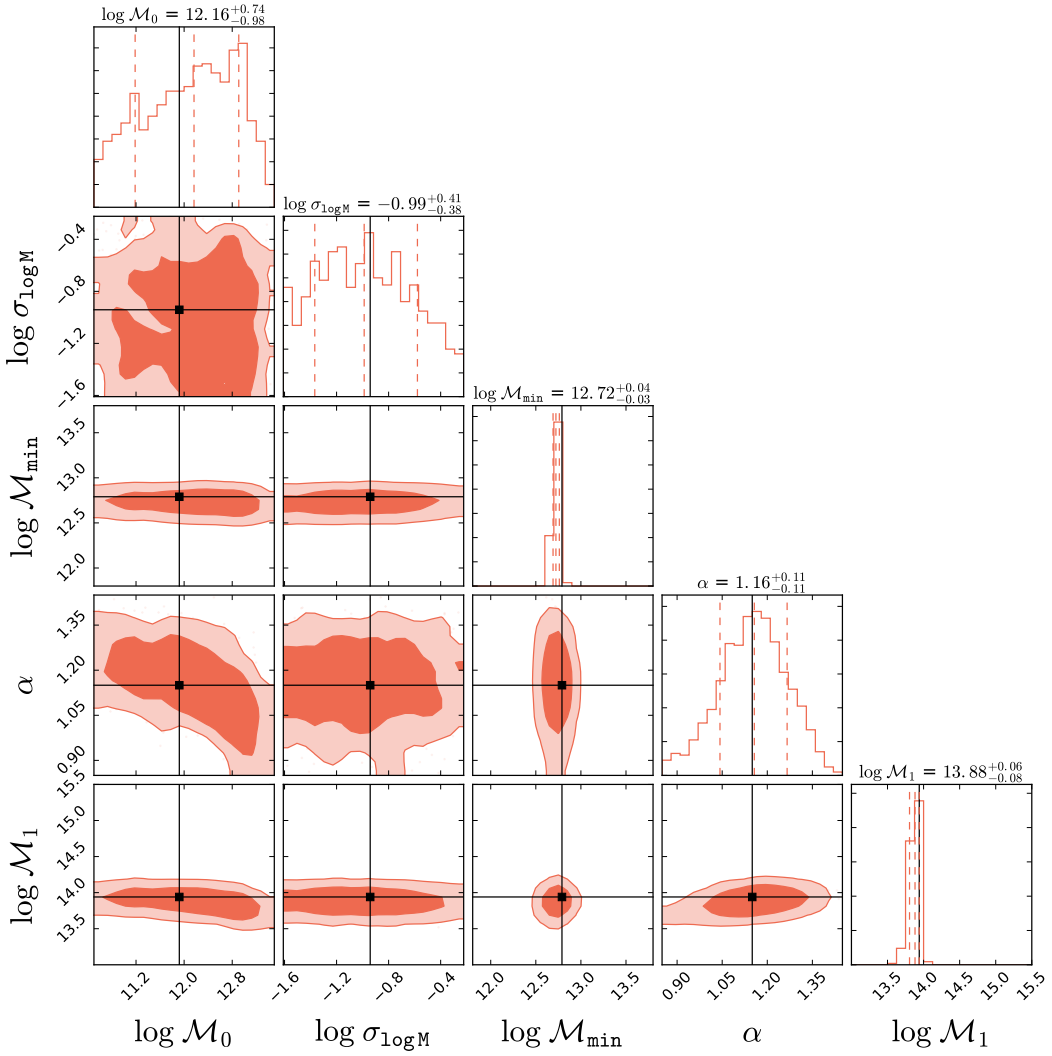

Figure 3.4: We present the constraints on the ? HOD model parameters obtained from our ABC-PMC analysis using $\bar{n}$ and $\xi_{\rm gg}(r)$ as observables. The diagonal panels plot the posterior distribution of each HOD parameter with vertical dashed lines marking the 50% quantile and 68% confidence 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 3.4.1, 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 center of the 68% 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% confidence interval. Ultimately, the ABC parameter constraints we obtain in our analysis are consistent with the "true" HOD parameters.

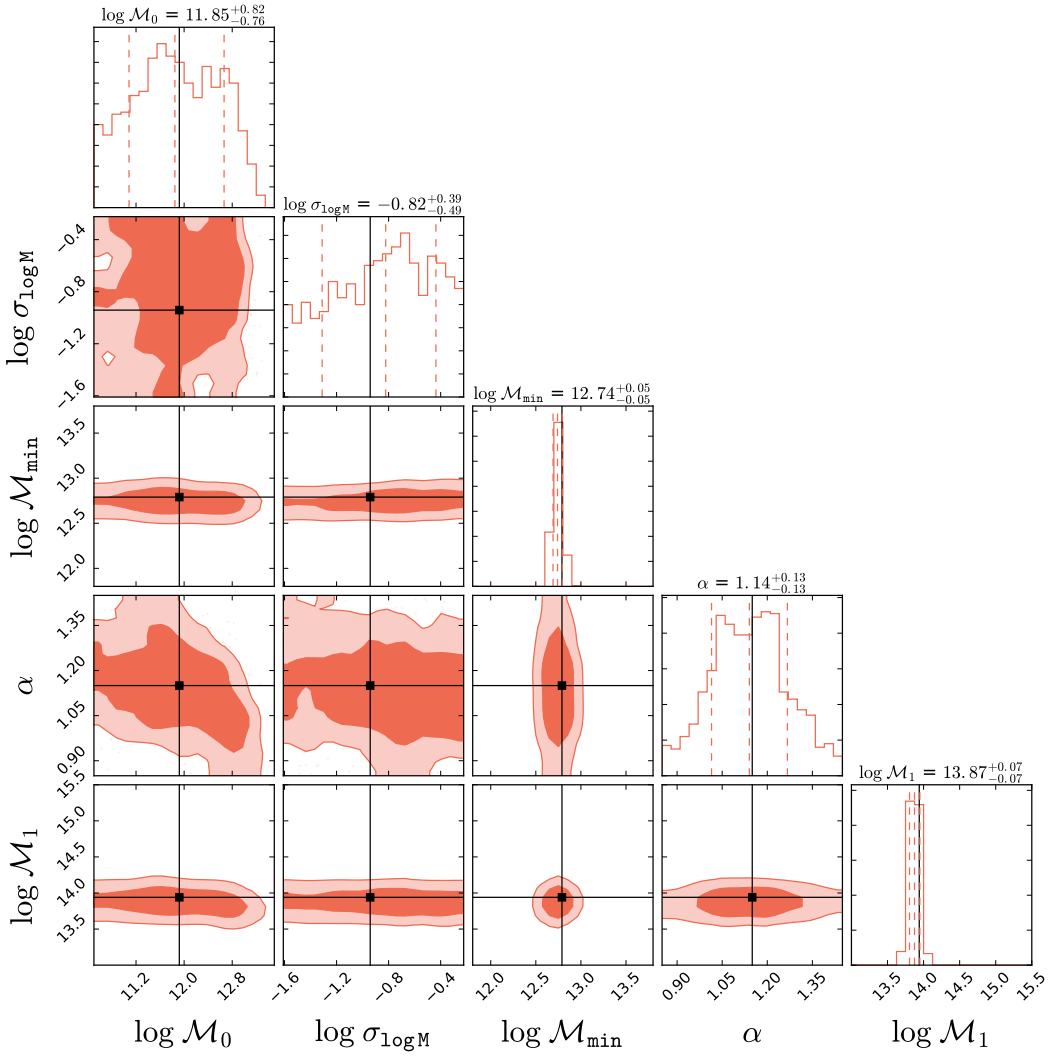

Figure 3.5: Same as Figure 3.4 but for our ABC analysis using $\bar{n}$ and $\zeta_{\rm g}(N)$ as observables. The ABC parameter constraints we obtain are consistent with the "true" HOD parameters.

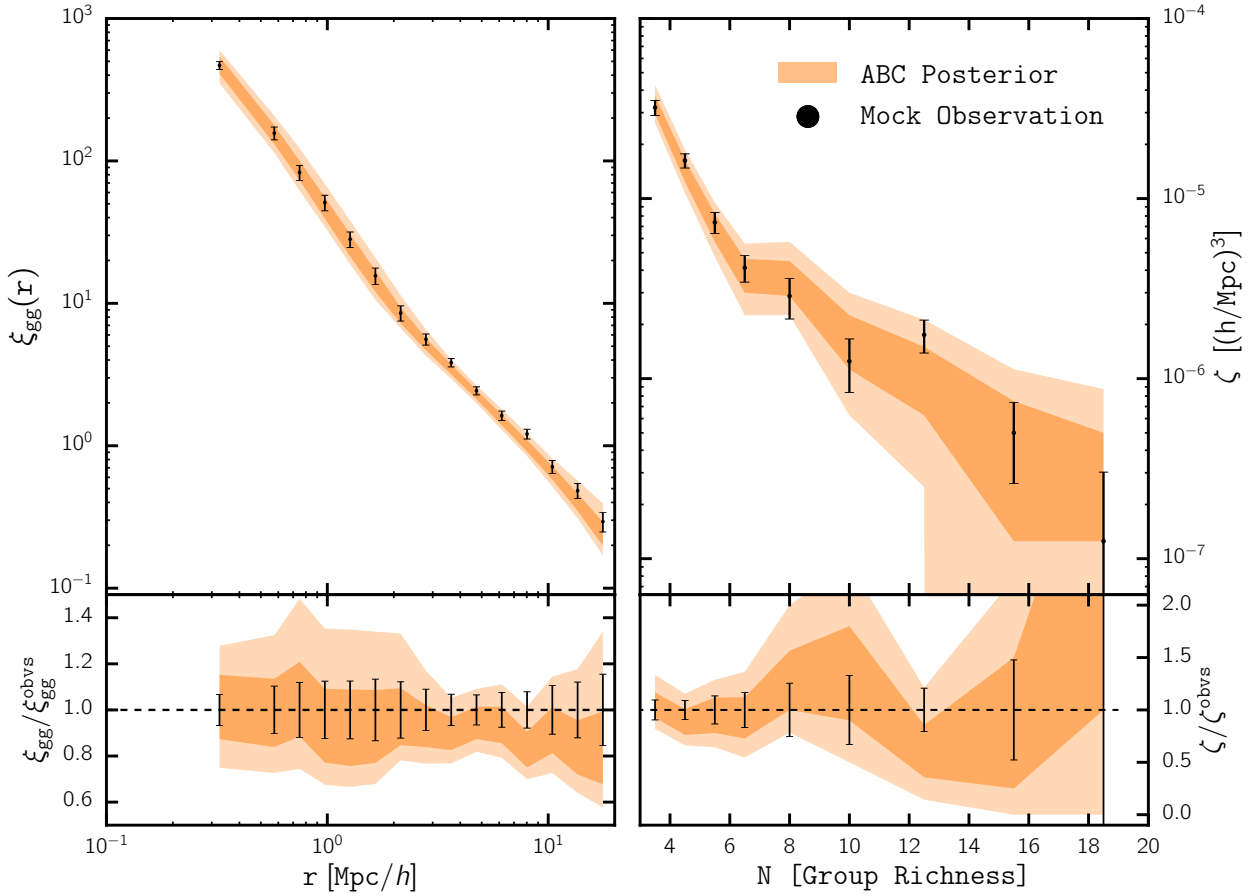

Figure 3.6: We compare the ABC-PMC posterior prediction for the observables $\xi_{gg}(r)$ (left) and $\zeta_g(N)$ (right) (orange; Section 3.4.3) to $\xi_{gg}(r)$ and $\zeta_g(N)$ of the mock observation (black) in the top panels. In the lower panels, we plot the ratio between the ABC-PMC posterior predictions for ξ_{gg} and ζ_g to the mock observation ξ^{obvs}_{gg} and ζ^{obvs}_g. The darker and lighter shaded regions represent the 68% and 95% confidence regions of the posterior predictions, respectively. The error-bars represent the square root of the diagonal elements of the error covariance matrix (equation 3.14) 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% confidence interval of the ABC-PMC posterior predictions.

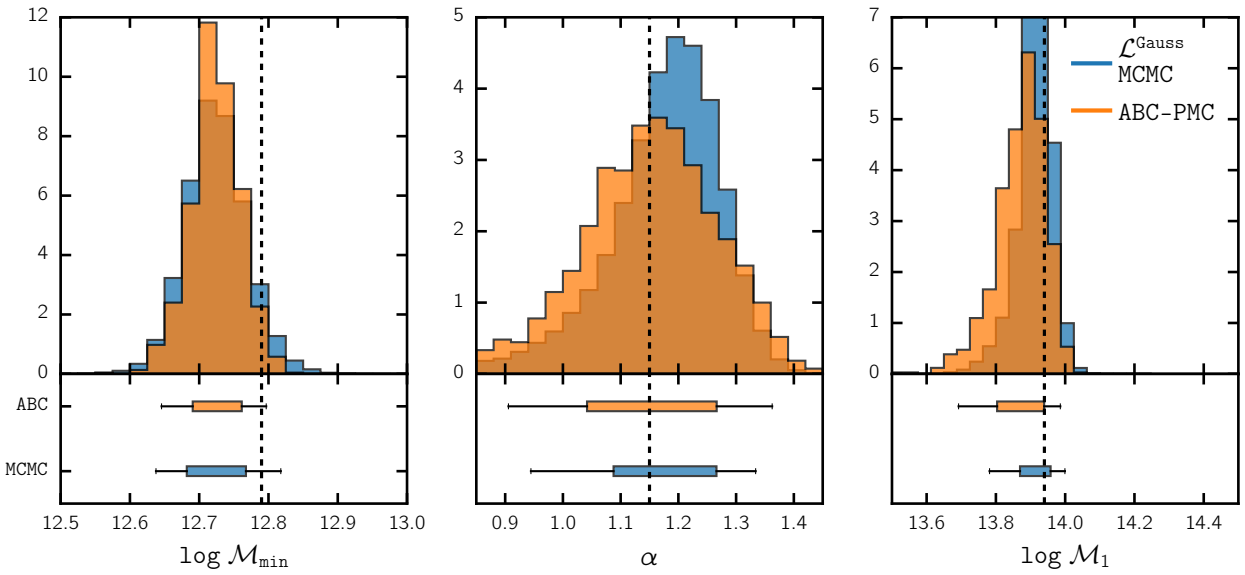

Figure 3.7: We compare the $\log \mathcal{M}_{\mathrm{min}}$, α, and $\log \mathcal{M}_1$ parameter constraints from ABC-PMC (orange) to the constraints from the Gaussian Likelihood MCMC (blue) using $\bar{n}_{\mathrm{g}}$ and $\xi_{\mathrm{gg}}(r)$ as observables. The *top* panels show the histograms of the 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% confidence interval while the "whiskers" represent the 95% confidence interval. We also plot the "true" HOD parameters with vertical black dashed line. Marginalized posterior PDFs obtained from the two methods are consistent with each other. The ABC-PMC constraints are slightly narrower for $\log \mathcal{M}_{\mathrm{min}}$, slightly wider for $\log \mathcal{M}_1$, and slightly less biased for α.

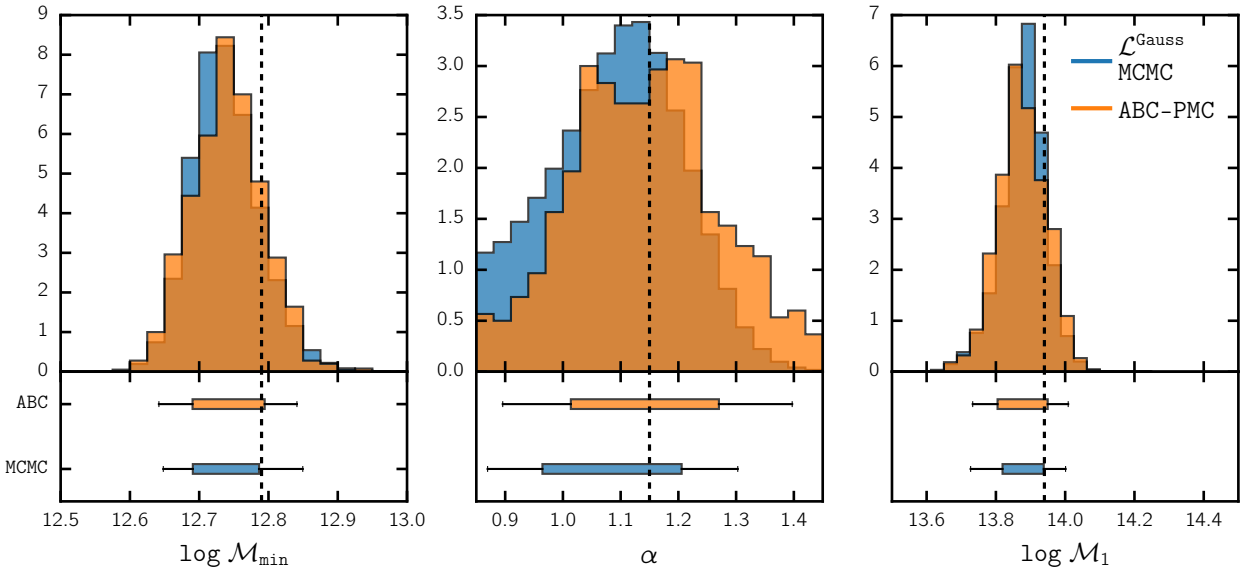

Figure 3.8: Same as Figure 3.7, but both the ABC-PMC analysis and the standard Gaussian Likelihood MCMC analysis are done using the observables $\bar{n}_{\rm g}$ and $\zeta_{\rm g}(N)$. The constraints are consistent with the "true" HOD parameters and both methods infer the region of allowed values to similar precision. The MCMC result for α is slightly more biased compared to ABC-PMC estimate. This may stem from the fact that the use of Gaussian-likelihood and its associated assumptions is more spurious when modeling the group multiplicity function.

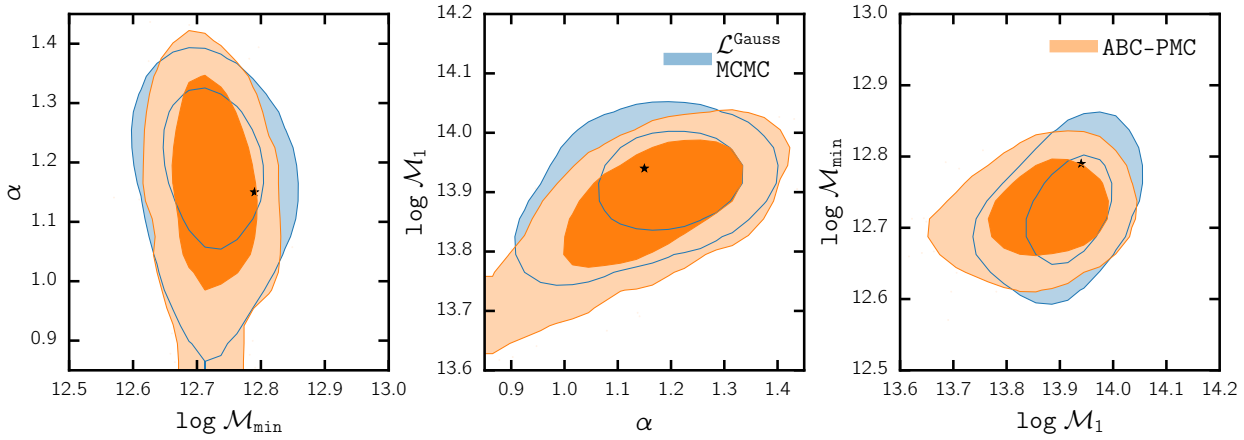

Figure 3.9: We compare the ABC-PMC (orange) and the Gaussian likelihood MCMC (blue) predictions of the 68% and 95% posterior confidence regions over the HOD parameters ($\log \mathcal{M}_{\rm min}$, α, and $\log \mathcal{M}_1$) using $\bar{n}_{\rm g}$ and $\xi_{\rm gg}(r)$ as observables. The "true" HOD parameters used to create the mock observations are plotted by black stars in each panel. The two approaches are consistent with the true values of the parameters used to generate the data.

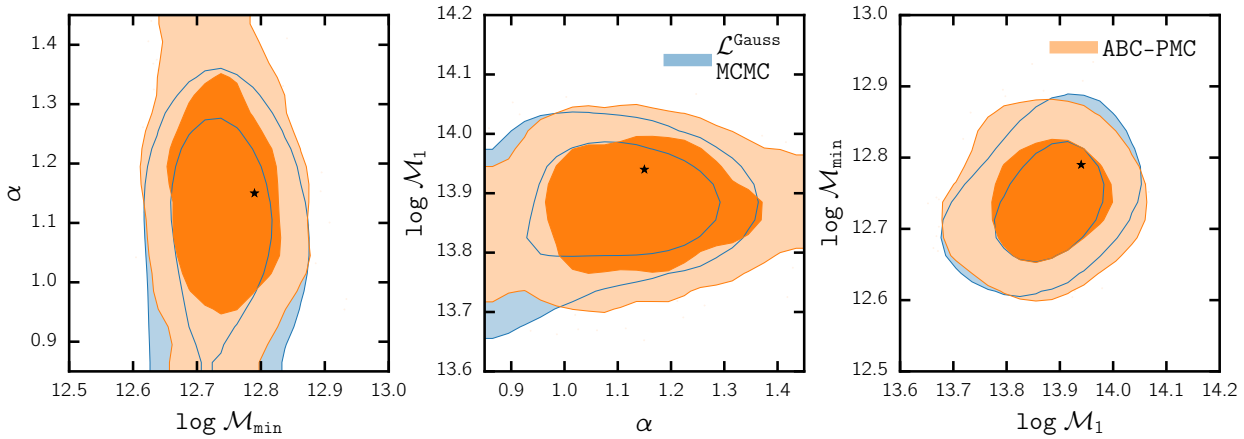

Figure 3.10: Same as Figure 3.9, but using $\bar{n}_{\rm g}$ and $\zeta_{\rm g}(N)$ as observables. Again, both methods accurately estimate the parameter value regions of the true values for the data. The MCMC estimation of α by use of a Gaussian-likelihood is biased when compared with the ABC-PMC contours. This may be due to the fact that the group multiplicity function is particularly unsuited to the use of a Gaussian-likelihood analysis.

Chapter 4

How are galaxies assigned to halos? searching for assembly bias in SDSS clustering measurements

This Chapter is joint work with ChangHoon Hahn (NYU) submitted to the *Astrophysical Journal* as **?**.

*arg max

4.1 chapter abstract

Clustering of dark matter halos has been shown to depend on halo properties beyond mass such as halo concentration, a phenomenon referred to as halo assembly bias. Standard halo occupation **models** (HOD) in large scale structure studies assume that halo mass alone is sufficient in characterizing the connection between galaxies and halos. **Modeling of galaxy clustering can face systematic effects if the number of galaxies are**

correlated with other halo properties. Using the Small MultiDark-Planck high resolution N-body simulation and the measurements of the projected two-point correlation function and the number density of Sloan Digital Sky Survey (SDSS) DR7 main galaxy sample, we investigate the extent to which the dependence of halo occupation on halo concentration can be constrained, and to what extent allowing for this dependence can improve our modeling of galaxy clustering.

From our constraints on HOD with assembly bias, suggests that satellite population is not correlated with halo concentration at fixed halo mass. **Furthermore, in terms of the occupation of centrals at fixed halo mass, our results favor lack of correlation with halo concentration in the most luminous samples ($M_{\rm r} < -21.5, -21$), modest levels of correlation for $M_{\rm r} < -20.5, -20, -19.5$ samples, lack of correlation for $M_{\rm r} < -19, -18.5$ samples, and anti-correlation for the faintest sample $M_{\rm r} < -18$.**

We show that in comparison with abundance-matching mock catalogs, our findings suggest qualitatively similar but modest levels of the impact of halo assembly bias on galaxy clustering. **The** effect is only present in the central occupation and becomes less significant in brighter galaxy samples. Furthermore, by performing model comparison based on information criteria, we find that in most cases, the standard mass-only HOD model is still favored by the observations.

4.2 Introduction

Most theories of cosmology and large-scale structure formation under consideration today rely on the central assumption that galaxies reside in dark matter halos. Detailed study of the galaxy–halo connection is therefore critical in **constraining** cosmological models (by modeling galaxy clustering at non-linear scales) as well as providing a window into

galaxy formation physics. One of the most powerful methods for describing the galaxy–halo connection is the halo occupation distribution (HOD, see **?????????**).

HOD is an empirical framework that provides an analytic prescription for the expected number of galaxies N that reside in halos by specifying a probability distribution function $P(N|x)$ where x is a property of the halo. **The standard HOD model assumes** that halo mass M alone is sufficient in determining the galaxy population of a halo. **In the standard model, the statistical properties of galaxies is governed by the halo mass.** Mathematically, this assumption can be written as $P(N|M, \{x\}) = P(N|M)$ where $\{x\}$ is the set of all possible halo properties beyond halo mass M.

Despite this simplifying assumption, the models of galaxy–halo connection based on HOD have been successfully used in fitting the measurements of a wide range of statistics such as the projected two-point correlation function of galaxies, small scale redshift space distortion, three-point function, and galaxy–galaxy lensing with remarkable success (e.g. **??????????**). HOD has been used in constraining the cosmological parameters through modeling the galaxy two-point correlation function (hereafter 2PCF) (**?**), combination of 2PCF with mass-to-light ratio of galaxies (**?**), redshift space distortions (**?**), mass-to-number ratio of galaxy clusters (**?**) galaxy-galaxy weak lensing (**????**) in the main sample of galaxies of the Sloan Digital Sky Survey (hereafter SDSS, **?**), and also the combination of galaxy clustering and galaxy-galaxy lensing (**?**) in SDSS III Baryon Oscillation Spectroscopic Survey (BOSS, **?**). Furthermore, HOD is implemented in producing mock galaxy catalogs in the BOSS survey (**??**). It has also been used in galaxy evolution studies (**??????**).

The complexity of structure formation however, is not sufficiently modeled under the standard HOD framework. Numerous N-body simulations that examine the clustering of dark matter halos have demonstrated that halo clustering is correlated with the formation history of halos. That is, at a fixed halo mass, the halo bias is correlated with properties of

halos beyond mass, such as the concentration, formation time, or etc. This phenomenon is known as halo assembly bias (see **???????????**). It has been claimed that there is support for halo assembly bias in observations of SDSS redMaPPer galaxy clusters (**?**).

Furthermore, the halo occupation may also depend on the formation history of halos. Then we may expect the spatial statistics of galaxies to be tied to the halo properties beyond mass such as the concentration of halos. There have been many attempts in the literature at examining the dependence of halo occupation on environment of the halos. But the results are mixed, and there is very little consensus. **?** show that the properties of the galaxies that reside in voids can be explained by the halo mass in which they live, and their properties are independent of their large scale environment of the halos. **?** proposes an extension of the standard HOD model $P(N|M)$ such that the number of galaxies residing in a halo not only depends on the mass of the halo, but also on the large scale density contrast $P(N|M, \delta)$. Based on modeling the clustering and void statistics of the SDSS galaxies, **?** shows that the dependence of the expected number of central galaxies on large scale density is not very strong. By randomly shuffling the galaxies among host halos of similar mass in the Millenium simulation, **?** shows that assembly bias significantly impacts the galaxy two-point correlation functions. They also show that the effect is different for the faint and the bright samples.

Another family of empirical galaxy–halo connection models is *Abundance Matching*. In Abundance Matching models, galaxies are assumed to live in halos and are assigned luminosities or stellar masses by assuming a monotonic mapping. In this monotonic mapping, the abundance of the halos are matched to the abundance of some property of galaxies (**????????????**). One of the most commonly used host halo properties in abundance matching is the maximum circular velocity of the host halo $V_{\rm max}$ that traces the depth of the gravitational potential well of the halo. Furthermore, a scatter is assumed in this map-

ping. Within this galaxy–halo connection framework, abundance matching models have been successfully used in modeling a wide range of the statistical properties of galaxies such as two-point correlation function (**???**) as well as the group statistics of galaxies (**?**).

It has been shown that the abundance matching mock catalogs that use $V_{\rm max}$ (see **??**), or the ones that use some combination of $V_{\rm max}$ and host halo virial mass $M_{\rm vir}$ (see **?**) exhibit significant levels of assembly bias. That is, halo occupation in these models depends not only on halo mass, but also on other halo properties. This has been demonstrated by randomizing the galaxies among host halos in bins of halo mass, such that the HOD remains constant, and then comparing the difference in the 2PCF of the randomized catalog and that of the original mock catalog.

Based on the projected 2PCF measurements of (**?**) galaxy catalogs, **?** showed that after fitting the 2PCF measurements of these catalogs with the standard *mass-only* HOD modeling, the inferred HOD does not match the *true* halo occupation of these catalogs. That is, in the presence of assembly bias in a galaxy sample, one can fit the clustering of this sample with the standard *mass-only* HOD, but that does not guarantee recovery of the true HOD parameters.

In this work, we aim to investigate the dependence of halo occupation on **halo concentration and how this dependence can be constrained in the low-redshift universe by 2PCF measurements of galaxies in a wide range of luminosities in the SDSS DR7 main galaxy sample.** In order to achieve this goal, we need to adopt a HOD model that takes into account a dependence on halo properties beyond mass. **A number of frameworks in the literature (???) have proposed environment-dependent HOD models that take into account the large-scale density contrast. In this investigation, we use the following case of the decorated HOD framework (?). In our decorated HOD framework, at fixed halo mass, halos are populated with galaxies**

according to the standard HOD model. Then using a secondary halo property, halos are split into two populations in halo mass bins: halos with the highest and lowest secondary property values. Afterwards, based on the assembly bias amplitude parameter, the number of galaxies in the two populations are enhanced or reduced. In this model, the assembly bias parameter is not be degenerate with the rest of the HOD parameters.

The advantage of this framework is that the more complex HOD model is identical to the underlying *mass-only* HOD model in every respect, except that at a fixed halo mass, halos receive enhancement (decrements) in the number of galaxies they host according to the value of their secondary property. In order to constrain assembly bias along with the rest of the HOD parameters, we make use of the publicly available measurements of the projected 2PCF and number density measurements made by (**?**). These measurements made use of the NYU Value-Added Galaxy Catalog (**?**).

Furthermore, we discuss how taking assembly bias into account in a more complex HOD model can improve our modeling of galaxy clustering in certain brightness limits. Then, we make a qualitative comparison between the levels of the impact of assembly bias in our best-fit decorated HOD model on galaxy clustering, and the impact of assembly bias present in (**?**) catalogs on galaxy clustering. Our comparison shows the levels of the impact of assembly bias on galaxy clustering seen in the predictions of both models follow the same trend. That is assembly bias is more prominent in lower luminosity-threshold samples and its impact on galaxy clustering is only significant on large scales (more than a few Mpc).

In order to investigate whether the additional complexity of the decorated HOD model is demanded by the galaxy clustering data, we perform a model comparison between the standard HOD model and the HOD model with assembly bias. We also discuss the effect of our choice of N-body simulation on our constraints, and previous works in the literature (**?**)

based on smaller N-body simulations. In addition to analysis of the luminosity-threshold samples presented in **?**, we consider the faintest ($M_{\rm r} < -18, -18.5$) and the brightest ($M_{\rm r} < -20.5$) galaxy samples. For the samples considered in both **?** and this investigation, we compare the constraints on the expected levels of assembly bias.

This paper is structured as follows: In Section 5.3 we discuss the N-body simulation, the two halo occupation modeling methods, and the details of the computation of model observables used in this investigation. Then in Section 4.4, we discuss the data used in this study. In Section 4.5 we discuss the details of our inference analysis as well as the results. This includes description of the details of our inference setup. In Section 4.6 we discuss the constraints and their implications. This includes presentation of the constraints on the parameters of the two models, interpretation of the predictions of our constraints and their possible physical ramifications, assessment of the levels of assembly bias as predicted by our model constraints and its comparison with abundance matching mock catalogs, and finally model comparison. Finally, we discuss and conclude in Section 4.7. Throughout this paper, unless stated otherwise, all radii and densities are stated in comoving units. Standard flat ΛCDM is assumed, and all cosmological parameters are set to the Planck 2015 best-fit estimates.

4.3 Method

In this section, we discuss the ingredients of our modeling one-by-one. First, we discuss the simulation used in this study. Afterwards, we talk about the forward modeling of galaxy catalogs in the standard HOD modeling framework as well as the decorated HOD framework. Then, we provide an overview of the two summary statistics of the galaxy catalogs that we used in our inference.

4.3.1 Simulation

For the simulations used in this work, we make use of the Rockstar (**?**) halo catalogs in the $z = 0$ snapshot of the Small MultiDark of Planck cosmology (referred to as `SMDP`) (**?**). This high resolution N-body simulation (publicly available at `https://www.cosmosim.org`) was carried out using the **GADGET-2 (see ? and references therein) code**, following the Planck ΛCDM cosmological parameters $\Omega_{\rm m} = 0.307$, $\Omega_{\rm b} = 0.048$, $\Omega_{\Lambda} = 0.693$, $\sigma_8 = 0.823$, $n_{\rm s} = 0.96$, $h = 0.678$. The Box size for this N-body simulation is 0.4 h^{-1}Gpc, the number of simulation particles is 3840^3, the mass per simulation particle $m_{\rm p}$ is 9.6×10^7 $h^{-1}M_{\odot}$, and the gravitational softening length ϵ is 1.5 h^{-1}kpc.

In the `SMDP` simulation, as discussed in **?**, the Rockstar algorithm can reliably resolve halos with $\geq$ 100 **particles, which corresponds to** $M_{\rm vir} \geq 9.6 \times 10^9$ $h^{-1}M_{\odot}$**. The** `SMDP` **simulation provides a number of advantages by satisfying both the size and resolution requirements of studying the galaxy–halo connection in a wide range of luminosity thresholds. For fainter galaxy samples, the faintest galaxies reside in lower mass halos, which requires high resolution. Meanwhile for luminous galaxy samples, their lower number densities requires a large comoving volume.** Furthermore, since we are studying the higher order halo occupation statistics, the concentration-dependence in particular, it is important to use a simulation that can resolve the internal structure of halos. **In the context of Subhalo-Abundance Matching models, which requires subhalo completeness in the low mass limit, the** `SMDP` **simulation has been used to model the faintest galaxy samples in the SDSS data (see ?).** The added advantage of using the `SMDP` simulation over some of the other industrial simulation boxes commonly used in the literature such as `Bolshoi` (**??**) simulation is its larger comoving volume. Larger volume makes this simulation more suitable for performing inference with $L_{\star}$ (corresponding to $M_{\rm r} \sim -20.44$, see **?**) and more luminous than $L_{\star}$ galaxy

samples that occupy larger comoving volumes.

4.3.2 Halo occupation modeling

4.3.2.1 standard model without assembly bias

For our standard HOD model, we assume the HOD parameterization from **?**. According to this model, a dark matter halo can host a central galaxy and some number of satellite galaxies. The occupation of the central galaxies follows a nearest-integer distribution, and the occupation of the satellite galaxies follows a Poisson distribution. The expected number of centrals and satellites as a function of the host halo mass of M_h are given by the following equations

$$\langle N_{\rm c}|M_h\rangle = \frac{1}{2}\Big[1+\Big(\frac{\log M_h - \log M_{\rm min}}{\sigma_{\log {\rm M}}}\Big)\Big], \tag{4.1}$$

$$\langle N_{\rm s}|M_h\rangle = \Big(\frac{M_h - M_0}{M_1}\Big)^{\alpha}. \tag{4.2}$$

For populating the halos with galaxies, we follow the procedure described in **?**, and **?**. The central galaxies are assumed to be at the center of the host dark matter halos. We assume that the central galaxies are at rest with respect to the bulk motion of the halos and their velocities are given by the velocity of the center of mass of their host halo. Note that this assumption is shown to be violated in brighter than $L_\star$ galaxy samples (see **?**). But since we are not considering the redshift space 2PCF multipoles in our study, we do not expect this velocity bias to impact our inference. We place the satellite galaxies within the virial radius of the halo following a Navarro-Frenk-White profile (hereafter NFW; **?**). This approach is different from other simulation-based halo occupation modeling techniques (see **??**) in that the positions of the satellites are not assigned to the dark matter particles

in the N-body simulation. **The concentration of the NFW profile is given by the empirical mass-concentration relation provided by ?.** The velocities of the satellite galaxies are given by two components. The first component is the velocity of the host halo. The second component is the velocity of the satellite galaxy with respect to the host halo which is computed following the solution to the NFW profile Jeans equations (**?**). We refer the readers to **?** for a more comprehensive and detailed discussion of the forward modeling of the galaxy mock catalogs.

4.3.2.2 Model with Assembly bias

Now let us provide a brief overview of HOD modeling with `Heaviside Assemblybias` (referred to as the decorated HOD) introduced in **?**. At a fixed halo mass M_h, halos are split into two populations: population of halos with the 0.5-percentile of highest concentration, and population of halos with 0.5 percentile of lowest concentration. For simplicity, we call the first population "type-1" halos, and the second population "type-2" halos. In the decorated HOD model, the expected number of central and satellite galaxies at a fixed halo mass M_h in the two populations are given by

$$\langle N_{c,i}|M_h, c\rangle = \langle N_c|M_h\rangle + \Delta N_{c,i},\ i = 1, 2 \tag{4.3}$$

$$\langle N_{s,i}|M_h, c\rangle = \langle N_s|M_h\rangle + \Delta N_{s,i},\ i = 1, 2 \tag{4.4}$$

where $\langle N_c|M_h\rangle$ and $\langle N_s|M_h\rangle$ are given by Eqs 4.1 and 4.2 respectively, and we have $\Delta N_{s,1} + \Delta N_{s,2} = 0$, and $\Delta N_{c,1} + \Delta N_{c,2} = 0$. These two conditions ensure the conservation of HOD. At a given host halo mass M_h, the central occupation of the the two populations follows a nearest-integer distribution with the first moment given by 4.3; and the satellite occupation

of the the two populations follows a Poisson distribution with the first moment given by 4.4.

In this occupation model, the allowable ranges that quantities $\Delta N_{c,1}$ and $\Delta N_{s,1}$ can take are given by

$$\max\{-\langle N_c|M_h\rangle, \langle N_c|M_h\rangle - 1\} \leq \Delta N_{c,i} \leq \min\{\langle N_c|M_h\rangle, 1 - \langle N_c|M_h\rangle\}, \tag{4.5}$$

$$-\langle N_s|M_h\rangle \leq \Delta N_{s,i} \leq \langle N_s|M_h\rangle. \tag{4.6}$$

Afterwards, the assembly bias parameter $\mathcal{A}$ is defined in the following way:

$$\Delta N_{\alpha,1}(M_h) = |\mathcal{A}_\alpha| \Delta N^{\max}_{\alpha,1}(M_h) \text{ if } \mathcal{A}_\alpha > 0, \tag{4.7}$$

$$\Delta N_{\alpha,1}(M_h) = |\mathcal{A}_\alpha| \Delta N^{\min}_{\alpha,1}(M_h) \text{ if } \mathcal{A}_\alpha < 0, \tag{4.8}$$

where the subscript $\alpha = c, s$ stands for the centrals and satellites respectively, and $\Delta N^{\max}_{\alpha,1}(M_h)$, $\Delta N^{\min}_{\alpha,1}(M_h)$ are given by Eqs. 4.5 and 4.6.

For a given $\mathcal{A}_\alpha$, once $\Delta N_{\alpha,1}$ is computed using equation (4.7)—if $\mathcal{A}_\alpha > 0$—or equation (4.8)—if $\mathcal{A}_\alpha < 0$—, $\Delta N_{\alpha,2} = 1 - \Delta N_{\alpha,1}$ is computed. At a fixed halo mass M_h, once the first moments of occupation statistics for the *type*-1 and *type*-2 halos are determined, we perform the same procedure described in 4.3.2.1 to populate the halos with mock galaxies.

4.3.2.3 Redshift-space distortion

Once the halo catalogs are populated with galaxies, the real-space positions and velocities of all mock galaxies are obtained. The next step is applying a redshift-space distortion transformation by assuming plane-parallel approximation. Our use of plane parallel approximation is justified because of the narrow redshift range of the SDSS main galaxy sample

considered in this study. If we assume that the $\hat{z}$ axis is the line-of-sight direction, then with the transformation $(X, Y, Z) \rightarrow (S_x, S_y, S_z) = (X, Y, Z + v_z(1+z)/H(z))$ for each galaxy with the real space coordinates (X, Y, Z), velocities (v_x, v_y, v_z), and redshift z, we obtain the redshift-space coordinate of the produced mock galaxies. Here we assume $z \simeq 0$, and therefore transformation is given by $(X, Y, Z) \rightarrow (X, Y, Z + v_z/H_0)$.

4.3.3 Model Observables

As described in **?** and **?**, this approach makes no appeal to the fitting functions used in the analytical calculation of the 2PCF. The accuracy of these fitting functions is limited (**???**). Our approach also does not face the known issues of the treatment of halo exclusion and scale-dependent bias that can lead to potential inaccuracies in halo occupation modeling (see **?**).

The projected 2PCF $w_p(r_p)$ can be computed by integrating the 3D redshift space 2PCF $\xi(r_p, \pi)$ along the line-of-sight (where r_p and π denote the projected and line-of-sight separation of galaxy pairs respectively):

$$w_p(r_p) = 2 \int_0^{\pi_{max}} \xi(r_p, \pi)\, d\pi \tag{4.9}$$

For our 2PCF calculations, we use the w_p measurement functionality of the fast and publicly available pair-counter code `CorrFunc` (**?** , available at https://github.com/manodeep/Corrfunc). To be consistent with the SDSS measurements described in Section 4.4, $w_\mathrm{p}(r_\mathrm{p})$ is obtained by the line-of-sight integration to $\pi_{max} = 40\ h^{-1}$Mpc. Note that $w_p(r_p)$ is measured in units of h^{-1}Mpc. To be consistent with **?**, we use the same binning (as specified in Section 4.4) to measure w_p. In addition to the projected 2PCF, we use the number density given by the number of mock galaxies divided by the comoving volume of the `SMDP`

simulation.

Note that a full forward model of the data requires running the simulation at different redshifts, generation of light cones, accounting for the complex survey geometry and systematic errors such as fiber collisions. Using the $z = 0$ output of the SMDP simulation in our forward model of the spatial distribution of galaxies is only an approximation. This approximation can be justified by the small redshift range of the SDSS DR7 main galaxy sample. As described in **?**, using random catalogs with angular window function of the data in measurements of galaxy clustering accounts for the geometry of the data. As described in Section 4.4, the fiber collision correction method of **?** is applied to the SDSS clustering measurements used in this study. Therefore we do not account for that effect in our forward model.

4.4 Data

We focus on the measurements made on the volume-limited luminosity-threshold main sample of galaxies in the SDSS spectroscopic survey. In this section, we briefly describe the measurements used in our study for finding constraints on the assembly bias as well as the HOD parameters.

The measurements consist of the number density n_g and the projected 2PCF $w_p(r_p)$ made by **?** for the volume-limited sample of galaxies in NYU Value Added Galaxy Catalog (**?**) constructed from the SDSS DR7 main galaxy sample (**?**). In particular, eight volume-limited luminosity-threshold samples are constructed with maximum absolute luminosity in r-band of -18, 18.5, -19, -19.5, -20, -20.5, -21, and -21.5. Qualitatively, these samples are constructed in a similar way to those constructed in **?**. For detailed differences between the samples in **?** and **?**, we refer the reader to the Table 1 and Table 2 in those papers respectively.

The projected 2PCFs are measured in 12 logarithmic r_p bins (in units of h^{-1}Mpc) of width $\Delta \log(r_{\rm p}) = 0.2$, starting from $r_{\rm p} = 0.1\ h^{-1}$Mpc. For all luminosity threshold samples, the integration along the line-of-sight (4.9) are performed to $\pi_{max} = 40\ h^{-1}$Mpc.

The 2PCF measurement of each luminosity-threshold sample is accompanied by a covariance matrix constructed using 400 jackknife sub-samples of the data. The number density measurements are also accompanied by uncertainties measured using the jackknife method. Furthermore, the covariance between the number denisty and the projected 2PCF measurements are neglected. As **?** shows, parameter estimation using jackknife covariance matrices is conservative as the jackknife method overestimates the errors in the observations.

The advantage of using these measurements is that the effects of fiber collision systematic errors on the two-point statistics are corrected for (with the method described in **?**), and therefore, these measurements provide accurate small scale clustering measurements. The assembly bias parameters introduced in section 5.3 can have a 10-percent level impacts on galaxy clustering (**?**). Presence of assembly bias in the satellite population impacts the very small-scale clustering (**?**). Moreover as **?** demonstrates, the scale-dependence of the halo assembly bias has a pronounced bump in the 1-halo to 2-halo transition regime (1$\sim$2 h^{-1}Mpc). This scale can be impacted by fiber collision systematics. Precise investigation of the possible impact of this signal on the galaxy clustering modeling requires accurate measurements of 2PCF on small scales. The method of **?** is able to recover the true $w_{\rm p}$ with $\sim 6\%$ accuracy in small scales ($r_{\rm p} = 0.1\ h^{-1}$Mpc) and with $\sim 2.5\%$ at relatively large scales $r_{\rm p} \sim 30\ h^{-1}$Mpc.

Note that the comoving volume of the N-body simulation used in this investigation is $64\ \times 10^6\ h^{-3}{\rm Mpc}^3$ which is larger than the comoving volume of all the luminosity-threshold samples in the SDSS data considered in this study except the two most luminous samples. The comoving volumes of the $M_{\rm r} < -21,\ 21.5$ samples are 71.74 and 134.65 (in units of

$10^6 h^{-3}\mathrm{Mpc}^3$) respectively. Since we are not studying very large scale clustering ($r_{p,\,\mathrm{max}} \leq 25\ h^{-1}\mathrm{Mpc}$), using a slightly smaller box for those samples is justified.

4.5 Analysis

4.5.1 Inference setup

Given the SDSS measurements described in Section 4.4, we aim to constrain the HOD model without assembly bias (described in 4.3.2.1), and the HOD model with assembly bias (described in 4.3.2.2) for each luminosity-threshold sample, by sampling from the posterior probability distribution $p(\theta|d) \propto p(d|\theta)\pi(\theta)$ where θ denotes the parameter vector and d denotes the data vector. In the standard HOD modeling θ is given by

$$\theta = \{\log M_{\mathrm{min}},\ \sigma_{\log \mathrm{M}},\ \log M_0,\ \alpha,\ \log M_1\}, \tag{4.10}$$

and in the HOD modeling with assembly bias we have

$$\theta = \{\log M_{\mathrm{min}},\ \sigma_{\log \mathrm{M}},\ \log M_0,\ \alpha,\ \log M_1,\ \mathcal{A}_{\mathrm{cen}},\ \mathcal{A}_{\mathrm{sat}}\}, \tag{4.11}$$

Furthermore, data (denoted by d) is the combination of $[n_g, w_p(r_p)]$. The negative log-likelihood (assuming negligible covariance between n_g and $w_p(r_p)$) is given by

$$-2\ln p(d|\theta) \quad = \quad \frac{[n_g^{\mathrm{data}} - n_g^{\mathrm{model}}]^2}{\sigma_n^2} + \Delta w_{\mathrm{p}}^{\mathrm{T}} \widehat{C^{-1}} \Delta w_{\mathrm{p}} \ +\ \mathrm{const.}, \tag{4.12}$$

where Δw_{p} is a 12 dimensional vector, $\Delta w_{\mathrm{p}}(r_{\mathrm{p}}) = w_{\mathrm{p}}^{\mathrm{data}}(r_{\mathrm{p}}) - w_{\mathrm{p}}^{\mathrm{model}}(r_{\mathrm{p}})$, and $\widehat{C^{-1}}$ is the estimate of the inverse covariance matrix that is related to the inverse of the jackknife

Table 4.1: **Prior Specifications**: The prior probability distribution and its range for each of the parameters. All mass parameters are in unit of $h^{-1}M_{\odot}$. The parameters marked by $*$ are only used in the Heaviside Assembly bias modeling and by definition are bounded between -1 and 1.

Parameter	Prior	Range
α	Uniform	[0.85, 1.45]
$\sigma_{\log \mathrm{M}}$	Uniform	[0.05, 1.5]
$\log M_0$	Uniform	[10.0, 14.5]
$\log M_{\rm min}$	Uniform	[10.0, 14.0]
$\log M_1$	Uniform	[11.5, 15.0]
$\mathcal{A}^*_{\rm cen}$	Uniform	[-1.0, 1.0]
$\mathcal{A}^*_{\rm sat}$	Uniform	[-1.0, 1.0]

covariance matrix (provided by **?**) $\widehat{C}^{-1}$, following **?**:

$$\widehat{C^{-1}} = \frac{N-d-2}{N-1}\,\widehat{C}^{-1}, \tag{4.13}$$

where $N = 400$ is the number of the jackknife samples, and $d = 12$ is the length of the data vector $w_{\rm p}$. Another important ingredient of our analysis is specification of the prior probabilities $\pi(\theta)$ over the parameters of the halo occupation models considered in this study. For both models, we use uniform flat priors for all the parameters. The prior ranges are specified in the Table 4.5.1. Note that a uniform prior between -1 and 1 is chosen for assembly bias parameters since these parameters are, by definition, bounded between -1 and 1.

For sampling from the posterior probability, given the likelihood function (see equation 4.12) and the prior probability distributions (see Table 4.5.1), we use the affine-invariant ensemble MCMC sampler (**?**) and its implementation `emcee` (**?**). In particular, we run the `emcee` code with 20 walkers and we run the chains for at least 10000 iterations. We discard the first one-third part of the chains as burn-in samples and use the reminder of the chains

as production MCMC chains. Furthermore, we perform Gelman-Rubin convergence test (**?**) to ensure that the MCMC chains have reached convergence.

4.6 Results and Discussion

4.6.1 Constraints and Interpretations

In this section, we present the constraints derived for the two assembly bias parameters: the satellite assembly bias parameter ($\mathcal{A}_{\rm sat}$) and the central assembly bias parameter $\mathcal{A}_{\rm cen}$. As shown in Figure 5.2, for all the eight luminosity-threshold samples in the SDSS DR7 data, our constraints on the parameter $\mathcal{A}_{\rm sat}$ are consistent with zero. On the other hand, our constraints on the parameter $\mathcal{A}_{\rm cen}$— albeit not tightly constrained— show a trend which can be summarized as the following. In the most luminous galaxy samples, i.e. $M_{\rm r} < -21.5$ and $M_{\rm r} < -21$, $\mathcal{A}_{\rm cen}$ is poorly constrained and the constraints are equivalent to zero. As we investigate less luminous samples, $M_{\rm r} < -20.5, -20, -19.5$, our constraints on $\mathcal{A}_{\rm cen}$ shift toward positive values, with the $M_{\rm r} < -20$ sample favoring the highest values for $\mathcal{A}_{\rm cen}$. Furthermore, the posterior constraints on the assembly bias parameters in the slightly fainter samples, i.e. $M_{\rm r} < -19$ and i.e. $M_{\rm r} < -18.5$, are consistent with zero. In the faintest galaxy sample, i.e. $M_{\rm r} < -18$, our constraints favor negative values of $\mathcal{A}_{\rm cen}$.

The underlying theoretical consideration for explaining the assembly bias of the central and satellite galaxies are different. The large scale clustering—or the two halo term in the galaxy clustering—is mainly governed by the clustering of the central galaxies. The central galaxy clustering can be thought as the weighted average over the halo clustering. The large scale bias of the dark matter halo clustering depends not only on mass, but also on the other properties of halos beyond mass, such as concentration (**???**), spin (**?**), formation time (**??**),

and maximum circular velocity of the halo $V_{\rm max}$ (**?**).

In particular, findings of **?** and **?** have demonstrated that for halos with mass bellow the collapse mass ($M \leq M_{\rm col} \simeq 10^{12.5}\ M_{\odot}$) the large scale bias of high-$V_{\rm max}$ (or equivalently high-c halos at a fixed halo mass) is larger than that of the low-$V_{\rm max}$ (low-c halos). This signal reverses and weakens for the high mass halos ($M \geq M_{\rm col}$). Note that the halo concentration traces the maximum circular velocity $V_{\rm max}$ such that halos with higher V_{max} have higher concentration and vice versa (see **?**). For halos described by NFW profile, at a fixed halo mass, halos with higher values of concentration have higher values of $V_{\rm max}$.

Furthermore, investigation of the scale dependence of halo assembly bias has shown that the ratio of the bias of high-$V_{\rm max}$ halos and the low-$V_{\rm max}$ halos has a bump-like feature in the quasi-linear scales $\sim 0.5\ {\rm Mpc}h^{-1} - 5{\rm Mpc}h^{-1}$. From the theoretical standpoint, this phenomenon has been attributed to a population of distinct halos with $M \sim 10^{11.7}\ h^{-1}M_{\odot}$ at the present time that are close to the most massive groups and clusters (**?**, and see **?** for the observational investigation of this signal by means of galaxy-galaxy lensing). The clustering of these population of halos is therefore dictated by that of the massive halos. Note that this scale-dependent bias feature vanishes in high mass halos.

Consequently, at a fixed halo mass less than $M_{\rm col}$, assignment of more central galaxies to the high-c halos (higher expected number of central galaxies in the high-c halos) gives rise to a boost in the galaxy clustering in the linear scales as well as in regimes corresponding to the one-halo to two-halo transition. For the more massive halos ($M \geq M_{\rm col}$), we expect the large-scale clustering boost to reverse sign, and the quasi-linear bump feature to vanish.

Figure 4.2 demonstrates the 68% and 95% posterior predictions for the projected 2PCF w_p from the occupation model without assembly bias (shown in red) and the occupation model with assembly bias (shown in blue) for all eight luminosity-threshold samples. For the brightest galaxies, $M_r < -21.5$ and $M_r < -21.0$, the posterior prediction of w_p from

the two models are consistent with one another. Note that these galaxies reside in the most massive halos ($M > M_{\rm col}$) for which the scale-dependence of the halo assembly bias and the difference between the large-scale bias of the high-c and low-c halos become negligible.

Figure 4.3 shows the fractional difference between the 68% and 95% posterior predictions of w_p and the SDSS data. It is evident from Figure 4.3 that some improvement on modeling the clustering of the samples of $L_\star$ and slightly less brighter than $L_\star$ galaxies can be achieved by employing the more complex halo occupation model with assembly bias. As a result of apportioning more central galaxies to the high-c halos relative to the low-c halos, in the samples with luminosity thresholds of $M_{\rm r} < -20.5, -20, -19.5$, the posterior predictions for w_p are slightly improved in the intermediate scales ($1 \sim 2$ Mpch^{-1}) and large scales. This can be also noted in significantly lower χ^2 values—at the cost more model flexibility and higher degrees of freedom—achieved by the assembly bias model in these luminosity-threshold samples (see Table 4.6.1).

In the sample of galaxies with the luminosity threshold $M_{\rm r} < -19, -18.5$, the constraints on $\mathcal{A}_{\rm cen}$ become consistent with zero with the tendency towards positive values for the $M_{\rm r} < -19$ sample and towards more negative values for the $M_{\rm r} < -18.5$ sample. As shown in Figures 4.2 and 4.3, for the $M_{\rm r} < -19$ ($M_{\rm r} < -18.5$) sample this results in slightly higher (lower) posterior predictions for w_p in the intermediate toward large scales. Overall, for these two samples, the assembly bias parameters remain largely unconstrained and the decorated HOD model does not yield better χ^2 values.

Finally, In the faintest sample ($M_{\rm r} < -18$), negative constraints on the parameter $\mathcal{A}_{\rm cen}$ results in higher expected number of centrals in the low-c halos, which at a fixed halo mass, cluster less strongly. This affects both the large scale bias and the intermediate regimes as a result of the scale-dependent bump feature (see Figures 4.2 and 4.3). Furthermore, the model with assembly bias provides better fit to the SDSS data in this luminosity-threshold

sample.

The luminosity dependent trend in the constraints on the central assembly bias for the six dimmest samples can be attributed to the fact that the halo concentration is highly correlated with the maximum circular velocity $V_{\rm max}$ which is a tracer of the potential well of dark matter halos (**?**). In a dark matter halo described by an NFW profile, the depth of the gravitational potential well of dark matter halos can be directly measured by the maximum circular velocity $V_{\rm max}$ (**?**). In particular, the magnitude of the potential well at the center of an NFW halo—where the central galaxy is assumed to reside—scales as $V_{\rm max}^2$. More specifically, we have:

$$\Phi(r=0) = -\Big(\frac{V_{\rm max}}{0.465}\Big)^2, \tag{4.14}$$

where $\Phi(r=0)$ is the central potential of an NFW profile. Note that $V_{\rm max}$ is also the quantity often used in the abundance matching technique in which the luminosity of galaxies is monotonically matched to $V_{\rm max}$ (see for example **????**). The trend between the constraint on $\mathcal{A}_{\rm cen}$ and the luminosity threshold of the samples may suggest that at a *fixed halo mass*, the central galaxies in the dimmest samples ($M_{\rm r}$ <-18 , -18.5) tend to reside in dark matter halos with shallower gravitational well. In brighter galaxy samples ($M_{\rm max}$ <-19 , -19.5, -20, -20.5), at a *fixed halo mass*, the central galaxies have a tendency to reside in dark matter halos with deeper gravitational potential well.

The satellite assembly bias can only significantly alter the galaxy clustering at small-to-intermediate scales. Assigning more satellite galaxies to lower (or higher) concentration halos affects the one-halo term through increasing the satellite-satellite pair counts $\langle N_{\rm s}N_{\rm s}\rangle$. This results in boosting the small-scale clustering. But as pointed out by **?**, the amount by which small-scale clustering increases also depends on the sign of the central assembly bias parameter $\mathcal{A}_{\rm cen}$. Formation history of the halos can lead to the dependence of the abundance of subhalos on halo concentration (**??**) at fixed halo mass, and since the occupation of the

satellite galaxies is related to the abundance of subhalos, the satellite occupation may depend on halo concentration.

However, our results suggest that for all the luminosity-threshold samples considered in this study, the satellite assembly bias parameter is largely unconstrained and consistent with zero. We do not expect the galaxy clustering data to be a sufficient statistics for obtaining constraints on the satellite assembly bias. Group statistics probes the high mass end of the galaxy–halo connection and is sensitive to the parameters governing the satellite population (see **??**). Therefore these measurements may shed some light on potential presence of assembly bias in satellite population.

It is important to note that the halo mass range in which the central assembly bias $\mathcal{A}_{\rm cen}$ affects the central galaxy population is the mass range in which the condition $0 < \langle N_{\rm c}|M\rangle < 1$ is met. Consequently, larger scatter parameter $\sigma_{\log \rm M}$ increases the dynamical mass range in which assembly bias affects the galaxy clustering. Note that in the luminosity regimes for which we obtain tighter constraints on $\mathcal{A}_{\rm cen}$, the best-estimate values of the scatter parameter $\sigma_{\log \rm M}$ appear to be higher in the model with assembly bias. This is evident in Figure 4.8. The model with assembly bias tends to push $\sigma_{\log \rm M}$ to higher values. This can be attributed to the tendency of this model to increase the effective dynamical mass range of central assembly bias.

As shown in Table 4.6.1, in the HOD model with assembly bias, the constraints found on the scatter parameter are not tight. This is in keeping with the results of **?** which uses the same SDSS measurements and finds that the scatter parameter remains largely unconstrained when only n_g and w_p are used as observables. Note that scatter is better constrained for the most luminous galaxy samples (this is attributed to the steep dependence of the halo bias and halo mass function on halo mass in the high mass end). But since these samples live in the most massive halos, we do not expect the tighter constraints on scatter to help us constrain

the central assembly bias parameter. **?** shows that by employing additional measurements such as the monopole (ξ_0), quadruple (ξ_2), and hexadecapole (ξ_4), one can obtain tighter constraints on the scatter parameter. Tightening the constraints on the scatter parameter can lead to more precise inference of the central assembly bias parameter.

As shown in Table 4.6.1, our constraints on the underlying standard HOD model obtained from the model with assembly bias and the model without assembly bias are in good agreement. The only cases in which there are mild tensions between the constraints found from the two models on the underlying HOD parameters, are the $M_{\rm r} < -20$ and the $M_{\rm r} < -20.5$ samples. However, these tensions are still within one-sigma level. For instance, Figure 4.8 shows that in the $M_{\rm r} < -20.5$ sample, the constraint on the parameter α found from the model without assembly bias favor slightly higher values than the constraint found from the model with assembly bias. Also the scatter parameter $\sigma_{\log \rm M}$ is more tightly constrained in the standard HOD model. However, it is important to emphasize that these constraints are still in agreement with each other within a one-sigma level.

? shows that in the mock catalogs that exhibit significant levels of assembly bias, using a simple mass-only occupation model can lead to considerable biases in inference of the galaxy–halo connection parameters. Although we cannot rule out moderate levels of assembly bias in our findings, we do not find any considerable discrepancy between the two models in terms of estimating the underlying HOD parameters.

A few galaxy–halo connection methods have been proposed in the literature that give rise to assembly bias in the galaxy population. **? demonstrates that the abundance matching techniques based on $V_{\rm max}$ (???) exhibit some levels of assembly bias.** We aim to provide a comparison between the impact of assembly bias on clustering in these mock catalogs and the mock catalogs predicted from our constraints on the decorated HOD model for $L_\star$-type galaxies. In particular, we consider the abundance matching catalogs produced

by **?**. These catalogs have been extensively studied for examining potential systematic effects of galaxy assembly bias on cosmological (**?**) and halo occupation (**?**) parameter inferences.

This abundance matching catalog was built based on the `Bolshoi` N-body simulation (**?**) using the adaptive refinement tree code (ART **?**). The Box size for this simulation is 250 h^{-1}Mpc, the number of simulation particles is 2048^3, the mass per simulation particle $m_{\rm p}$ is 1.35×10^8 $h^{-1}M_{\odot}$, and the gravitational softening length ϵ is 1 h^{-1}kpc. The halos and subhalos in this simulation are identified using the ROCKSTAR algorithm (**?**). The **?** catalogs make use of $V_{\rm peak}$ (maximum $V_{\rm max}$ throughout the assembly history of halo) as the subhalo property to be matched to galaxy luminosity.

As noted by **?** and **?**, these galaxy mock catalogs show significant levels of assembly bias in the central galaxy population. This has been demonstrated by investigating the difference in w_p between the randomized mock catalogs and the original mock catalogs. Randomization is performed in a procedure described in **?** which we briefly summarize here: First, halos are divided into different bins of halo mass with width of 0.1 dex. Then all central galaxies are shuffled among all halos within each bin. Once the centrals have been shuffled, within each bin, the satellite systems are shuffled among all halos in that mass bin, preserving their relative distance to the center of halo. This procedure preserves the HOD, but erases any dependence of the galaxy population on the assembly history of halos. Therefore, assembly bias is erased in the randomized galaxy catalog.

For $L_{\star}$ galaxies, the difference in w_p between the randomized and the original catalogs of **?** is shown in Figure 4.4 with the red curves. As demonstrated in Figure 4.4 (and as previously noted by **??**), the relative difference in w_p is only significant in relatively large scales ($r_p > 1$Mpc h^{-1}). This implies that in these catalogs, only the central occupation is affected by assembly bias. This is in agreement with our findings.

Furthermore, we investigate whether the impact of assembly bias on galaxy clustering

predicted by our findings are in agreement with the abundance matching catalogs of **?**. First, we make random draws from the posterior probability distribution function over the parameters of the model with assembly bias. Then, we create mock catalogs with these random draws, and then we compute the difference in w_p between the randomized catalogs and the original catalogs. The relative difference in w_p predicted from our constraints are shown with blue curves in Figure 4.4.

We note that our findings follow the same trend. That is, we see negligible difference in the small scale clustering and more considerable differences in w_p on larger scales ($r_p >$ 1 Mpch^{-1}). Similar to findings of **?**, **?**, and **?**, we see that the impact of assembly bias on galaxy clustering becomes less significant in brighter galaxy samples. Furthermore, we notice that our mock catalogs favor more moderate changes in galaxy clustering as a result of assembly bias.

4.6.2 model comparison

We want to address this question that whether the constraints on the model with assembly bias and the model without assembly bias given the galaxy clustering data lead us to claim that assembly bias is strongly supported by the observations or not. Within the standard HOD framework, the distribution of the galaxies is modeled using a simple description based on the *mass*-only ansatz: $P(N|M)$. The decorated HOD model however, provides a more complex description of the data by adding a secondary halo property (halo concentration in this study) and a more flexible occupation model: $P(N|M, c)$.

In order to investigate whether the higher level of model complexity is demanded by the observations or not, we present model comparison between the models with and without assembly bias. In particular, we make use of two simple methods for model comparison: *Akaike Information Criterion* (AIC, **?** , see **?** for detailed discussion on AIC), and *Bayesian*

Table 4.2: **Constraints**: Constraints on the parameters of the HOD models with and without assembly bias. All mass parameters are in unit of $h^{-1}M_{\odot}$. The best-estimates and the error bars correspond to the 50% quantile and 68% confidence intervals obtained from the marginalized posterior probability pdfs. The last column is χ^2 per degrees of freedom (dof), where $dof = N_{data} - N_{par}$

$M_{\rm r,lim}$	$\log M_{\rm min}$	$\sigma_{\log \rm M}$	$\log M_0$	α	$\log M_1$	$\mathcal{A}_{\rm cen}$	$\mathcal{A}_{\rm sat}$	χ^2/dof
-18	$11.56^{+0.21}_{-0.25}$	$1.05^{+0.31}_{-0.52}$	$10.84^{+0.77}_{-0.59}$	$0.98^{+0.05}_{-0.05}$	$12.50^{+0.10}_{-0.10}$	–	–	14.51/8
-18	$11.53^{+0.23}_{-0.21}$	$1.08^{+0.28}_{-0.51}$	$10.86^{+0.81}_{-0.62}$	$0.97^{+0.05}_{-0.04}$	$12.56^{+0.09}_{-0.10}$	$-0.67^{+0.55}_{-0.25}$	$-0.30^{+1.09}_{-0.54}$	7.52/6
-18.5	$11.67^{+0.29}_{-0.25}$	$0.83^{+0.45}_{-0.53}$	$10.85^{+0.64}_{-0.60}$	$1.02^{+0.04}_{-0.04}$	$12.69^{+0.08}_{-0.08}$	–	–	6.17/8
-18.5	$11.60^{+0.31}_{-0.20}$	$0.74^{+0.49}_{-0.46}$	$10.73^{+0.69}_{-0.51}$	$1.01^{+0.05}_{-0.05}$	$12.72^{+0.09}_{-0.10}$	$0.02^{+0.67}_{-0.62}$	$0.07^{+0.53}_{-0.59}$	6.23/6
-19	$11.74^{+0.37}_{-0.18}$	$0.62^{+0.52}_{-0.39}$	$10.82^{+0.62}_{-0.56}$	$1.04^{+0.04}_{-0.04}$	$12.87^{+0.09}_{-0.09}$	–	–	8.69/8
-19	$11.71^{+0.37}_{-0.16}$	$0.58^{+0.53}_{-0.38}$	$10.75^{+0.66}_{-0.52}$	$1.03^{+0.04}_{-0.05}$	$12.90^{+0.10}_{-0.09}$	$0.36^{+0.44}_{-0.62}$	$-0.01^{+0.56}_{-0.54}$	8.87/6
-19.5	$11.78^{+0.37}_{-0.13}$	$0.51^{+0.53}_{-0.31}$	$11.09^{+0.69}_{-0.73}$	$1.06^{+0.03}_{-0.04}$	$13.03^{+0.08}_{-0.07}$	–	–	6.80/8
-19.5	$11.82^{+0.41}_{-0.17}$	$0.62^{+0.54}_{-0.44}$	$11.11^{+0.68}_{-0.77}$	$1.03^{+0.04}_{-0.06}$	$13.03^{+0.10}_{-0.08}$	$0.52^{+0.32}_{-0.47}$	$-0.01^{+0.66}_{-0.49}$	5.56/6
-20	$12.01^{+0.17}_{-0.08}$	$0.32^{+0.32}_{-0.19}$	$11.69^{+0.54}_{-0.99}$	$1.08^{+0.03}_{-0.05}$	$13.32^{+0.08}_{-0.07}$	–	–	13.45/8
-20	$12.23^{+0.39}_{-0.24}$	$0.76^{+0.41}_{-0.43}$	$11.66^{+0.62}_{-0.94}$	$1.00^{+0.08}_{-0.05}$	$13.26^{+0.08}_{-0.09}$	$0.81^{+0.12}_{-0.26}$	$-0.15^{+0.33}_{-0.31}$	8.12/6
-20.5	$12.31^{+0.06}_{-0.06}$	$0.21^{+0.14}_{-0.11}$	$12.36^{+0.27}_{-0.77}$	$1.11^{+0.08}_{-0.08}$	$13.56^{+0.09}_{-0.09}$	–	–	11.40/8
-20.5	$12.37^{+0.16}_{-0.09}$	$0.44^{+0.28}_{-0.26}$	$12.38^{+0.27}_{-0.54}$	$1.05^{+0.08}_{-0.07}$	$13.60^{+0.10}_{-0.08}$	$0.81^{+0.15}_{-0.43}$	$-0.11^{+0.26}_{-0.27}$	6.82/6
-21	$12.73^{+0.14}_{-0.07}$	$0.32^{+0.22}_{-0.17}$	$12.62^{+0.48}_{-1.36}$	$1.17^{+0.10}_{-0.16}$	$14.01^{+0.08}_{-0.10}$	–	–	7.34/8
-21	$12.81^{+0.28}_{-0.11}$	$0.47^{+0.41}_{-0.25}$	$12.51^{+0.61}_{-1.17}$	$1.08^{+0.17}_{-0.13}$	$14.02^{+0.08}_{-0.10}$	$0.33^{+0.51}_{-0.64}$	$-0.18^{+0.41}_{-0.35}$	7.21/6
-21.5	$13.44^{+0.11}_{-0.09}$	$0.60^{+0.09}_{-0.11}$	$12.57^{+0.77}_{-1.27}$	$1.33^{+0.07}_{-0.21}$	$14.53^{+0.05}_{-0.07}$	–	–	3.29/8
-21.5	$13.43^{+0.12}_{-0.06}$	$0.60^{+0.12}_{-0.08}$	$12.59^{+0.64}_{-1.25}$	$1.27^{+0.11}_{-0.19}$	$14.54^{+0.05}_{-0.05}$	$-0.24^{+0.50}_{-0.40}$	$-0.30^{+0.86}_{-0.49}$	3.37/6

Information Criteria (BIC, **?**). BIC and AIC are more computationally tractable than alternatives approaches such as computing the fully marginalized likelihood. The underlying assumption of these information criteria is that models that yield higher likelihoods are more preferable, but at the same time, models with more flexibility are penalized.

Suppose that $\mathcal{L}^\star$ is the maximum likelihood achieved by the model, $N_{\rm par}$ is the number of free parameters in the model, and $N_{\rm data}$ is the number of data points in the data set. Then we have

$$\mathrm{BIC} = -2\ \ln \mathcal{L}^\star + N_{\rm par}\ \ln N_{\rm data}\ , \tag{4.15}$$

$$\mathrm{AIC} = -2\ \ln \mathcal{L}^\star + 2N_{\rm par}\ . \tag{4.16}$$

Given a data set, models with lower value of AIC and BIC are more desired. That is, in order for the higher model complexity (given by $N_{\rm par}$) to be justified, $\mathcal{L}^\star$ must be sufficiently higher. Therefore in this formulation, models that deliver lower information criteria scores are more preferable.

Figure 4.5 shows the comparison between the BIC and AIC scores for the model with assembly bias and the model without assembly bias. We note that the model without assembly bias is still preferable by the galaxy clustering observations. That is, although some improvements to fitting the clustering data can be achieved as a result of using a more complicated occupation model, these improvements are not significant enough to justify the use of a more complicated model that takes assembly bias into account.

We note that both AIC and BIC scores improve in the luminosity-threshold samples for which, we have tighter constraints over the central assembly bias parameter. In particular, the model with assembly bias deliver *only slightly* lower AIC scores for the $M_{\rm r} < -18, -20$

samples. This supports our intuition that AIC and BIC penalize unconstrained parameters. Also note that, the difference between both BIC and AIC scores are marginal.

4.6.3 choice of simulation

Given the SDSS clustering measurements described in Section 4.4, We repeat the inference of the assembly bias parameters $\mathcal{A}_{\rm sat}$ and $\mathcal{A}_{\rm cen}$ with the `BolshoiP` simulation (**?**). This N-body simulation is carried out with similar setting as the `Bolshoi` simulation with the exception that in the `BolshoiP` simulation, Planck cosmology is adapted and the mass per simulation particle is $1.49 \times 10^8\ h^{-1} M_\odot$.

The summary of constraints are shown in Figure 4.6. In Figure 4.6, the constraints from the `SMDP` and the `BolshoiP` simulations are shown with circles and crosses respectively. Additionally, the upper and lower bounds on the inferred parameters reported by **?** are shown in shaded blue regions. In the case of central assembly bias, all three constraints are consistent. For the luminosity-thresholds samples $M_{\rm r} < -20.5, -20, -19.5$, where the central assembly bias parameters are strongly positive, the constraints obtained from the `SMDP` simulation are tighter.

In the case of the satellite assembly bias parameters however, constraints from the `BolshoiP` simulation for the luminosity threshold samples $M_{\rm r} < -20.5, -19$ favor more positive values of the parameter, while our constraints from the `SMDP` simulation for these two luminosity thresholds favor zero satellite assembly bias. As it is shown in the lower panel of Figure 4.6, our constraints from the `BolshoiP` simulations for the $M_{\rm r} < -21, -20.5, -20, -19.5, -19$ samples are consistent with the lower and upper bounds (shown with the shaded blue region) reported by **?** that uses the same simulation but different w_p measurements (**?**). Therefore, there is some discrepancy between our $\mathcal{A}_{\rm sat}$ constraints using the `SMDP` simulation and the `BolshoiP` simulations.

For the $M_{\rm r} < -19, -20.5$ samples, the marginalized posterior PDFs over $\mathcal{A}_{\rm sat}$ from the two simulations are shown in Figure 4.7. Note that $\mathcal{A}_{\rm sat}$ is poorly constrained in both simulations and for both luminosity-threshold samples. For the $M_{\rm r} < -19$ sample, considering how poorly constrained the parameters are, the discrepancy between the constraints is not very stark. Note that the tension is still at a one-sigma level. The discrepancy however, becomes more pronounced in the $M_{\rm r} < -20.5$ sample.

In terms of the effect of assembly bias on galaxy clustering, note that mocks created using the inferred parameters with the `SMDP` simulation show the same behavior as we observe in the abundance matching catalogs presented in **?** and **?**. That is, the difference in w_p between the mock catalogs and the randomized catalogs is mostly on large scales where the clustering is governed by the central galaxies. That is, the impact of assembly bias on the satellite occupation is negligible and only the central occupation is affected.

4.7 Summary and Conclusion

In this investigation, we provide constraints on the concentration-dependence of halo occupation for a wide range of galaxy luminosities in the SDSS data. In particular, the modeling is done in the context of the decorated HOD model **?**, and the data used in this investigation is the projected 2PCF measurements published by **?**. We make use of `SMDP` high resolution N-body simulation that enables us to reliably perform inference for the faintest galaxy samples in the SDSS DR7 catalog that live in low mass halos, and the brightest galaxy samples that occupy large comoving volumes.

Our findings suggest that the satellite assembly bias remains consistent with zero. However, our constraints on the central assembly bias parameter exhibit a trend with the luminosity limits of the galaxy samples. For the brightest samples, central assembly bias is

consistent with zero, which is in agreement with this picture that the halo assembly bias becomes negligible for the most massive halos.

For the $M_{\rm r} < -20.5, -20, -19.5$ samples, at a fixed halo mass, we find positive correlation between the central population and halo concentration at fixed halo mass. For $M_{\rm r} < -19, -18.5$ we find no correlation, and for the faintest sample, we find negative correlation. Given the large scale halo assembly bias and the scale-dependent feature of assembly bias in the quasi-linear scales, our constraints on the more flexible HOD model lead to improvement in modeling the galaxy clustering. However, we do not find these improvements to be sufficient to lower the information criteria scores associated with the more complex model. The exceptions are the $M_{\rm r} < -20, -18$ luminosity-threshold samples for which we find the strongest constraints on the central assembly bias. For these two samples, the HOD model with assembly bias yields lower BIC score than the model without assembly bias.

We compare the impact of assembly bias on galaxy clustering between the catalogs constructed from our results and the abundance matching catalogs presented in **??**. We demonstrate that the effect of assembly bias on galaxy clustering predicted from our results is similar to (but more moderate than) the effects seen in the abundance matching catalogs of **?**. That is, assembly bias mostly affects the large scales and the quasi-linear clustering and the small scale clustering remains unaltered. In addition, the effect of assembly bias on galaxy clustering vanishes in the brightest galaxy samples.

Moreover, we repeat our inference using the `BolshoiP` simulation. We find that our findings based on the `BolshoiP` simulation are consistent with constraints reported by **?** (in the $M_{\rm r} < -21, -20.5, -20, -19.5, -19$ luminosity-threshold samples) that predicts positive satellite assembly bias (correlation between the expected number of satellites and $V_{\rm max}$ at fixed host halo mass) for the $M_{\rm r} < -19, -20.5$ samples. However, we note that the results based on the `SMDP` simulation are more consistent with the picture provided by the previous

models based on the abundance matching technique (e.g. **??**). That is, only the large-scale clustering, governed by the centrals, is affected by assembly bias.

Acknowledgments

We are grateful to David W. Hogg, Alex I. Malz, Andrew Hearin, Andrew Zentner, Chia-Hsun Chuang, Michael R. Blanton, and Kilian Walsh for discussions related to this work. This work was supported by the NSF grant AST-1517237. All the computations in this work were carried out in the New York University High Performance Computing Mercer facility. We thank Shenglong Wang, the administrator of the NYU HPC center for his continuous and consistent support throughout the completion of this study.

The CosmoSim database used in this paper is a service by the Leibniz-Institute for Astrophysics Potsdam (AIP). The MultiDark database was developed in cooperation with the Spanish MultiDark Consolider Project CSD2009-00064. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) and the Partnership for Advanced Supercomputing in Europe (PRACE, www.prace-ri.eu) for funding the MultiDark simulation project by providing computing time on the GCS Supercomputer SuperMUC at Leibniz Supercomputing Centre (LRZ, www.lrz.de).

All of the code written for this project is available in an open-source code repository at `https://github.com/mjvakili/gambly`. The SDSS clustering measurements and the covariance matrices used in this work are available at `http://sdss4.shao.ac.cn/guoh/files/wpxi_measurements_Guo15.tar.gz`. Description of the `SMDP` and the `BolshoiP` halo catalogs used in this investigation can be found at `https://www.cosmosim.org/cms/simulations`. The RockStar halo catalogs of the `SMDP` and the `BolshoiP` simulations are publicly available at `http://yun.ucsc.edu/sims/SMDPL/hlists/index.html` and `http:`

//yun.ucsc.edu/sims/Bolshoi_Planck/hlists/index.html respectively. We thank Peter Behroozi for making the halo catalogs publicly available. In this work we have made use of the publicly available codes: corner (**?**), emcee (**?**), halotools (**?**), Corrfunc (**?**), and changtools (https://github.com/changhoonhahn/ChangTools). The abundance matching mock catalogs used in this study are available at (http://logrus.uchicago.edu/~aphearin/).

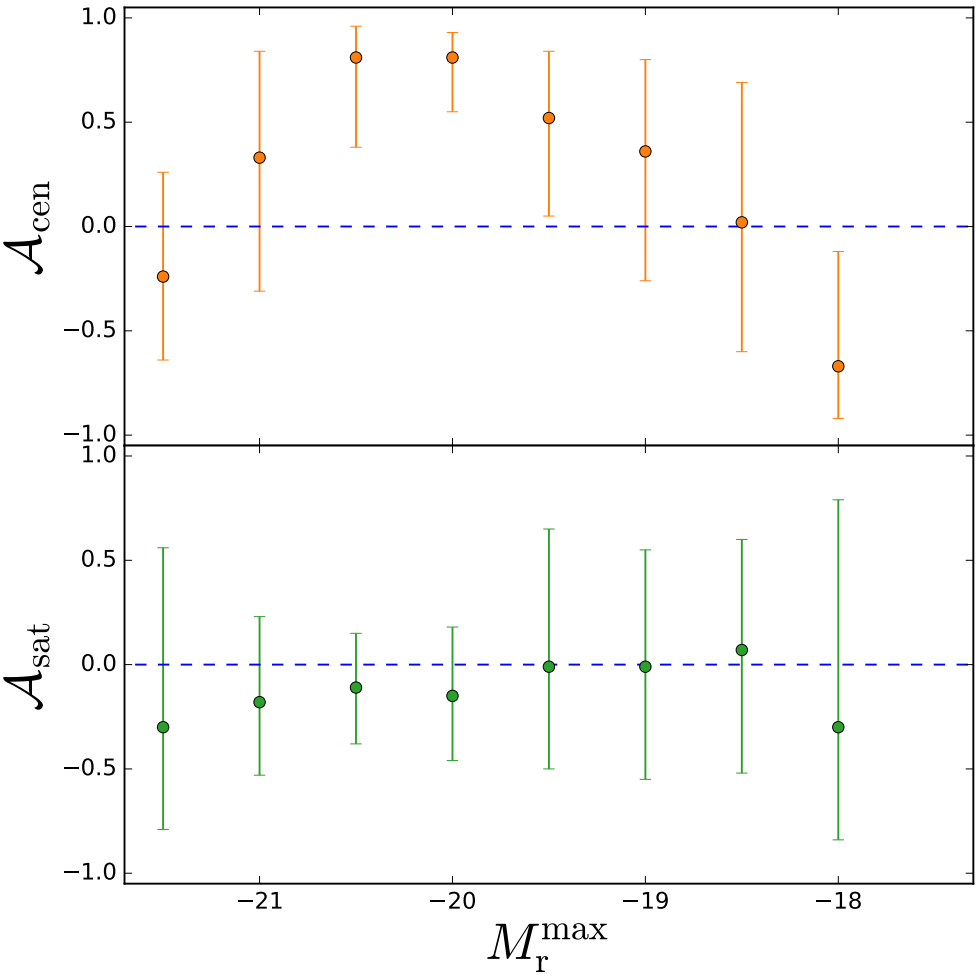

Figure 4.1: Constraints on the central assembly bias $\mathcal{A}_{\rm cen}$ (Top panel) and the satellite assembly bias $\mathcal{A}_{\rm sat}$ (Bottom panel) parameters. The $\mathcal{A}_{\rm cen}$ constraints for the $M_{\rm r} < -20.5, -20, -19.5$ samples favor positive values of $\mathcal{A}_{\rm cen}$ with the tightest constraint coming from the $M_{\rm r} < -20$ sample. The $\mathcal{A}_{\rm cen}$ constraints for the $M_{\rm r} < -18$ sample favor negative values of $\mathcal{A}_{\rm cen}$. All the $\mathcal{A}_{\rm sat}$ constraints are consistent with no satellite assembly bias.

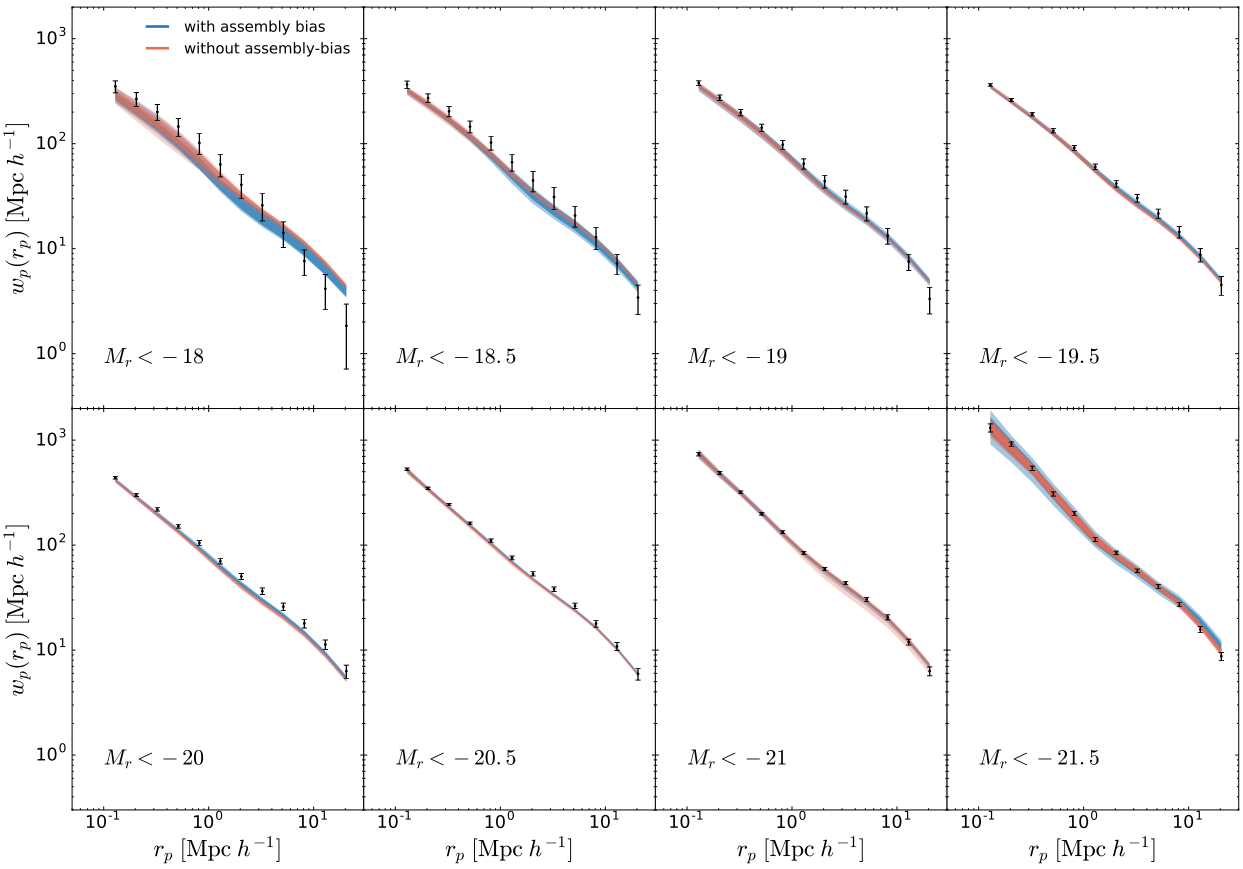

Figure 4.2: Comparison between the posterior predictions of $w_p(r_p)$ and the SDSS $w_p(r_p$ measurements. Predictions from the standard HOD model (HOD model with assembly bias) are shown in red (blue). The Dark and light shaded regions mark the 68% and the 95% confidence intervals. The errorbars are from the diagonal elements of the covariance matrix.

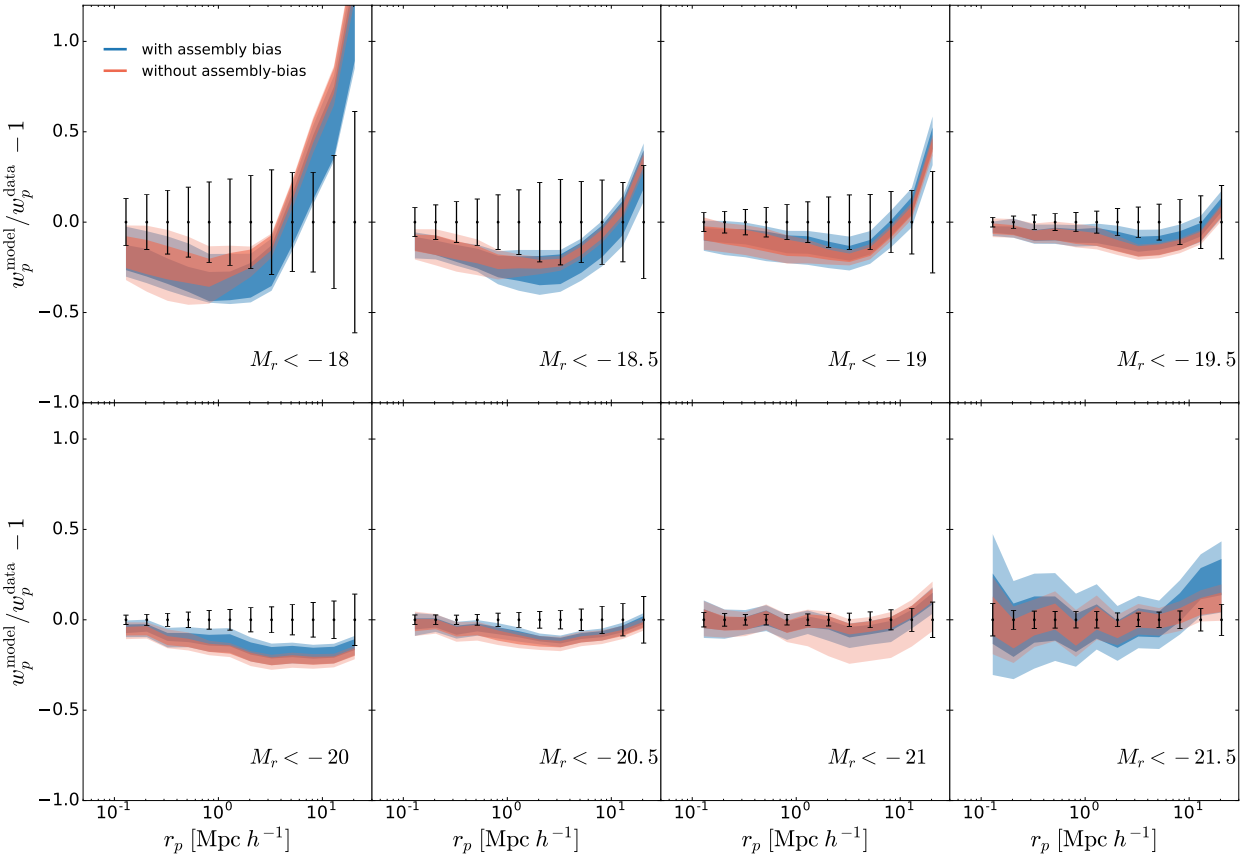

Figure 4.3: Same as Figure 4.2, but showing the fractional difference between the posterior predictions and the observed projected 2PCF for all the luminosity threshold samples. In all luminosity threshold samples, predictions of the two models for small scale clustering are consistent. In the samples that favor more positive values of the central assembly bias parameter ($M_{\rm r} < -19.5, -19, -20, -20.5$), modeling of the intermediate and large scale clustering is slightly improved. The large scale clustering modeling of the $M_{\rm r} < -18$ sample is also improved because of negative constraints on $\mathcal{A}_{\rm cen}$ which is equivalent to allocation of more central galaxies in low concentration halos at fixed halo mass.

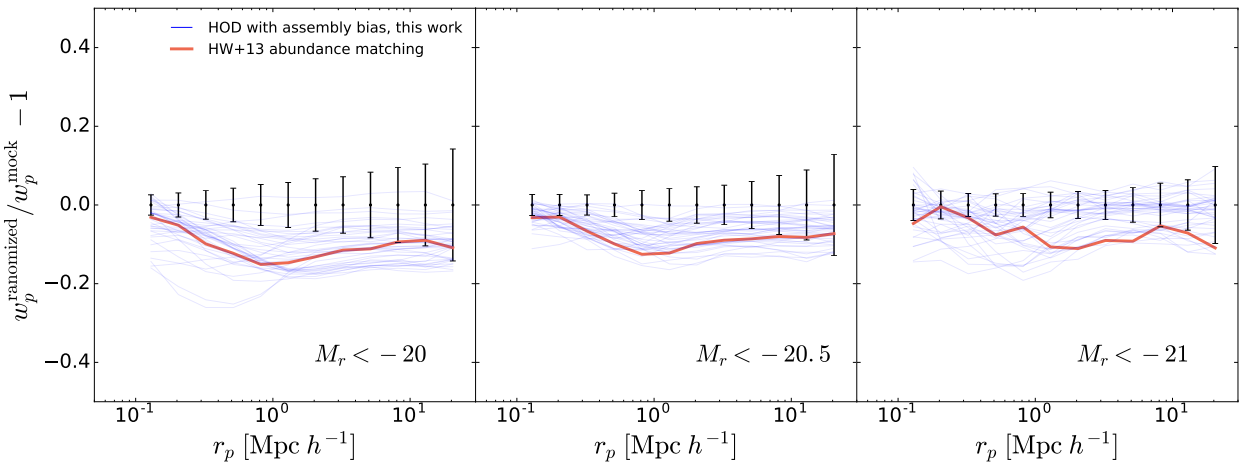

Figure 4.4: Demonstration of the relative difference in w_p between randomized and non-randomized catalogs for different luminosity threshold samples: $M_{\rm r} < -20, -20.5, -21$. The errorbars are from the diagonal elements of the covariance matrix. The blue lines correspond to the random draws from the posterior probability (summarized in Table 4.6.1) over the parameters of the HOD model with assembly bias. The red line corresponds to the subhalo abundance matching catalog (**??**). Our constraints favor *more moderate* levels of the impact of assembly bias on galaxy clustering than the levels seen in the abundance matching mock catalogs. Within both models, the small scale clustering remains unaltered after randomizing the catalogs, signaling the lack of correlation between the satellite occupation and the halo concentration at a fixed mass in the two models.

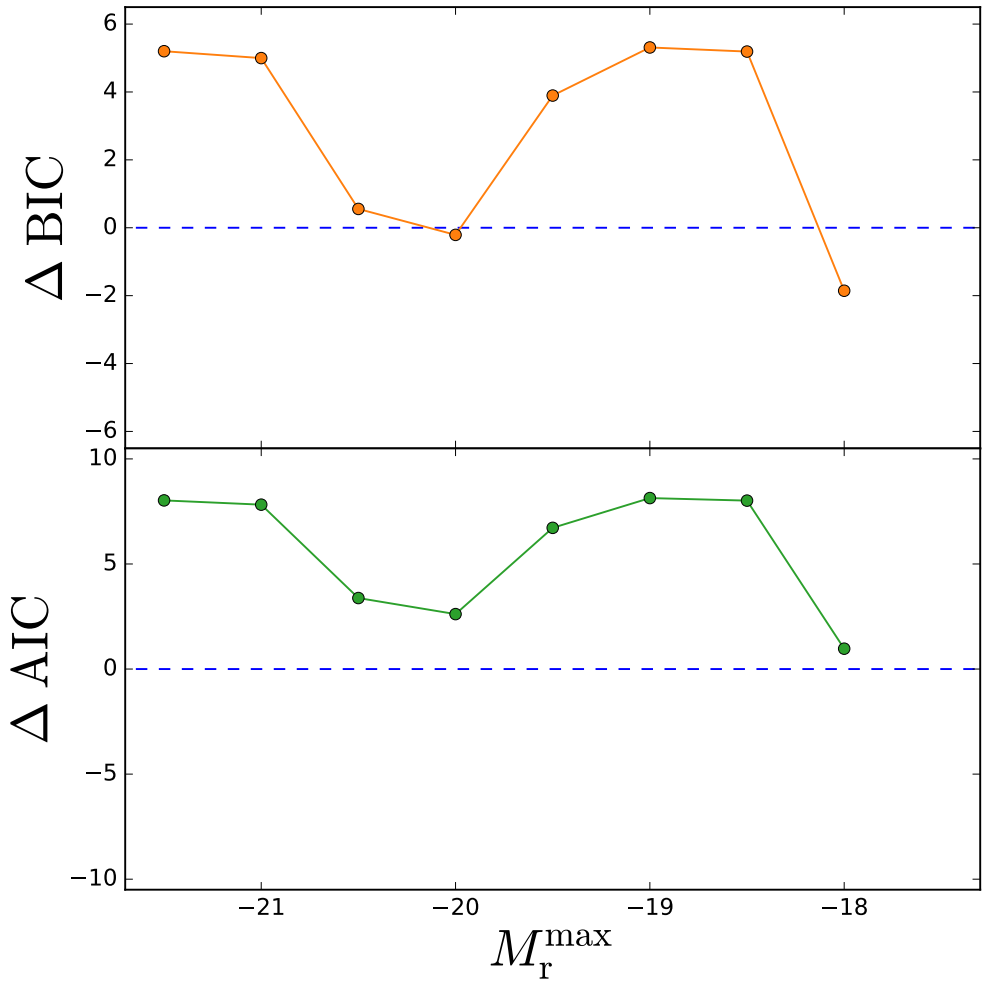

Figure 4.5: Difference in the information criteria between the HOD model with assembly bias and the model without assembly bias. **Top**: ΔBIC = BIC(with assembly bias) - BIC(without assembly bias). **Bottom**: ΔAIC = AIC(with assembly bias) - AIC(without assembly bias). According to BIC (AIC), the more complex model with assembly bias is favored once ΔBIC$<$ 0 (ΔAIC$<$ 0). Both ΔBIC and ΔAIC are lower for the samples with tighter constraints over the central assembly bias parameter $\mathcal{A}_{\rm cen}$, with ΔBIC being (marginally) negative only for $M_{\rm r} < -20, -18$ samples that yield strongest constraints on $\mathcal{A}_{\rm cen}$.

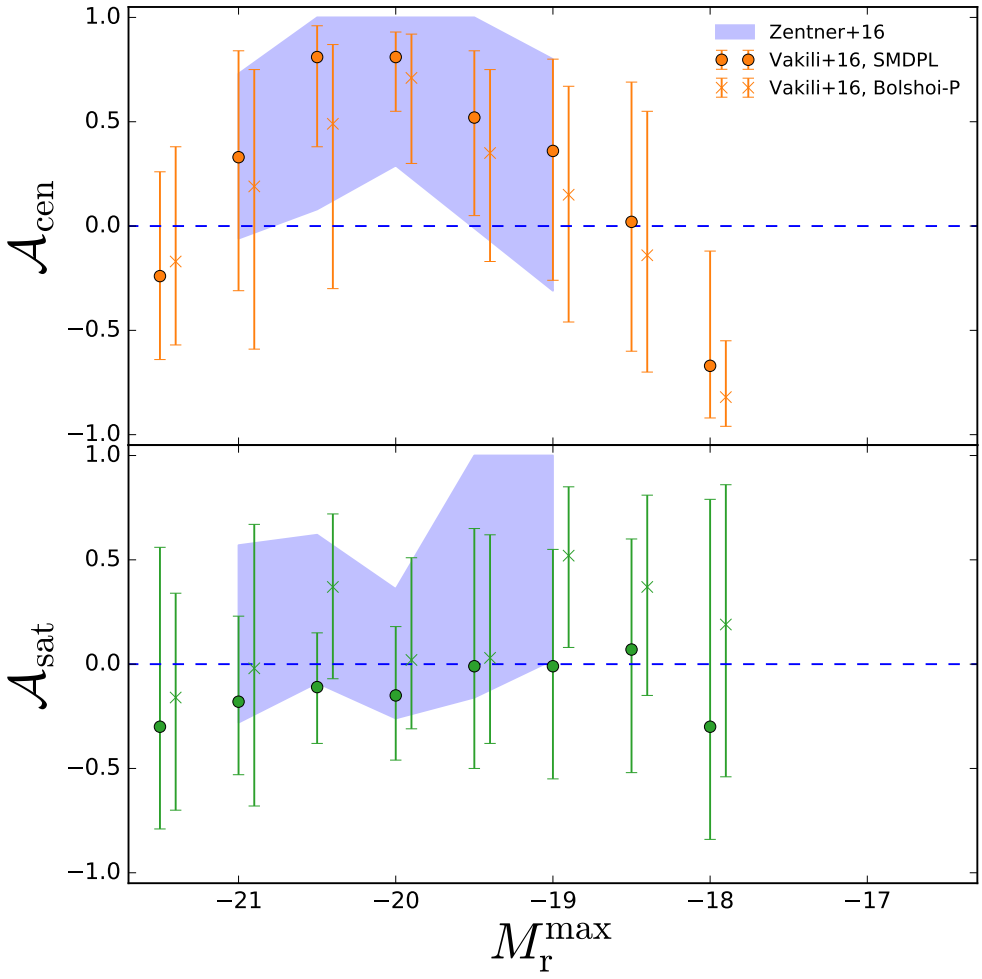

Figure 4.6: Comparison between the constraints on the assembly bias parameters $\mathcal{A}_{\rm cen}$ (shown in the top panel) and $\mathcal{A}_{\rm sat}$ (shown in the bottom panel) for different simulations: `SMDP` (shown with circle), and `BolshoiP` (shown with cross). The errorbars mark the 68% uncertainty over the parameters. Shaded blue regions show the upper and lower bounds reported by **?** that uses the `BolshoiP` and clustering measurements of **?**. For the confidence intervals corresponding to the shaded blue regions, we refer the readers to Table 2 of **?**. The central assembly bias constraints found from the two simulations are consistent, with the constraints for from the `SMDP` simulation being tighter for the most luminous samples. The constraints on $\mathcal{A}_{\rm sat}$ from the two simulations are largely in agreement with the exception of $M_{\rm r} < -19,\ -20.5$ samples that favor more positive values of $\mathcal{A}_{\rm sat}$ when the `BolshoiP` simulation is used.

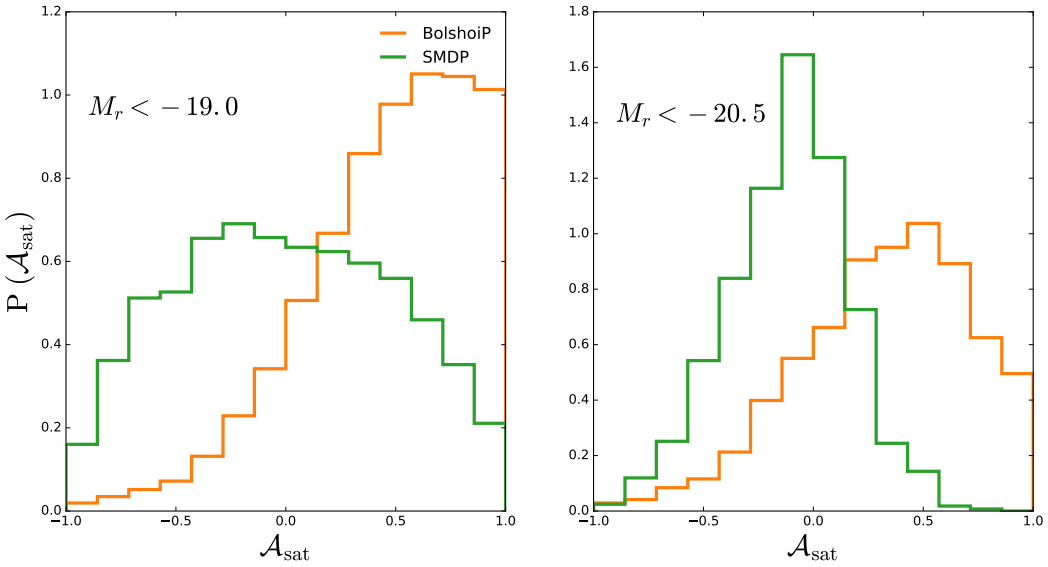

Figure 4.7: Constraints over the satellite assembly bias parameters from luminosity-threshold samples $M_{\rm r} < -19$, -20.5, for two different simulations: `BolshoiP` (yellow), and `SMDPL` (green). The $\mathcal{A}_{\rm sat}$ constraints found using the `BolshoiP` simulation favor more positive values of $\mathcal{A}_{\rm sat}$, while the constraints found using the `SMDP` simulation favor zero satellite assembly bias.

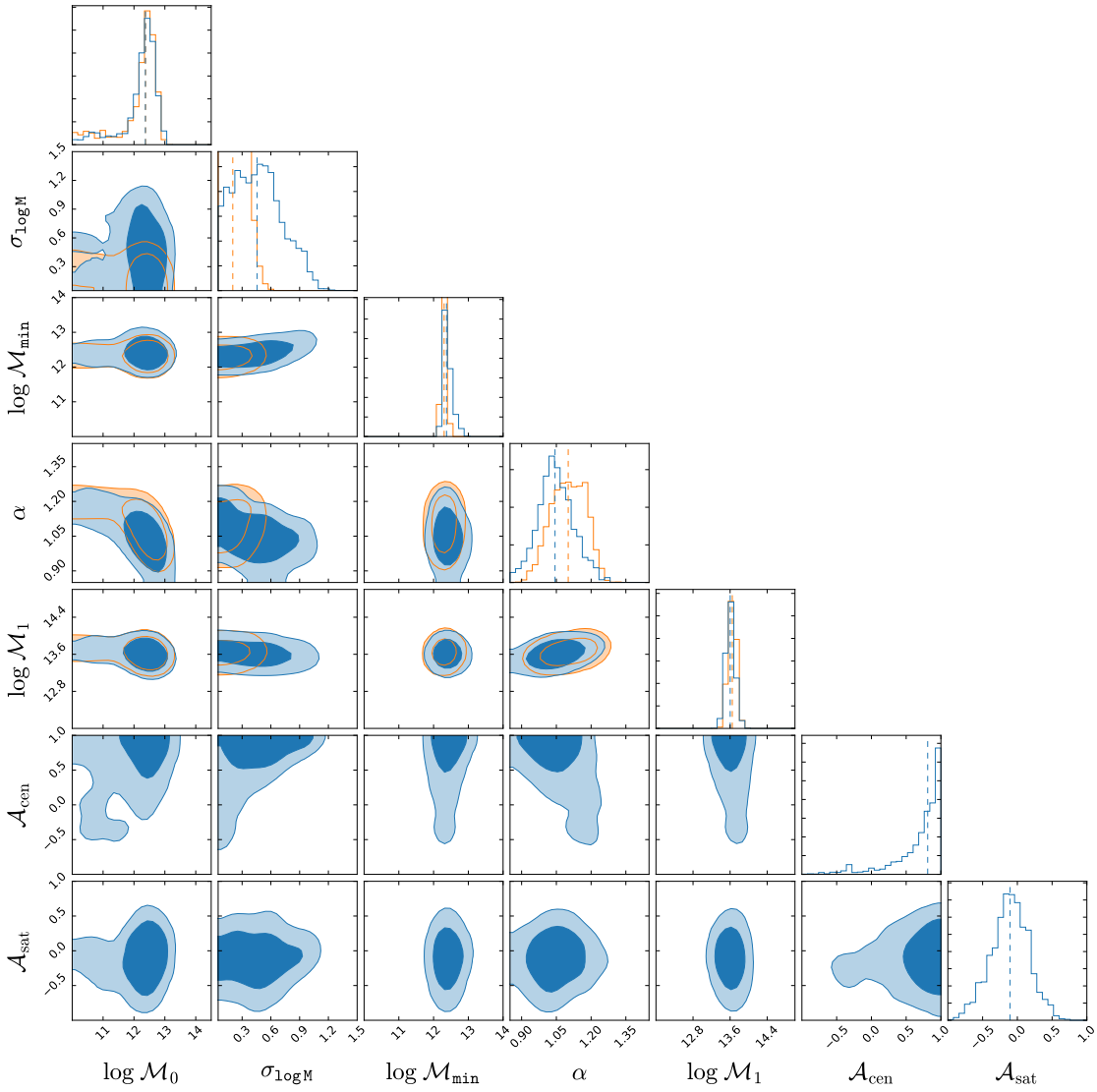

Figure 4.8: An example of posterior probability distribution over the parameters of the standard HOD model with no assembly bias (shown with yellow), and the HOD model with assembly bias (shown in blue). These constraints are obtained from the clustering measurements of the $M_{\rm r} < -20.5$ luminosity threshold sample. The dark (light) blue shaded regions show the 68% (95 %) confidence intervals. The constraints on $\mathcal{A}_{\rm cen}$ and $\mathcal{A}_{\rm sat}$ show positive correlation between the central occupation and the halo concentration at fixed halo mass, and lack of correlation between the satellite occupation and halo concentration at fixed halo mass.

Chapter 5

Accurate galaxy-halo mocks with automatic bias estimation and particle-mesh gravity solvers

This Chapter is joint work with Francisco-Shu Kitaura (IAC), Yu Feng (Berkeley), Gustavo Yepes (UAM), Cheng Zhao (Tsinghua), Chia-Hsun Chuang (Leibniz), ChangHoon Hahn (NYU) submitted to the *Monthly Royal Astronomical Society Notice* as **?**.

5.1 abstract

Reliable extraction of cosmological information from clustering measurements of galaxy surveys requires estimation of the error covariance matrices of observables. The accuracy of covariance matrices is limited by our ability to generate sufficiently large number of independent mock catalogs that can describe the physics of galaxy clustering across a wide range of scales. Furthermore, galaxy mock catalogs are required to study systematics in galaxy

surveys and to test analysis tools. In this investigation, we present a fast and accurate approach for generation of mock catalogs for the upcoming galaxy surveys. Our method relies on low-resolution approximate gravity solvers to simulate the large scale dark matter field, which we then populate with halos according to a flexible nonlinear and stochastic bias model. In particular, we extend the PATCHY code with an efficient particle mesh algorithm to simulate the dark matter field (the FASTPM code), and with an efficient and robust MCMC method relying on the EMCEE code for constraining the parameters of the bias model. Using the halos in the BigMultiDark high-resolution N-body simulation as a reference catalog, we demonstrate that our technique can model the bivariate probability distribution function, power spectrum, and bispectrum of halos in the reference catalog. Specifically, we show that the new ingredients permit us to reach percentage accuracy in the power spectrum up to $k \sim 0.4\ h\,\mathrm{Mpc}^{-1}$ (within 5% up to $k \sim 0.6\ h\,\mathrm{Mpc}^{-1}$) with accurate bispectra improving previous results based on Lagrangian perturbation theory.

5.2 Introduction

The current and the next generation of galaxy surveys such as EBOSS[1] (Extended Baryon Oscillation Spectroscopic Survey, **?**), DESI[2] (Dark Energy Spectroscopic Instrument, **?**), EUCLID[3] (**?**), LSST[4] (**?**), and WFIRST[5] (**?**) are expected to achieve unprecedented constraints on the cosmological parameters, growth of structure, expansion history of the universe, and modified theories of gravity. Accurate cosmological inferences with these surveys requires accurate computation of the likelihood function of the observed data given a

[1] http://www.sdss.org/surveys/eboss/
[2] http://desi.lbl.gov/
[3] http://www.euclid-ec.org/
[4] http://www.lsst.org/
[5] https://www.nasa.gov/wfirst

cosmological model. This goal can be achieved provided that the uncertainties, in the form of error covariance matrices in the likelihood functions, are reliably estimated. Therefore, covariance matrices are essential ingredients in extraction of cosmological information from the data.

The most commonly used technique in estimation of the covariance matrix for galaxy clustering observables requires generation of a large number of simulated galaxy mock catalogs. These mock catalogs need to reproduce the cosmic volume probed by the galaxy surveys. They also need to describe the clustering observables with high accuracy in a wide range of scales. It has been demonstrated that both the precision and the accuracy of constraints on the cosmological parameters, regardless of the details of a given galaxy survey, depend on the number of realizations of the survey (**??**). The requirement on the number of independent realizations of the survey becomes more stringent as the number of data points in a given analysis grows (**?**). The most pressing challenges ahead of simulating a large number of catalogs are: simulation of large volumes for sampling the Baryonic Acoustic feature in the galaxy clustering, accurate description of the clustering signal at small scales, accurate clustering not only at the level of two-point statistics but also at the level of higher order statistics. and resolving low mass halos that host fainter galaxy samples.

High-resolution N-body simulations are ideal venues for reproducing the dark matter clustering accurately. But production of a large number of density field realizations with N-body simulations is not computationally feasible. In order to alleviate the computational cost of N-body simulations, several methods based on approximate gravity solvers have been introduced. Methods based on higher order Lagrangian perturbation theory (**??????**), Zeldovich approximation (**?**), and approximate N-body simulations (**???????**) have been demonstrated to be promising for fast generation of dark matter density field. Sampling the structures such as galaxies and halos from the dark matter density field requires an additional

step. Identification of virialized regions of matter overdensity is either done through a biasing scheme (**??**) or is done through application of friends-of-friend algorithm (**???**). Methods that employ a biasing scheme need to be calibrated such that they are statistically consistent with accurate N-body simulations or observations.

The PATCHY method (**??**) produces mock catalogs by first generating dark matter field with Lagrangian Perturbation Theory modified with spherical collapse model on small scales ($r \leq 2\ h^{-1}\,\mathrm{Mpc}$) and then sampling galaxies (halos) from the density field using nonlinear stochastic biasing introduced in **?**. This method has been shown to reproduce the two-point clustering down to $k \sim 0.3\ h\,\mathrm{Mpc}^{-1}$ and the counts-in-cell of the massive halos in an accurate N-body simulation. **?** demonstrate that the mock catalogs generated using this technique are capable of accurately describing the halo bispectrum in the reference N-body simulations. Furthermore, **?** used this method for massive production of mock catalogs for the cosmological analysis of the completed SDSS III Baryon Oscillation Spectroscopic Survey DR12 galaxy sample.

Alternatively, computation of error covariance matrices can be delivered with analytical models (**??????**). These methods are promising, though still need further investigation especially including sysytematic effects, such as the survey geometry. They will potentially permit us to use a smaller number of mock catalogs to obtain accurate covariance matrices.

In recent years, development of the shrinkage methods (**?????**) have been proved promising for alleviating the requirement on the number of mocks. In principle, one could use a combination of the shrinkage methods and a smaller number of mock catalogs to reach the same level of accuracy needed for large scale structure inferences.

Moreover, production of mocks will be a useful tool for investigation of possible sources of systematic errors as well as verification of covariance matrices derived from analytical methods.

In this investigation, we introduce an MCMC method for calibration of the bias model of the PATCHY code. This method constrains the bias parameters by the halo power spectrum and the halo counts-in-cells (hereafter halo PDF) of a reference halo catalog constructed from an accurate N-body simulation.

Furthermore, we replace the dark matter gravity solver of the code with the fast particle-mesh approximate N-body solver implemented in the FASTPM code (**?**). The advantage of the FASTPM algorithm over other methods based on particle-mesh is its low memory requirements as well as accurate large scale growth. In addition, the dark matter density field produced by the FASTPM code yields better nonlinear clustering than that of the perturbation theory.

As a proof of concept, we make use of the halos in the BigMultiDark Planck high-resolution N-body simulation (**?**). This catalog has been extensively used for validation, comparison and production of galaxy mock catalogs (**????**). In addition, we will make a statistical comparison between our PATCHY mocks and the reference catalog. We present the number density, halo PDF, and halo two-point statistics. We also present our results in terms of the three-point statistics since it is rising as a major complementary approach in various large-scale structure analyses (**???????**).

The remainder of this paper is structured as follows: In section §5.3, we present our method for generating and calibrating mock catalogs. This includes description of the structure formation model, nonlinear stochastic bias model of the PATCHY code, and our MCMC method for constraining the bias parameters. We illustrate the performances using a reference halo catalog constructed from an accurate N-body simulation in section §5.4, and we discuss the main results and present our conclusions in section §5.5.

5.3 Methodology

Our method consists of producing the large scale dark matter field on a mesh and then populating it with halos (or galaxies) with a given bias model. The parameters of that bias model are constrained with a reference catalog in an automatic statistical way. Our approach is agnostic about the method used for identification of halos in the reference catalog. The PATCHY code permits us to sample galaxies directly from the density field. For instance **?** samples mock galaxy catalogs based on an accurate reference mock galaxy catalog (**?**). Let us first describe in §5.3.1 the new implementation of the structure formation in PATCHY, followed in §5.3.2 by the bias model, and finally in §5.3.3 our novel MCMC sampling procedure to obtain the bias parameters.

5.3.1 Structure formation model

Originally, the PATCHY code used Augmented Lagrangian Perturbation Theory (ALPT, **?**) as a structure formation model. In this model the second order Lagrangian perturbation theory is modified by employing a spherical collapse model on small comoving scales ($r \leq 2\ h^{-1}\,\mathrm{Mpc}$). Any LPT based approximation will lack the one halo term in the clustering. This can be partially compensated within the bias model, however, at the price of obtaining a less accurate description of the biasing relation. Therefore we introduce in this work within the PATCHY code the fast particle mesh code FastPM (**?**). In FastPM, the kick and drift steps of the PM codes are modified such that the linear growth of structure is exact. **?** demonstrate that the memory requirements of this algorithm are much lower than those of the COmoving Lagrangian Acceleration N-body solver (COLA, **?**).

Moreover, **?** shows that running the code with relatively few time steps, and applying a friends-of-friend (hereafter fof) halo finder (**?**) to the density field, one can accurately recover

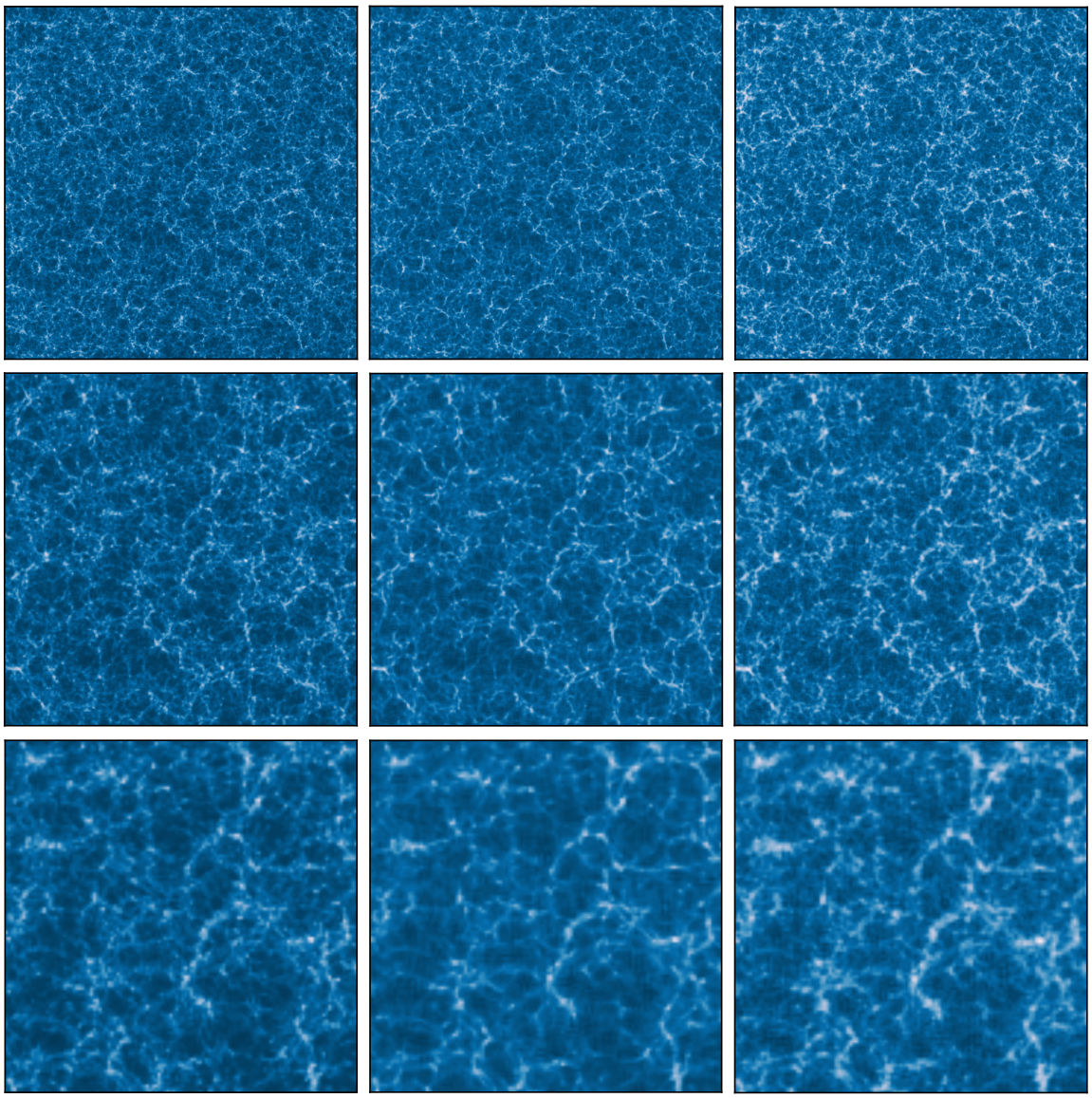

Figure 5.1: Dark matter overdensity $\delta = \rho_{\rm m}/\rho - 1$ slices of 20 $h^{-1}\,$Mpc from the high-resolution BigMultiDark simulation (left panels) , the low-resolution FastPM simulation (central panels) and from the ALPT simulation (right panels), taking a subvolume of $(1250\ h^{-1}\,{\rm Mpc})^3$ (top panels), $(625\ h^{-1}\,{\rm Mpc})^3$ (middle panels), and $(312.5\ h^{-1}\,{\rm Mpc})^3$ (bottom panels). The structures in the high-resolution N-body simulation and the low-resolution FastPM simulation look very similar inspite of having very different resolutions (3840^3 vs 960^3 particles). The low-resolution ALPT simulation looks more diffuse.

the redshift space power spectrum of the fof halos of TreePM accurate N-body solver (**?**) down to $k \sim 0.5\ h\,\mathrm{Mpc}^{-1}$. The linking length of 0.2 was chosen to be consistent with other works in the literature (**?**). In this work, we run the FastPM code with 10 time steps.

In this work we will use as a reference the high-resolution N-body BigMultiDark simulation described in more detail in section §5.4.1. A comparison of the dark matter density fields obtained with the different methods is shown in Fig. 5.1. While the structures in the high-resolution N-body simulation and the low-resolution FastPM simulation look very similar inspite of having very different resolutions (3840^3 vs 960^3 particles), the low-resolution ALPT simulation looks more diffuse due to the exaggerated shell crossing inherent to LPT based methods. We will study the impact of this inaccuracy in more detail in section §5.4.3.

5.3.2 Sampling halos from the density field

In this section, we describe the statistical bias model of the PATCHY code. This model generates halos/galaxies from a given dark matter density field and consists of: deterministic bias, stochastic bias, and an additional step for applying redshift space distortions (RSDs) to the catalogs. We describe the bias steps below and leave RSDs for a later work.

5.3.2.1 Deterministic bias

The expected number of halos $\langle\rho_\mathrm{h}\rangle$ in a given volume element $\mathrm{d}V$ (cosmic cell) can be described in general by a deterministic bias relation $B(\rho_\mathrm{h}|\rho_\mathrm{m})$:

$$\langle\rho_\mathrm{h}\rangle_{\mathrm{d}V} = f_\mathrm{h}\, B(\rho_\mathrm{h}|\rho_\mathrm{m})\,, \tag{5.1}$$

where ρ_m is the matter density field. The prefactor f_h is an overall normalization factor which can be determined by requiring the halo density field to have the number density of

the reference sample $n_{\rm h}$, i.e., $n_{\rm h} = \langle\langle\rho_{\rm h}\rangle_{{\rm d}V}\rangle_V$. Formally, this can be written as

$$f_{\rm h} = \frac{n_{\rm h}}{\langle B(\rho_{\rm h}|\rho_{\rm m})\rangle_V}\,, \tag{5.2}$$

where $\langle\ .\ \rangle_V$ is an ensemble volume average. In particular, we will adopt the following compact deterministic bias model:

$$\begin{aligned} B(\rho_{\rm h}|\rho_{\rm m}) \;&=\; \underbrace{\rho_{\rm m}^{\alpha}}_{\text{nonlinear bias}} \\ &\times\; \underbrace{\theta\big(\rho_{\rm m}-\rho_{\rm th}\big)}_{\text{threshold bias}} \times \underbrace{\exp\big(-(\rho_{\rm m}/\rho_\epsilon)^\epsilon\big)}_{\text{exponential cutoff}}\,, \end{aligned} \tag{5.3}$$

where $\rho_{\rm th}$ is the density threshold which suppresses halo formation in under-dense regions, and α is a nonlinear bias parameter. The threshold bias (**????**) is modeled by a step function $\theta\big(\rho_{\rm m}-\rho_{\rm th}\big)$ (**?**) and an exponential cutoff $\exp\big(-(\rho/\rho_\epsilon)^\epsilon\big)$ (**?**). Therefore, for this particular bias model we have a normalisation of

$$f_{\rm h} = \frac{n_{\rm h}}{\langle\theta(\rho_{\rm m}-\rho_{\rm th})\ \rho_{\rm m}^{\alpha}\ \exp\big(-(\rho_{\rm m}/\rho_\epsilon)^\epsilon\big)\rangle_V}\,. \tag{5.4}$$

The advantage of this kind of bias model is that it is flexible and it is able to incorporate additional terms and each of the terms have a physical interpretation. The power law bias stands for one of the simplest possible nonlinear bias models: a linear Lagrangian bias in a comoving framework, which can be derived from the lognormal approximation (see **?**), and it resumes in one single bias parameter an infinite Taylor expansion of the dark matter density field (**???**).

The threshold bias and the exponential cut-off describe the fact that halos (or galaxies) can only reside in regions which contain a minimum mass. They also represent the loss of

information with respect to the full cosmic density field from selecting only gravitationally collapsed objects.

5.3.2.2 Stochastic bias

The number of halos in each cell is drawn from a Negative Binomial (NB) distribution which can be characterized by the expected number of halos in the cell $\lambda_{\rm h} = \langle \rho_{\rm h} \rangle_{{\rm d}V} \times {\rm d}V$, and a parameter β which quantifies the stochasticity (deviation of the distribution from Poissonity) in the halo distribution. According to this model, the probability of having $N_{\rm h}$ objects in a volume element is given by

$$P(N_{\rm h}|\lambda_{\rm h},\beta) = \underbrace{\frac{\lambda_{\rm h}^{N_{\rm h}}}{N_{\rm h}!}\, e^{-\lambda_{\rm h}}}_{\text{Poisson distribution}} \times \underbrace{\frac{\Gamma(\beta+N_{\rm h})}{\Gamma(\beta)(\beta+\lambda_{\rm h})^{N_{\rm h}}} \times \frac{e^{\lambda_{\rm h}}}{(1+\lambda_{\rm h}/\beta)^{\beta}}}_{\text{Deviation from Poissonity}} . \tag{5.5}$$

For $\beta \to \infty$ we can show that the second raw in the above equation goes to one. Since $\Gamma(\beta) = \frac{\Gamma(\beta+1)}{\beta} = \frac{\Gamma(\beta+N_{\rm h})}{\beta(\beta+1)\cdots(\beta+N_{\rm h}-1)}$, the first factor can be written as $\frac{\Gamma(\beta+N_{\rm h})}{\Gamma(\beta)(\beta+\lambda_{\rm h})^{N_{\rm h}}} = \frac{\beta(\beta+1)\cdots(\beta+N_{\rm h}-1)}{(\beta+\lambda_{\rm h})^{N_{\rm h}}} = \frac{(1+1/\beta)\cdots(1+(N_{\rm h}-1)/\beta)}{(1+\lambda_{\rm h}/\beta)^{N_{\rm h}}}$. It is now straightforward to see that this goes to one for $\beta \to \infty$. The same happens for the second factor $\frac{e^{\lambda_{\rm h}}}{(1+\lambda_{\rm h}/\beta)^{\beta}} \to 1$, since $(1+\lambda_{\rm h}/\beta)^{\beta} \to e^{\lambda_{\rm h}}$ in that limit.

Given a dark matter density field $\rho_{\rm m}$, the halo density field can be constructed by drawing samples from the expected halo density field $\rho_{\rm h}$ with the Negative-Binomial (hereafter NB) distribution (Eq. 5.5). This is inspired by the fact that the excess probability of finding halos in high density regions generates over-dispersion (**??**). This over-dispersion is modeled by a NB distribution (**??**).

The stochastic bias stands for the shot noise from the transition of the continuous dark matter field to the discrete halo (or galaxy) distribution. As predicted by **?**, it produces a

dispersion larger than Poisson, as long as the two-point correlation function remains positive below the scale of the cell size. This is captured by the negative binomial PDF (Eq. 5.5).

5.3.3 Constraining the bias model

Production of approximate mock catalogs with PATCHY requires a reference catalog constructed from the observations or based on an accurate N-body simulation. We aim at constraining the parameters describing the deterministic bias $\{\delta_{\rm th}, \alpha, \rho_\epsilon, \epsilon\}$, and the parameter that governs the stochasticity of the halo population $\{\beta\}$.

The bias parameters are estimated such that the statistical summaries of the halos (galaxies) in the PATCHY mocks match the statistical summaries of the halos (galaxies) in the reference catalog. The set of statistical summaries of the catalog can in principle include number density, bivariate probability distribution function or number of counts-in-cells ρ (hereafter halo PDF), two-point statistics ξ_2, and higher-order statistics such as the three-point statistics ξ_3.

By construction, the PATCHY mocks reproduce the exact number density of objects in the reference catalog. This comes from the particular choice of normalization in the deterministic bias relation (see Eqs. 5.3,5.4). In this work, we follow **?** and constrain the bias parameters with the halo PDF and the two-point statistics ξ_2. These two quantities can be computed very fast and the skewness of the halo PDF determines the three point statistics. Given the bias parameters found fitting the PDF and the two-point statistics, we will demonstrate a comparison between the approximate mocks and the reference catalog in terms of the two- and three-point statistics.

We simultaneously fit the real-space power spectrum $P(k)$ and the PDF $\rho(n)$ of the PATCHY halo density field to $P(k)$ and $\rho(n)$ measured for the BigMuliDark halo catalog. Specifically, constraints on $\theta = \{\delta_{\rm th}, \alpha, \rho_\epsilon, \epsilon, \beta\}$ are found by sampling from the posterior

probability $p(\theta|\text{data}) \propto p(\text{ref}|\theta)p(\theta)$, where ref denotes the combination $\{P_{\rm ref}(k), \rho_{\rm ref}(n)\}$, and the likelihood $p(\text{ref}|\theta)$ is given by

$$\begin{aligned} -2\ln p(\text{ref}|\theta) &= \Sigma_k \frac{\big(P_{\rm ref}(k) - P_{\rm mock}(k)\big)}{\sigma_k^2} \\ &+ \Sigma_n \frac{\big(\rho_{\rm ref}(n) - \rho_{\rm mock}(n)\big)}{\sigma_n^2}. \end{aligned} \tag{5.6}$$

For the purpose of estimating the bias parameters, we find it sufficient to assume simple uncorrelated noise terms $\{\sigma_k, \sigma_n\}$ in the above likelihood (5.6). We assume σ_k^2 to be $4\pi^2 P_{\rm ref}^2(k)/(V_{\rm box}k^2\Delta k)$, and σ_n^2 to be N_n where N_n is the number of cells containing n number of halos (including parent halos and subhalos). Furthermore, we choose a flat prior for all parameters of the bias model with the following lower and upper bounds: $-1 < \delta_{\rm th} < 2$, $0 < \alpha < 1$, $0 < \beta < 1$, $0 < \rho_\epsilon < 1$, and $0 < \epsilon < 1$.

For sampling from the posterior probability, given the likelihood function (Eq. 5.6) and the prior, we use the affine-invariant ensemble MCMC sampler (**?**) and its implementation EMCEE (**?**). In particular, we run the EMCEE code with 10 walkers and we run the chains for at least 2000 iterations. We discard the first 500 chains as burn-in samples and use the remainder of the chains as production MCMC chains. Furthermore, we perform Gelman-Rubin convergence test (**?**) to ensure that the MCMC chains have reached convergence.

5.3.4 Comparison with other approximate methods

Pioneering fast halo/galaxy generating methods have relied on approximate gravity solvers based on Lagrangian perturbation theory (LPT) to compute the position and mass of the objects, such as PINOCCHIO (Zeldovich: **??**, 3LPT: **?**) and PTHALOS (2LPT: **??????**). This has the disadvantage of being affected by an inaccurate description of the small scale clustering, and, in particular, of missing the one halo term contribution. As a consequence,

the power spectra of such catalogs have systematic deviations towards high values of k, already deviating about 10% at $k \sim 0.2\ h\,\mathrm{Mpc}^{-1}$ (**?**).

While fast particle mesh solvers, such as COLA or FastPM, are much more precise than LPT based approaches, they are still computationally too expensive to be suitable for massive production, if one is trying to resolve all the necessary structures required to model next generation of galaxy surveys. Therefore three methods were recently proposed: PATCHY (**?**), QPM (**?**), and EZmocks (**?**), which do not try to resolve halos (nor galaxies) with the approximate gravity solvers, but just get a reliable large scale dark matter field, which can then be populated with some bias prescription. The gravity solver thus only needs to be accurate on a certain scale, then the halo/galaxy-dark matter connection is exploited to reach a high accuracy, as described above. These methods use both different gravity solvers and different bias models. While PATCHY originally relies on ALPT, QPM uses a quick particle mesh solver, and EZmocks uses the Zeldovich linear LPT. But more importantly the bias prescription follows very different philosophies. QPM uses a rank ordering scheme relating the halo mass to density peaks. However, a recent study demonstrated that the dependence of the halo mass to its environment is not trivial (see **?**). EZmocks, on the other hand, first modifies the initial power spectrum introducing a tilt to adjust the final two point statistics, correcting hereby the missing one halo term of the approximate gravity solver. Second it imposes the halo PDF, which was shown to determine the 3pt statistics (**?**). PATCHY on the other hand follows a more physical approach, relying on an effective analytical stochastic bias model. In this sense the statistics is not directly imposed as in EZmocks, but fitted through the bias parameters. In fact PATCHY was shown to be considerably more accurate than EZmocks when assigning halo masses (**?**), and than QPM when fitting the two and three point statistics of the luminous red galaxies (LRGs) in the Baryon Oscillation Spectroscopic Survey (BOSS) (**?**).

Moreover, the approach we follow in PATCHY tests the validity range of effective bias prescriptions commonly used in large scale structure analysis methods (see e.g. **?**).

Now for the first time we include a robust MCMC sampling scheme to determine the bias parameters, and have improved the gravity solver with FASTPM.

5.4 Demonstration on an accurate N-body based halo catalog

In this section we present the application of the above described method to a well studied case: the halo distribution required to describe the CMASS LRG sample of the BOSS survey (**??**). First, we briefly describe the reference catalog and then present a detailed statistical analysis of the results.

5.4.1 Reference catalog

For the reference simulation used in this work we rely on the Bound-Density-Maxima (BMD, **?**) halo catalogs in the $z = 0.5618$ snapshot of the BigMultiDark-Planck high resolution N-body simulation (**?**). This simulation was carried out using the L-Gadget2 code (**?**), following the Planck ΛCDM cosmological parameters $\Omega_{\rm m} = 0.307$, $\Omega_{\rm b} = 0.048$, $\Omega_\Lambda = 0.693$, $\sigma_8 = 0.823$, $n_{\rm s} = 0.96$, $h = 0.678$. The box size for this N-body simulation is 2500 $h^{-1}\,$Mpc, the number of simulation particles is 3840^3, the mass per simulation particle $m_{\rm p}$ is $2.359 \times 10^{10}\ h^{-1} M_\odot$, and the gravitational softening length ϵ is 30 h^{-1}kpc at high-z and 10 h^{-1}kpc at low-z.

A minimum mass cut of $0.5 \times 10^{13}\ h^{-1} M_\odot$ has been applied to the halo catalog so that it matches with the number density of the SDSS III-BOSS CMASS galaxy catalog (**??**). After applying the mass cut, the number density of the final catalog is 3.5 $\times 10^{-4}$ $(\,h\,{\rm Mpc}^{-1})^3$.

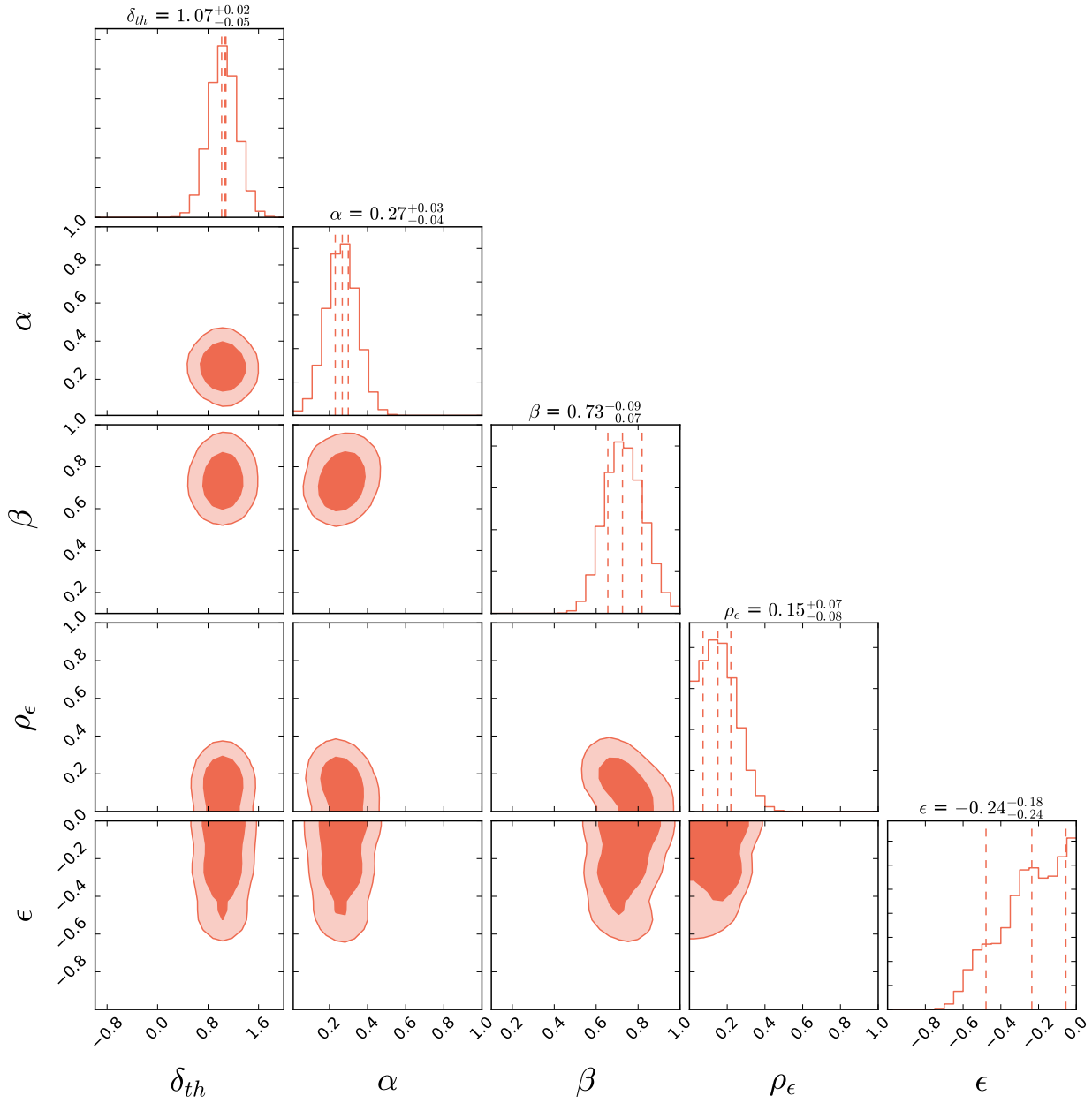

Figure 5.2: Posterior probability distribution of the PATCHY bias parameters $\{\delta_{\rm th}, \alpha, \beta, \rho_\epsilon, \epsilon\}$. The contours mark the 68% and the 95% confidence intervals of the posterior probabilities. This plot is made using the open-source software CORNER (**?**).

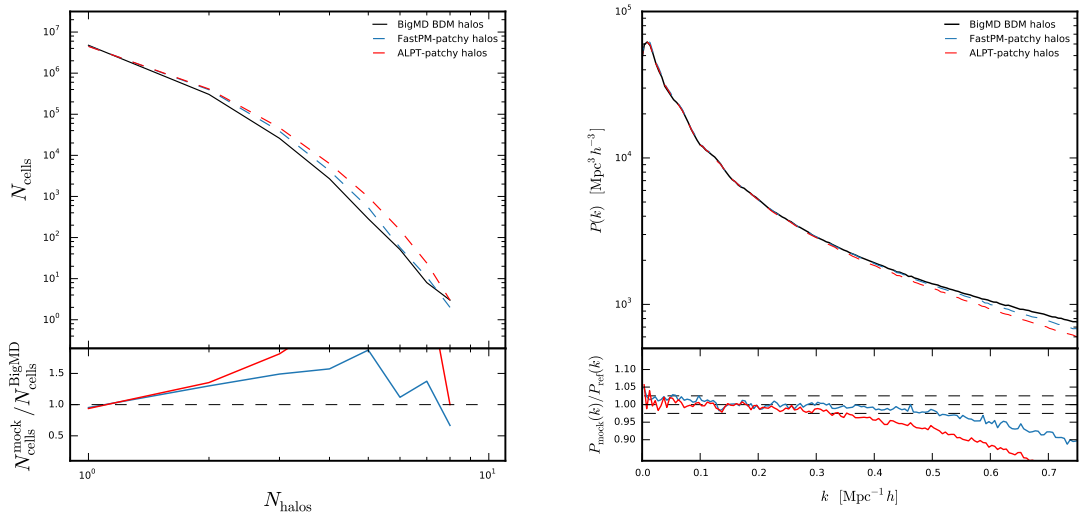

Figure 5.3: Top: Demonstration of the halo bivariate probability distribution function of halos (halo counts-in-cells) in the BigMultiDark simulation (shown in black) and in the FastPM-patchy simulation (shown in blue) and in the ALPT-patchy simulation (shown in red) on the left. Comparison between the real-space power spectrum of the BDM halos (shown in black) in the reference BigMultiDark simulation and that of the halos in the FastPM-patchy (ALPT-patchy) simulation shown in blue (red) on the right. Bottom: Ratio between the halo PDFs of the approximate mocks and halo PDF of the BigMultiDark simulation on the left. Ratio between the halo power spectra of the approximate mocks and the halo power spectrum of the BigMultiDark simulation on the right.

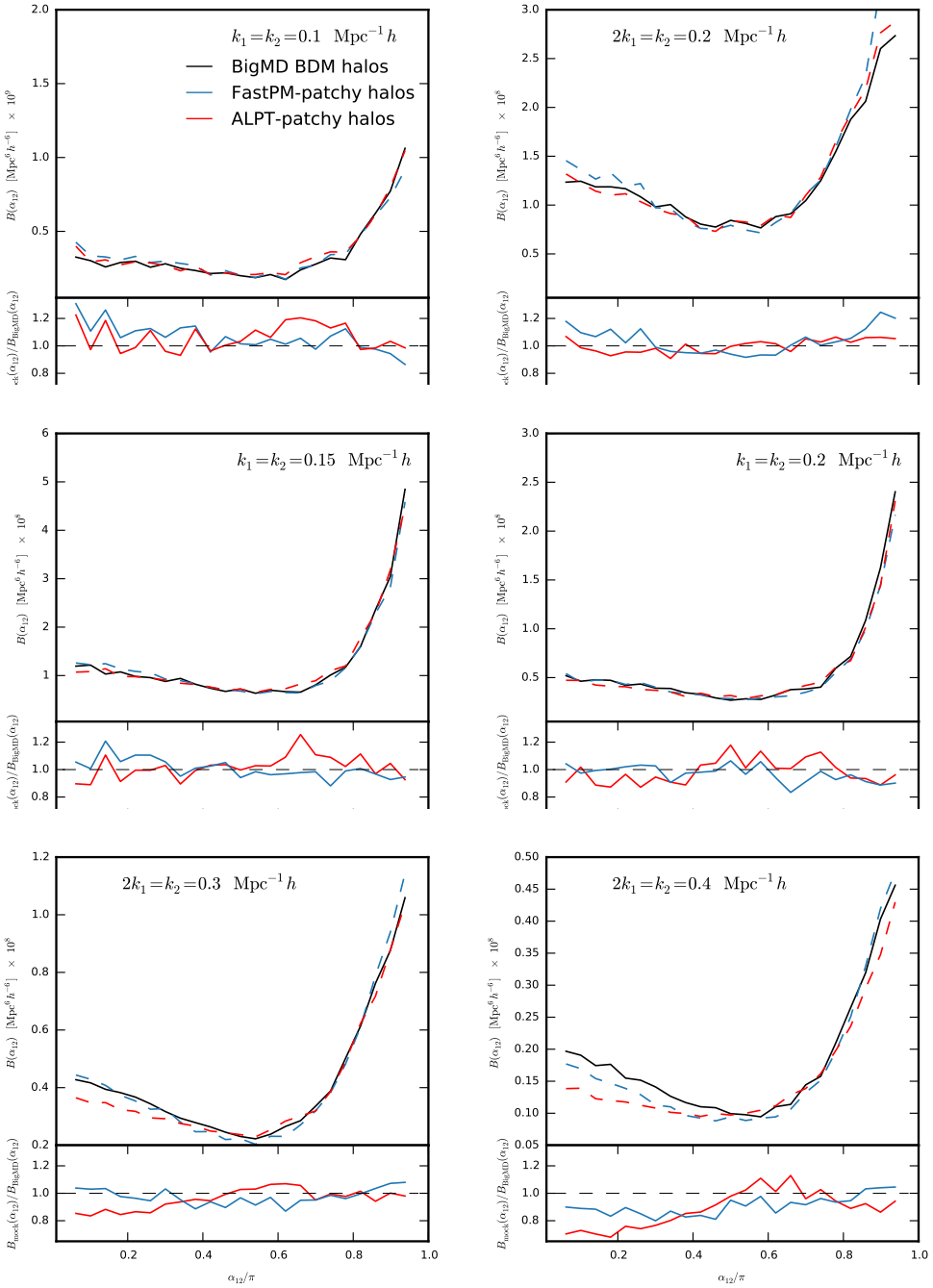

Figure 5.4: Real-space bispectrum of the BigMD BDM halos and that of the approximate mocks as a function of angle α_{12} between $\mathbf{k}_1$ and $\mathbf{k}_2$ for $k_1 = k_2 = 0.1\ h\,\mathrm{Mpc}^{-1}$ (upper left), $2k_1 = k_2 = 0.2\ h\,\mathrm{Mpc}^{-1}$ (upper right), $k_1 = k_2 = 0.15\ h\,\mathrm{Mpc}^{-1}$ (middle left), $k_1 = k_2 = 0.2\ h\,\mathrm{Mpc}^{-1}$ (middle right), $2k_1 = k_2 = 0.3\ h\,\mathrm{Mpc}^{-1}$ (lower left), and $2k_1 = k_2 = 0.4\ h\,\mathrm{Mpc}^{-1}$ (lower right). The BigMD is represent by the solid black line, while ALPT-PATCHY is represented by the dashed red line, and FASTPM-PATCHY is represented by the dashed blue line.

The MultiDark-PATCHY galaxy catalogs (**?**) are calibrated against BOSS-HAM catalogs which were constructed by populating the halos in different snapshots of the BigMultiDark simulation using halo abundance matching (**?**).

Evaluation of $P(k)$ and $\rho(n)$ for a set of bias parameters requires running the forward model of generating halos from the matter density field. Therefore, in order to speed up the fitting procedure we run the PATCHY code with a smaller box size of 625 h^{-1} Mpc and grid size of 240 in each dimension. This choice of box and grid size preserves the resolution. Furthermore, running the PATCHY code and computing the statistics of the halo catalogs in a smaller box size significantly reduces the computational time needed for constraining the bias parameters.

5.4.2 Bias parameters

The first step in our pipeline consists of producing the large scale dark matter field on a mesh. We use the down-sampled white noise of the BigMultiDark simulation from 3840^3 to 960^3 cells to estimate the initial conditions used for both FASTPM and ALPT runs, as shown in Fig. 5.1. The dark matter particles are then assigned to a mesh of 960^3 cells with clouds-in-cells (CIC), which we define as the large scale dark matter density field $\rho_{\rm m}$ required for Eqs. 5.3, 5.4, 5.5.

After running the MCMC chains with the method described in section §5.3, we find constraints on the bias parameters of such equations. These constraints are summarized in Fig. 5.2. The threshold bias parameter $\delta_{\rm th}$ is found to be 1.07 which is equivalent to sampling halos from the regions of high matter overdensity. This supports our intuition that massive halos are generated from high density regions. Our estimated value of the nonlinear bias parameter α is ~ 0.2. These values are qualitatively consistent with ALPT (**?**), although the threshold bias is slightly reduced and the power law bias is slightly higher (parameters with

ALPT: $\delta_{\rm th} \sim 1.2$ and $\alpha \sim 0.12$).

The parameter that governs the deviation from Poissonity β is found to be 0.73. This value is significantly larger than the one found with ALPT (about 0.6), i.e., indicating that the deviation from Poissonity is not so pronounced, as previously found. The reason for this, is that Lagrangian perturbation theory does not manage to model the one halo term, as done with FastPM. Therefore a larger deviation of Poissonity had to be assumed to fit the power spectrum towards small scales, as is demonstrated here. In this sense, a more accurate description of the large scale dark matter field permits us to reduce the stochasticity in the halo distribution.

Furthermore, parameters corresponding to the exponential cutoff term in the deterministic bias relation $\{\rho_\epsilon, \epsilon\}$ are estimated to be $\sim \{0.15, -0.24\}$. While the constraints on both parameters of the exponential cutoff bias are consistent with zero, their presence, albeit being small, is essential in a more accurate modeling of the halo bivariate PDF and the halo bispectrum. By including these extra parameters we demonstrate the flexibility and efficiency of the code to incorporate complex bias models. Furthermore, we believe that the exponential cutoff term will become crucial when considering smaller mass halos, which have a non negligible probability of residing in low density regions (**?**).

5.4.3 Statistical comparison

In this section we discuss the statistical comparisons between the BDM halo catalog of the BigMultiDark simulation and the halo catalog generated from our method. In particular, the FastPM-PATHCY mock is generated using the best-fit bias parameters (see Fig. 5.2). For the ALPT-PATHCY mocks we rely on the parameters found from previous PATCHY studies (**?**). The halo statistical summaries presented in this work are the number density, the bivariate halo probability distribution function (PDF), the real-space power spectrum and

the real-space bispectrum.

By construction our method reproduces the exact number density of halos in the reference catalog (Eq. 5.4). We observe that the bivariate PDF (or halo counts-in-cells) of the reference catalog can be reproduced with good accuracy (Fig. 5.3).

In terms of the agreement between halo PDF of approximate mock catalog and that of the BigMultiDark simulation, we find that significant improvement can be achieved when halos are sampled from the FastPM dark matter density field.

Furthermore, we present our comparison in terms of the power spectrum P and the bispectrum B which are the two-point function and the three-point function in Fourier space. Given the Fourier transform of the halo density field $\delta_h(\mathbf{k})$, the power spectrum and the bispectrum are defined as follows

$$\langle \delta_h(\mathbf{k}_1)\delta_h(\mathbf{k}_2)\rangle = (2\pi)^3 P(k_1)\delta^D(\mathbf{k}_1+\mathbf{k}_2), \tag{5.7}$$

$$\langle \delta_h(\mathbf{k}_1)\delta_h(\mathbf{k}_2)\delta_h(\mathbf{k}_3)\rangle = (2\pi)^3 B(\mathbf{k}_1,\mathbf{k}_2)\delta^D(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3), \tag{5.8}$$

where δ^D is the Dirac delta function. The shot-noise contribution to the power spectrum and bispectrum is modeled in the following way:

$$P_{\rm sn}(k) = \frac{1}{\bar{n}}, \tag{5.9}$$

$$B_{\rm sn}(\mathbf{k}_1,\mathbf{k}_2) = \frac{1}{\bar{n}}[P(k_1)+P(k_2)+P(k_3)] + \frac{1}{\bar{n}^2}, \tag{5.10}$$

where $\bar{n}$ is the halo number density and $k_3 = |\mathbf{k1}+\mathbf{k2}|$.

Our methodology is able to reproduce the halo power spectrum of the reference with percentage level accuracy to $k \sim 0.4\ h\,\mathrm{Mpc}^{-1}$ (within 5% up to $k \sim 0.6\ h\,\mathrm{Mpc}^{-1}$) which

corresponds to nonlinear regimes (Fig. 5.3). We have also run our method ignoring the PDF in the posterior sampling, yielding accurate power spectra up to $k \sim 1\ h\,\mathrm{Mpc}^{-1}$. **?** also reported accurate power spectra up to high k, however, using an arbitrary threshold bias of zero. In a later work additionally fitting the PDF, it was found that the power spectra are accurate within 2% up to $k \sim 0.3\ h\,\mathrm{Mpc}^{-1}$ (**?**), in agreement with what is found here using ALPT. An even higher accuracy will require a more complex bias model and a proper modelling of the clustering on sub-Mpc scales, differentiating between centrals and satellites. The current version of PATCHY randomly assigns dark matter particle positions to halos sampled in a given cell. The bias model could be augmented with nonlocal bias terms following **?**. We have neglected in this study the perturbation bias term used in **?** (in an attempt to compensate for the missing power towards small scales), where the limit in the accuracy was found to be around $k \sim 0.3\ h\,\mathrm{Mpc}^{-1}$. Omitting the perturbation theory term also allows for a fair comparison with the study presented in **?** and is not necessary when using FASTPM.

Fig. 5.3 showed an improved PDF when relying on FASTPM. This is expected to have an impact in the three point statistics, which in fact yields better fits towards small scales, as we discuss below. We show our results in terms of bispectrum for six different values of $|\mathbf{k}_1|$ and $|\mathbf{k}_2|$ as a function of the angle between the two vectors $\alpha_{12} = \angle(\mathbf{k}_1, \mathbf{k}_2)$. The adopted wave numbers are $k_1 = k_2 = 0.1,\ 2k_1 = k_2 = 0.2,\ k_1 = k_2 = 0.15,\ k_1 = k_2 = 0.2,\ 2k_1 = k_2 = 0.3,\ 2k_1 = k_2 = 0.4$ (all wave numbers are expressed in units of $h\,\mathrm{Mpc}^{-1}$).

We find that in general for both ALPT and FASTPM there is good agreement between the bispectrum measured from our approximate mock catalogs and that of the BigMultiDark simulation (Fig. 5.4). Deviations as large as 15-20% are expected, as we are using a down-sampled white noise of the BigMultiDark simulation from 3840^3 to 960^3 cells and are on the level of what was found in **?**.

For configurations corresponding to smaller scales ($2k_1 = k_2 = 0.3\ h\,\mathrm{Mpc}^{-1}$, $2k_1 = k_2 = 0.4\ h\,\mathrm{Mpc}^{-1}$), the agreement between the bispectra of our approximate mock halo catalogs and the BigMultiDark halos improves when we sample halos from the FastPM density field. This improvement is dramatic when compared to EZmocks (see real-space lines in the lower panels in Fig. 5 of **?**).

5.5 Summary and Discussion

This work presents a major step in fast and accurate generation of mock halo/galaxy catalogs, extending in particular, the patchy code. We have introduced an efficient MCMC technique to automatically obtain the bias parameters relating the halo/galaxy population to the underlying large scale dark matter field based on a reference catalog.

This technique is flexible and admits incorporation of different bias models, and number of bias parameters. This permits us to robustly assess the degeneracies and confidence regions of the different bias parameters.

Furthermore we have introduced in the patchy code a particle mesh structure formation model (the FastPM code, see **?**) in addition to the previous LPT based schemes.

As a demonstration of the performance of this method, we used the halo catalog of the BigMultiDark N-body simulation as a reference catalog. Our calibration method makes use of the halo two-point statistics and the counts-in-cells to estimate the bias parameters.

Based on the dark matter field obtained with FastPM, which includes an improved description towards small scales, and in particular, the enhanced power caused by the one halo term, we have found that previous studies were overestimating the contribution to the power due to deviation from Poissonity. Though present, this deviation turns out to be less pronounced. Also, we have managed to extend the accuracy of the power spectra from

$k \sim 0.3\ h\,\mathrm{Mpc}^{-1}$ to $k \sim 0.6\ h\,\mathrm{Mpc}^{-1}$, being at the level of percentage accuracy up to $k \sim 0.4\ h\,\mathrm{Mpc}^{-1}$.

We have demonstrated that the novel implementation of the PATCHY code reaches higher accuracy in terms of the bispectrum towards small scales with respect to LPT based schemes, such as ALPT, and even more so with respect to EZMOCKS, which relies on the Zeldovich approximation.

The assignment of halo masses must be done in a post-processing step taking into account the underlying dark matter density field. **?** demonstrated that the mass assignment is more precise when the underlying dark matter field is more accurate (ALPT vs Zeldovich). We therefore expect that using FASTPM contributes to further reduce the scatter. We leave the investigation of mass assignment for a later work.

We have also left the analysis of redshift space distortions for a future work, as it turns out that the two and three point statistics are apparently more easily described in redshift space (see e.g. **???**). However, a better description of the quadrupole on small scales is not trivial and requires further investigation (see **?**).

As we have now implemented a PM solver into our approach, we expect that certain high mass range of halos are correctly described and could be found with a friends-of-friends algorithm, the halos which are not properly resolved could be augmented with the method presented here (see methods to extend the resolution of N-body simulations, **???**).

It is important to note that our investigation in this work has been focused on the generation of high mass halo (and subhalo) catalogs. One of the main challenges toward generation of mock galaxy catalogs is sampling of low mass halos. These host fainter galaxies which will dominate the observed galaxy samples in upcoming galaxy survey datasets.

We leave a thorough investigation of production of low mass halo catalogs to a future work. This will presumably require more sophisticated bias models including also nonlocal

bias terms. The robust, automatic, and efficient methodology presented in this work should be capable of dealing with this.

In summary, the work presented here contributes to set the basis for a method able to generate galaxy mock catalogs needed to meet the precision requirements of the next generation of galaxy surveys.

Acknowledgments

We are grateful to David W. Hogg, Michael R. Blanton, Uros Seljak, and Jeremy L. Tinker for discussions related to this work. MV is particularly thankful to David W. Hogg for his continuous support during the completion of this work. We thank David W. Hogg anf Alex I. Malz for reading and commenting on the manuscript. FSK thanks Uros Seljak for hospitality at UC Berkeley and LBNL during January to July 2016. During this time he met MV and was able to collaborate with YF. This work was supported by the NSF grant AST-1517237. GY acknowledges financial support from MINECO/FEDER (Spain) under research grant AYA2015-63810-P. Most of the computations in this work were carried out in the New York University High Performance Computing Mercer facility. We thank Shenglong Wang, the administrator of the NYU HPC center, for his consistent support throughout the completion of this study. We also thank Jeremy L. Tinker, Dan Foreman-Mackey, and Mulin Ding for providing some of the computational facilities used in this investigation.

The CosmoSim database used in this paper is a service by the Leibniz-Institute for Astrophysics Potsdam (AIP). The MultiDark database was developed in cooperation with the Spanish MultiDark Consolider Project CSD2009-00064. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) and the Partnership for

Advanced Supercomputing in Europe (PRACE, www.prace-ri.eu) for funding the MultiDark simulation project by providing computing time on the GCS Supercomputer SuperMUC at Leibniz Supercomputing Centre (LRZ, www.lrz.de).

Conclusion

In this dissertation, we study ....

Bibliography

White, M., & Scott, D. 1996, Comments on Astrophysics, 18,

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009

Oh, S. P., Spergel, D. N., & Hinshaw, G. 1999, ApJ, 510, 551

Guth, A. H., & Pi, S.-Y. 1982, Physical Review Letters, 49, 1110

Bardeen, J. M., Steinhardt, P. J., & Turner, M. S. 1983, Phys. Rev. D, 28, 679

Knox, L. 1995, Phys. Rev. D, 52, 4307

Ross, A. J., Percival, W. J., Sánchez, A. G., et al. 2012, MNRAS, 424, 564

Leach, S. M., Cardoso, J.-F., Baccigalupi, C., et al. 2008, A&A, 491, 597

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, Phys. Rep., 367, 1

Dressler, A. 1980, ApJ, 236, 351

Santiago, B. X., & Strauss, M. A. 1992, ApJ, 387, 9

Steidel, C. C., Adelberger, K. L., Dickinson, M., et al. 1998, ApJ, 492, 428

Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51

Kaiser, N. 1984, ApJ, 284, L9

Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709-728

Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, ApJ, 724, 878

Lemson, G., & Kauffmann, G. 1999, MNRAS, 302, 111

Watson, W. A., Iliev, I. T., D'Aloisio, A., et al. 2013, MNRAS, 433, 1230

Anderson, L., Aubourg, E., Bailey, S., et al. 2012, MNRAS, 427, 3435

Eriksen, H. K., O'Dwyer, I. J., Jewell, J. B., et al. 2004, ApJS, 155, 227

Tinker, J. L., Leauthaud, A., Bundy, K., et al. 2013, ApJ, 778, 93

Tinker, J., Wetzel, A., & Conroy, C. 2011, arXiv:1107.5046

Rodríguez-Torres, S. A., Chuang, C.-H., Prada, F., et al. 2016, MNRAS, 460, 1173

Wandelt, B. D., Larson, D. L., & Lakshminarayanan, A. 2004, Phys. Rev. D, 70, 083511

Ishida, E. E. O., Vitenti, S. D. P., Penna-Lima, M., et al. 2015, Astronomy and Computing, 13, 1

Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347

Somerville, R. S., Lemson, G., Sigad, Y., et al. 2001, MNRAS, 320, 289

Casas-Miranda, R., Mo, H. J., Sheth, R. K., & Boerner, G. 2002, MNRAS, 333, 730

Filippi, S., Barnes, C., Cornebise, J., & Stumpf, M. P. H. 2011, arXiv:1106.6280

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., & Robert, C. P. 2008, arXiv:0805.2256

Del Moral, P., Doucet, A., & Jasra, A. 2012, arXiv:1203.0464

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44

Levi, M., Bebek, C., Beers, T., et al. 2013, arXiv:1308.0847

Bolton, A. S., Schlegel, D. J., Aubourg, É., et al. 2012, AJ, 144, 144

Heitmann, K., White, M., Wagner, C., Habib, S., & Higdon, D. 2010, ApJ, 715, 104

Heitmann, K., Higdon, D., White, M., et al. 2009, ApJ, 705, 156

Lawrence, E., Heitmann, K., White, M., et al. 2010, ApJ, 713, 1322

Ata, M., Kitaura, F.-S., & Müller, V. 2015, MNRAS, 446, 4250

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

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

Tinker, J. L., Weinberg, D. H., Zheng, Z., & Zehavi, I. 2005, ApJ, 631, 41

van den Bosch, F. C., Mo, H. J., & Yang, X. 2003, MNRAS, 345, 923

More, S., van den Bosch, F. C., Cacciato, M., et al. 2013, MNRAS, 430, 747

Miyatake, H., More, S., Mandelbaum, R., et al. 2015, ApJ, 806, 1

van den Bosch, F. C., More, S., Cacciato, M., Mo, H., & Yang, X. 2013, MNRAS, 430, 725

Dutton, A. A., & Macciò, A. V. 2014, MNRAS, 441, 3359

More, S., van den Bosch, F. C., & Cacciato, M. 2009, MNRAS, 392, 917

Conroy, C., & Wechsler, R. H. 2009, ApJ, 696, 620

Leauthaud, A., Tinker, J., Bundy, K., et al. 2012, ApJ, 744, 159

Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013, ApJ, 770, 57

Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013, ApJ, 762, 109

Klypin, A., & Holtzman, J. 1997, arXiv:astro-ph/9712217

Knebe, A., Knollmann, S. R., Muldrew, S. I., et al. 2011, MNRAS, 415, 2293

Kravtsov, A. V., Klypin, A. A., & Khokhlov, A. M. 1997, ApJS, 111, 73

Carlson, J., White, M., & Padmanabhan, N. 2009, Phys. Rev. D, 80, 043531

White, M., Tinker, J. L., & McBride, C. K. 2014, MNRAS, 437, 2594

Tassev, S., Zaldarriaga, M., & Eisenstein, D. J. 2013, J. Cosmology Astropart. Phys., 6, 036

Feng, Y., Chu, M.-Y., Seljak, U., & McDonald, P. 2016, MNRAS, 463, 2273

Klypin, A. A., Trujillo-Gomez, S., & Primack, J. 2011, ApJ, 740, 102

Riebe, K., Partl, A. M., Enke, H., et al. 2011, arXiv:1109.0003

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760

Leauthaud, A., Tinker, J., Behroozi, P. S., Busha, M. T., & Wechsler, R. H. 2011, ApJ, 738, 45

Hearin, A. P., Zentner, A. R., van den Bosch, F. C., Campbell, D., & Tollerud, E. 2016, MNRAS, 460, 2552

Hearin, A., Campbell, D., Tollerud, E., et al. 2016, arXiv:1606.04106

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Lin, C.-A., & Kilbinger, M. 2015, A&A, 583, A70

Lin, C.-A., Kilbinger, M., & Pires, S. 2016, A&A, 593, A88

Weyant, A., Schafer, C., & Wood-Vasey, W. M. 2013, ApJ, 764, 116

Akeret, J., Refregier, A., Amara, A., Seehars, S., & Hasner, C. 2015, J. Cosmology Astropart. Phys., 8, 043

Ishida, E. E. O., Vitenti, S. D. P., Penna-Lima, M., et al. 2015, Astronomy and Computing, 13, 1

Cameron, E., & Pettitt, A. N. 2012, MNRAS, 425, 44

Berlind, A. A., Frieman, J., Weinberg, D. H., et al. 2006, ApJS, 167, 1

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

Peebles, P. J. E. 1980, Research supported by the National Science Foundation. Princeton, N.J., Princeton University Press, 1980. 435 p.,

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, A&A, 571, A16

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13

Seljak, U. 2000, MNRAS, 318, 203

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, ApJ, 546, 20

Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791

Bishop, C. M., & Nasrabadi, N. M. 2007, Journal of Electronic Imaging, 16, 049901

Cooray, A., & Sheth, R. 2002, Phys. Rep., 372, 1

Chuang, C.-H., Zhao, C., Prada, F., et al. 2015, MNRAS, 452, 686

Heitmann, K., Lukić, Z., Fasel, P., et al. 2008, Computational Science and Discovery, 1, 015003

Silk, D., Filippi, S., & Stumpf, M. P. H. 2012, arXiv:1210.3296

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A20

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A17

Navarro, J. F., Hayashi, E., Power, C., et al. 2004, MNRAS, 349, 1039

Padmanabhan, N., White, M., Zhou, H. H., & O'Connell, R. 2016, MNRAS, 460, 1567

Ahn, K., Iliev, I. T., Shapiro, P. R., Srisawat, C. 2015, MNRAS, 450, 1486

Angulo, R. E., Baugh, C. M., Frenk, C. S., Lacey, C. G. 2014, MNRAS, 442, 3256

Ata, M., Kitaura, F.-S., Müller, V. 2015, MNRAS, 446, 4250

Bagla, J. S. 2002, Journal of Astrophysics and Astronomy, 23, 185

Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15

Blot, L., Corasaniti, P. S., Amendola, L., & Kitching, T. D. 2016, MNRAS, 458, 4462

Bouchet, F. R., Colombi, S., Hivon, E., & Juszkiewicz, R. 1995, A&A, 296, 575

Buchert, T., & Ehlers, J. 1993, MNRAS, 264,

Catelan, P. 1995, MNRAS, 276, 115

Casas-Miranda, R., Mo, H. J., Sheth, R. K., & Boerner, G. 2002, MNRAS, 333, 730

Cen, R., & Ostriker, J. P. 1993, ApJ, 417, 415

Chuang, C.-H., Kitaura, F.-S., Prada, F., Zhao, C., & Yepes, G. 2015, MNRAS, 446, 2621

Chuang, C.-H., Zhao, C., Prada, F., et al. 2015, MNRAS, 452, 686

Crocce, M., Cabré, A., & Gaztañaga, E. 2011, MNRAS, 414, 329

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44

de la Torre, S., & Peacock, J. A. 2013, MNRAS, 435, 743

Dodelson, S., & Schneider, M. D. 2013, Phys. Rev. D, 88, 063537

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23

Feng, Y., Chu, M.-Y., Seljak, U., & McDonald, P. 2016, MNRAS, 463, 2273

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 24, Url = http://dx.doi.org/10.5281/zenodo.45906

Frieman, J., & Dark Energy Survey Collaboration 2013, American Astronomical Society Meeting Abstracts #221, 221, 335.01

Fry, J. N., & Gaztanaga, E. 1993, ApJ, 413, 447

Gelman, A., & Rubin, D. B. 1992, Statistical Science, 457

Gil-Marín, H., Percival, W. J., Verde, L., et al. 2017, MNRAS, 465, 1757

Gil-Marín, H., Noreña, J., Verde, L., et al. 2015, MNRAS, 451, 539

Gil-Marín, H., Verde, L., Noreña, J., et al. 2015, MNRAS, 452, 1914

Goodman, J., & Weare, J. 2010, Communications in applied mathematics and computational science, 5, 65

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., & Dalla Vecchia, C. 2016, MNRAS, 457, 1577

Guo, H., Zheng, Z., Behroozi, P. S., et al. 2016, ApJ, 831, 3

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399

Howlett, C., Manera, M., & Percival, W. J. 2015, Astronomy and Computing, 12, 109

Izard, A., Crocce, M., & Fosalba, P. 2016, MNRAS, 459, 2327

Joachimi, B. 2016, arXiv:1612.00752

Kaiser, N. 1984, ApJ, 284, L9

Kitaura, F.-S., & Heß, S. 2013, MNRAS, 435, L78

Kitaura, F.-S., Yepes, G., & Prada, F. 2014, MNRAS, 439, L21

Kitaura, F.-S., Gil-Marín, H., Scóccola, C. G., et al. 2015, MNRAS, 450, 1836

Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., et al. 2016, MNRAS, 456, 4156

Kalus, B., Percival, W. J., & Samushia, L. 2016, MNRAS, 455, 2573

Klypin, A., & Holtzman, J. 1997, arXiv:astro-ph/9712217

Klypin, A., Yepes, G., Gottlöber, S., Prada, F., & Heß, S. 2016, MNRAS, 457, 4340

Knebe, A., Knollmann, S. R., Muldrew, S. I., et al. 2011, MNRAS, 415, 2293

Koda, J., Blake, C., Beutler, F., Kazin, E., & Marin, F. 2016, MNRAS, 459, 2118

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193

Ledoit, O., & Wolf, M. 2004, Journal of Multivariate Analysis, 88, 365

Ledoit, O., & Wolf, M. 2012, The Annals of Statistics, 40, 1024

Levi, M., Bebek, C., Beers, T., et al. 2013, arXiv:1308.0847

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201

Manera, M., Scoccimarro, R., Percival, W. J., et al. 2013, MNRAS, 428, 1036

Manera, M., Samushia, L., Tojeiro, R., et al. 2015, MNRAS, 447, 437

McDonald, P., Roy, A. 2009, J. Cosmology Astropart. Phys., 8, 20

Mo, H. J., & White, S. D. M. 2002, MNRAS, 336, 112

Monaco, P., Theuns, T., Taffoni, G. et al. 2002, ApJ, 564, 8

Monaco, P., Sefusatti, E., Borgani, S., et al. 2013, MNRAS, 433, 2389

Monaco, P. 2016, Galaxies, 4, 53

Morrison, C. B., & Schneider, M. D. 2013, J. Cosmology Astropart. Phys., 11, 009

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Neyrinck, M. C., Aragón-Calvo, M. A., Jeong, D., & Wang, X. 2014, MNRAS, 441, 646

Padmanabhan, N., White, M., Zhou, H. H., & O'Connell, R. 2016, MNRAS, 460, 1567

Peebles, P. J. E. 1980, Research supported by the National Science Foundation. Princeton, N.J., Princeton University Press, 1980. 435 p.,

Pope, A. C., & Szapudi, I. 2008, MNRAS, 389, 766

Rodríguez-Torres, S. A., Chuang, C.-H., Prada, F., et al. 2016, MNRAS, 460, 1173

Schneider, P., van Waerbeke, L., Kilbinger, M., & Mellier, Y. 2002, A&A, 396, 1

Scoccimarro, R., Sheth, R. K. 2002, MNRAS, 329, 629

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

Simpson, F., Blake, C., Peacock, J. A., et al. 2016, Phys. Rev. D, 93, 023525

Slepian, Z., Eisenstein, D. J., Beutler, F., et al. 2015, arXiv:1512.02231

Slepian, Z., Eisenstein, D. J., Blazek, J. A., et al. 2016, arXiv:1607.06098

Slepian, Z., Eisenstein, D. J., Brownstein, J. R., et al. 2016, arXiv:1607.06097

Smith, R. E., Scoccimarro, R., & Sheth, R. K. 2008, Phys. Rev. D, 77, 043525

Somerville, R. S., Lemson, G., Sigad, Y., et al. 2001, MNRAS, 320, 289

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv:1503.03757

Springel, V. 2005, MNRAS, 364, 1105

Sun, L., Wang, Q., & Zhan, H. 2013, ApJ, 777, 75

Tassev, S., Zaldarriaga, M., & Eisenstein, D. J. 2013, J. Cosmology Astropart. Phys., 6, 036

Tassev, S., Eisenstein, D. J., Wandelt, B. D., & Zaldarriaga, M. 2015, arXiv:1502.07751

Taylor, A., Joachimi, B., & Kitching, T. 2013, MNRAS, 432, 1928

Taylor, A., & Joachimi, B. 2014, MNRAS, 442, 2728

White, M., Blanton, M., Bolton, A., et al. 2011, ApJ, 728, 126

White, M., Tinker, J. L., & McBride, C. K. 2014, MNRAS, 437, 2594

Zhao, C., Kitaura, F.-S., Chuang, C.-H., et al. 2015, MNRAS, 451, 4266