



This article has been accepted for publication in Monthly Notices of the Royal Astronomical Society ©: 2019 The Authors. Published by Oxford University Press on behalf of the Royal Astronomical Society. All rights reserved.

UNIT project: Universe N -body simulations for the Investigation of Theoretical models from galaxy surveys

Chia-Hsun Chuang¹  ¹★ Gustavo Yepes^{2,3} ^{2,3}★ Francisco-Shu Kitaura^{4,5}
 Marcos Pellejero-Ibanez^{4,5} Sergio Rodríguez-Torres^{2,3} Yu Feng⁶
 Robert Benton Metcalf^{7,8} Risa H. Wechsler^{1,9,10} Cheng Zhao^{11,12}
 Chun-Hao To^{1,9,10} Shadab Alam¹³ Arka Banerjee^{1,9,10} Joseph DeRose^{1,9,10}
 Carlo Giocoli^{7,8,14,15} Alexander Knebe¹⁶  ^{2,3,16} and Guillermo Reyes^{2,3}

Affiliations are listed at the end of the paper

Accepted 2019 April 29. Received 2019 April 22; in original form 2018 November 14

ABSTRACT

We present the UNIT N -body cosmological simulations project, designed to provide precise predictions for non-linear statistics of the galaxy distribution. We focus on characterizing statistics relevant to emission line and luminous red galaxies in the current and upcoming generation of galaxy surveys. We use a suite of precise particle mesh simulations (FASTPM) as well as with full N -body calculations with a mass resolution of $\sim 1.2 \times 10^9 h^{-1} M_{\odot}$ to investigate the recently suggested technique of Angulo and Pontzen to suppress the variance of cosmological simulations. We study redshift-space distortions, cosmic voids, higher order statistics from $z = 2$ down to 0. We find that both two- and three-point statistics are unbiased. Over the scales of interest for baryon acoustic oscillations and redshift-space distortions, we find that the variance is greatly reduced in the two-point statistics and in the cross-correlation between haloes and cosmic voids, but is not reduced significantly for the three-point statistics. We demonstrate that the accuracy of the two-point correlation function for a galaxy survey with effective volume of $20 (h^{-1} \text{Gpc})^3$ is improved by about a factor of 40, indicating that two pairs of simulations with a volume of $1 (h^{-1} \text{Gpc})^3$ lead to the equivalent variance of ~ 150 such simulations. The N -body simulations presented here thus provide an effective survey volume of about seven times the effective survey volume of Dark Energy Spectroscopic Instrument or *Euclid*. The data from this project, including dark matter fields, halo catalogues, and their clustering statistics, are publicly available.

Key words: large-scale structure of the Universe.

1 INTRODUCTION

The large-scale structure measured in galaxy surveys represents one of the most powerful probes of present-day cosmology and the nature of dark matter and dark energy in the Universe. To this end, a considerable observational effort is being put forward to map the 3D galaxy distribution in the Universe at unprecedented scales with large photometric and spectroscopic surveys that will measure the positions of tens to hundreds of millions of galaxies. The current largest photometric and spectroscopic surveys are the Dark Energy Survey¹ and the Extended Baryon Oscillation Spectroscopic

Survey,² respectively. The total volume of the Universe mapped with galaxy surveys is dramatically increasing, with several large upcoming ground- and space-based experiments being planned, including as 4MOST³ (4-metre Multi-Object Spectroscopic Telescope, de Jong et al. 2012), DESI⁴ (Dark Energy Spectroscopic Instrument, Schlegel et al. 2011; Levi et al. 2013), HETDEX⁵ (Hobby–Eberly Telescope Dark Energy Experiment, Hill et al. 2008), J-PAS⁶ (Javalambre Physics of accelerating universe As-

²<http://www.sdss.org/sdss-surveys/eboss/>

³<http://www.4most.eu/>

⁴<http://desi.lbl.gov/>

⁵<http://hetdex.org>

⁶<http://j-pas.org>

* E-mail: chuangch@stanford.edu (CHC); gustavo.yepes@uam.es (GY)

¹<http://www.darkenergysurvey.org>

Table 1. Overview of the set of simulations performed for this study and their corresponding parameter settings, including 800 FASTPM and two pairs of GADGET simulations. LR and HR refer to low and high resolutions, respectively.

Simulation code	Amplitude	Phases	Box side length	Number of particles	Particle $M [h^{-1} M_{\odot}]$	Force resolution	Number of boxes
GADGET \bar{G}	Fixed	Regular	$1 h^{-1} \text{Gpc}$	4096^3	1.2×10^9	$6 h^{-1} \text{kpc}$	2
GADGET \bar{G}	Fixed	Inverse phase of \bar{G}	$1 h^{-1} \text{Gpc}$	4096^3	1.2×10^9	$6 h^{-1} \text{kpc}$	2
FASTPM \bar{A}	Non-fixed	Regular reference LR	$1 h^{-1} \text{Gpc}$	1024^3	7.68×10^{10}	$1.46 h^{-1} \text{Mpc}$	100
FASTPM \bar{A}	Non-fixed	Inverse phase of \bar{A}	$1 h^{-1} \text{Gpc}$	1024^3	7.68×10^{10}	$1.46 h^{-1} \text{Mpc}$	100
FASTPM \bar{B}	Fixed	Regular	$1 h^{-1} \text{Gpc}$	1024^3	7.68×10^{10}	$1.46 h^{-1} \text{Mpc}$	100
FASTPM \bar{B}	Fixed	Inverse phase of \bar{B}	$1 h^{-1} \text{Gpc}$	1024^3	7.68×10^{10}	$1.46 h^{-1} \text{Mpc}$	100
FASTPM \bar{C}	Non-fixed	Regular reference HR	$250 h^{-1} \text{Mpc}$	1024^3	1.2×10^9	$0.36 h^{-1} \text{Mpc}$	100
FASTPM \bar{C}	Non-fixed	Inverse phase of \bar{C}	$250 h^{-1} \text{Mpc}$	1024^3	1.2×10^9	$0.36 h^{-1} \text{Mpc}$	100
FASTPM \bar{D}	Fixed	Regular	$250 h^{-1} \text{Mpc}$	1024^3	1.2×10^9	$0.36 h^{-1} \text{Mpc}$	100
FASTPM \bar{D}	Fixed	Inverse phase of \bar{D}	$250 h^{-1} \text{Mpc}$	1024^3	1.2×10^9	$0.36 h^{-1} \text{Mpc}$	100

trophysical Survey, Benitez et al. 2014), PFS⁷ (Subaru Prime Focus Spectrograph, Takada et al. 2014), LSST⁸ (Large Synoptic Survey Telescope, Abell et al. 2009), *Euclid*⁹ (Laureijs et al. 2011), and *WFIRST*¹⁰ (*Wide-Field Infrared Survey Telescope*, Spergel et al. 2013).

In order to extract cosmological constraints from these surveys as well as the data allow, the systematic errors associated with theoretical models that for example characterize the galaxy power spectrum or correlation function as a function of cosmological model must be well below the statistical uncertainties caused by cosmic variance and shot noise. Some pioneering analytical models have been developed to compute the theoretical expected correlation function. To date, these models have limited accuracy, as they rely on analytical gravity models (e.g. Zel’dovich 1970), simplified biasing descriptions, and approximate redshift-space distortion models (RSDs, see White 2015, and references therein). These models have not achieved the accuracy possible with a numerical computation using a full gravity solver. To meet the goals of current galaxy surveys, simulations with much larger effective volumes than those probed by the surveys *and* with enough mass resolution to resolve the dark matter haloes hosting the typical galaxies detected in those surveys are required. Yet, the computational resources needed to accomplish this task are at the edge of the current (petaflop) computational power. A single simulation with the required halo mass resolution ($\sim 4 \times 10^{11} h^{-1} M_{\odot}$, Cochrane et al. 2017) that covers the whole volume sampled by *Euclid* ($\sim 70 \text{Gpc}^3$) would demand an enormous number of particles (more than 16000^3 in a $4 \text{Gpc} h^{-1}$ box). The largest N -body simulations performed so far, e.g., MillenniumXXL (Angulo et al. 2012), MICE (Fosalba et al. 2015), MultiDark (Klypin et al. 2016), Dark Sky (Skillman et al. 2014), OuterRim (Habib et al. 2016), and FLAGSHIP (Potter, Stadel & Teyssier 2016), are still well below this particle number despite their computational expense.

In this project, we explore an alternative way of reaching the same level of required accuracy using far fewer computational resources. Recently, Angulo & Pontzen (2016) proposed a new method to dramatically reduce cosmic variance arising from the sparse sampling of large-scale wave modes in cosmological simulations. The method uses pairs of simulations (Pontzen et al. 2016) with initial Fourier-mode amplitudes that are fixed to the ensemble-averaged

power spectrum and initial modes that are exactly out of phase (one of the pair of simulations has opposite phases with respect to its companion). Using this methodology, one can potentially obtain a result that is statistically equivalent to the mean of many independent simulations from a single pair of simulations.

To date, the method has been tested only on the dark matter distribution at high redshifts ($z = 1$) obtained with particle mesh (PM) gravity solvers, with a low-resolution (LR) particle mass of $1.7 \times 10^{12} h^{-1} M_{\odot}$. Given that fixed initial conditions cease to be formally a Gaussian field with a fixed power spectrum, there has been concern about potential biases this approach could introduce in the clustering statistics, although Angulo & Pontzen (2016) gave analytical arguments why the biases should be negligible. Villaescusa-Navarro et al. (2018) further tested this method using hydrodynamical simulations. These simulations were done with small volumes ($20 h^{-1} \text{Mpc}$ on a side) or with LR (particle mass of $6.6 \times 10^{11} h^{-1} M_{\odot}$) compared to the requirements of current and upcoming surveys.

There is a need to directly test the usefulness and applicability of this approach to key large-scale structure analyses, including baryon acoustic oscillations (BAOs) and RSDs, that are expected with upcoming large surveys. That is the goal of this work. Here, we extend these studies to the statistics, redshift range, galaxy samples, resolution, and volume required by surveys such as DESI and *Euclid*. We use volumes of $(1 h^{-1} \text{Gpc})^3$ in our studies. Such volumes have been claimed to be large enough to account for large-scale mode coupling (Klypin & Prada 2018), although this likely needs to be further investigated; missing modes may need to be accounted for in a post-processing step (Chuang et al. in preparation).

We have designed the simulations in this work to focus on the key cosmology samples of upcoming surveys, which require robust modelling that encompasses the expected halo masses for emission-line galaxies (ELGs, $\sim 10^{11} h^{-1} M_{\odot}$, González-Pérez et al. 2018) and $H\alpha$ galaxies ($\sim 4 \times 10^{11} h^{-1} M_{\odot}$, Cochrane et al. 2017). The simulation boxes are $1 h^{-1} \text{Gpc}$ on a side, with 4096^3 particles and a mass resolution of $1.2 \times 10^9 h^{-1} M_{\odot}$. We are thus able to safely resolve all haloes with masses larger than $1.2 \times 10^{11} h^{-1} M_{\odot}$, using 100 particles per halo.

We demonstrate that the resulting errors in the statistical correlation function measurements using the suppressed variance method are equivalent to having more than seven times the effective volume sampled by DESI or *Euclid* galaxies ($\sim 20 (h^{-1} \text{Gpc})^3$). We generate halo catalogues and merger trees using the publicly available ROCKSTAR halo finder (Behroozi, Wechsler & Wu 2013a), together with density and velocity fields on a mesh for later construction of light-cone distributions of galaxies and weak-lensing maps. In

⁷<https://pfs.ipmu.jp>

⁸<http://www.lsst.org/lst/>

⁹<http://www.euclid-ec.org>

¹⁰<http://wfirst.gsfc.nasa.gov>

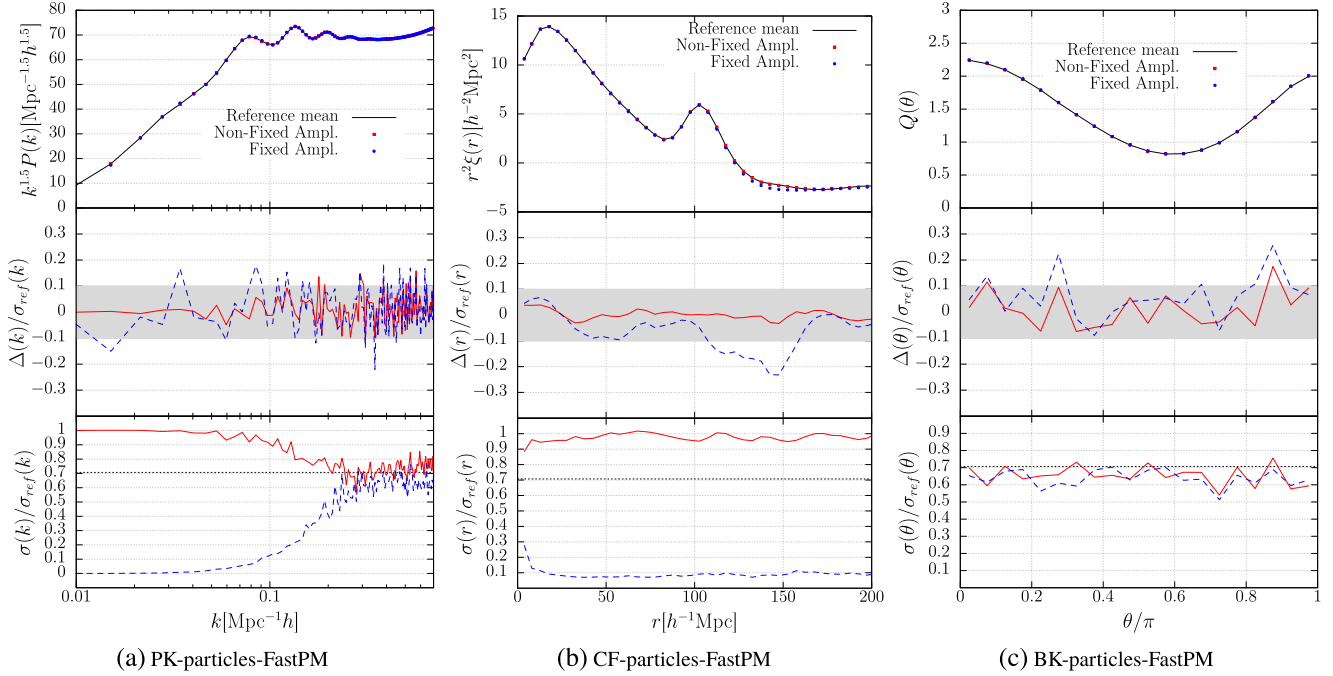


Figure 1. Performance of the SVM for dark matter particles in real space in the two- (Fourier and configuration space) and three-point statistics. We show the clustering statistics for dark matter particles from FASTPM runs with box size = $1 h^{-1}$ Gpc and 1024^3 particles at $z = 1$. Left-hand, centre, and right-hand panels present the power spectra, correlation functions, and bispectra, respectively. The regular simulation set up is shown in black; the set of paired non-fixed-amplitude simulations in red, and the set of paired-fixed-amplitude simulations in blue. The middle row shows the ratio between $\Delta(k)$ and the standard deviation from the reference LR set. Since the uncertainty on the mean should be inverse proportional to $\sqrt{100}$, deviations between the means should be considered as unbiased if they agree within $0.1\sigma_{\text{ref}}$ (grey region). We confirm that the SVM does not introduce any statistic significant bias at any scale in the considered range. The correlations among the data points of a correlation function are large at larger scales, so that the deviations shown in the centre plot are not statistically significant either. The bottom row shows the ratio between the standard deviations from each paired set, $\sigma(k)$, and the reference LR set, $\sigma_{\text{ref}}(k)$. Since we compare the sets of paired simulation with the reference simulations, if the uncertainty is reduced by only $1/\sqrt{2} \sim 0.7$, it indicates no improvement. We find significant improvement of the uncertainty from the set of paired-fixed-amplitude simulations. However, power spectrum shows that the improvement significantly depends on the scale.

future work, we will use these simulations to produce thousands of catalogues, including mock galaxies with various techniques. The corresponding data will be made publicly available through data bases and web portals for the general use of the astrophysical community¹¹.

This paper is organized as follows. First, we present our study of the potential systematic biases from suppressed variance methods (hereafter SVM; Section 2). In Section 3, we present our suppressed variance simulation products including a clustering analysis and a robust assessment of the improvement. We summarize and conclude in Section 4. Throughout this work, we use the following cosmological parameters: $\Omega_m = 0.3089$, $h \equiv H_0/100 = 0.6774$, $n_s = 0.9667$, and $\sigma_8 = 0.8147$ (see table 4 in Planck Collaboration XIII et al. 2016).

2 ASSESMENT OF POTENTIAL SYSTEMATIC BIASES IN SVM

We begin by studying the potential systematic biases and the improvement introduced by the SVM. To this end, we want to generate a large total volume and number of simulations that permit us to estimate the uncertainties of the measurements and to quantify the improvements in the uncertainties on different clustering measurements.

To create large simulated volumes, we rely on accelerated PM solvers, which have been recently shown to produce accurate halo populations compared to full N -body calculations, when enhanced with various techniques (see the COLA code, Tashev, Zaldarriaga & Eisenstein 2013 or the FASTPM code, Feng et al. 2016).

2.1 Setup

We use the C implementation of the FASTPM software, which employs a pencil domain-decomposition Poisson solver and a Fourier-space four-point differential kernel to compute the force. The time integration scheme is modified from a vanilla leap-frog scheme to account for the acceleration of velocity during a step and thus to correctly track the linear growth of large-scale modes regardless of the number of time steps.

This permits us to efficiently perform 800 paired simulations to benchmark the method. Half of them (400) have LR (1024^3 particles) but large volume ($1 h^{-1}$ Gpc side), and the other half (400) have enough resolution for the purpose of this study (1024^3 particles, see Section 3) but smaller volume ($250 h^{-1}$ Mpc side). In each case, we run half of the paired simulations with normal Gaussian random field initial conditions and half with fixed amplitudes, yielding 100 pairs of simulations for each case (see Table 1).

The simulations are started at $a \equiv 1/(1+z) = 0.01$ ($z = 99$), and evolved to $a = 1$ ($z = 0$) with 100 time-steps. We save snapshots of the simulations at $z = 2, 1$, and 0. These regular simulation

¹¹see <http://www.unitsims.org>

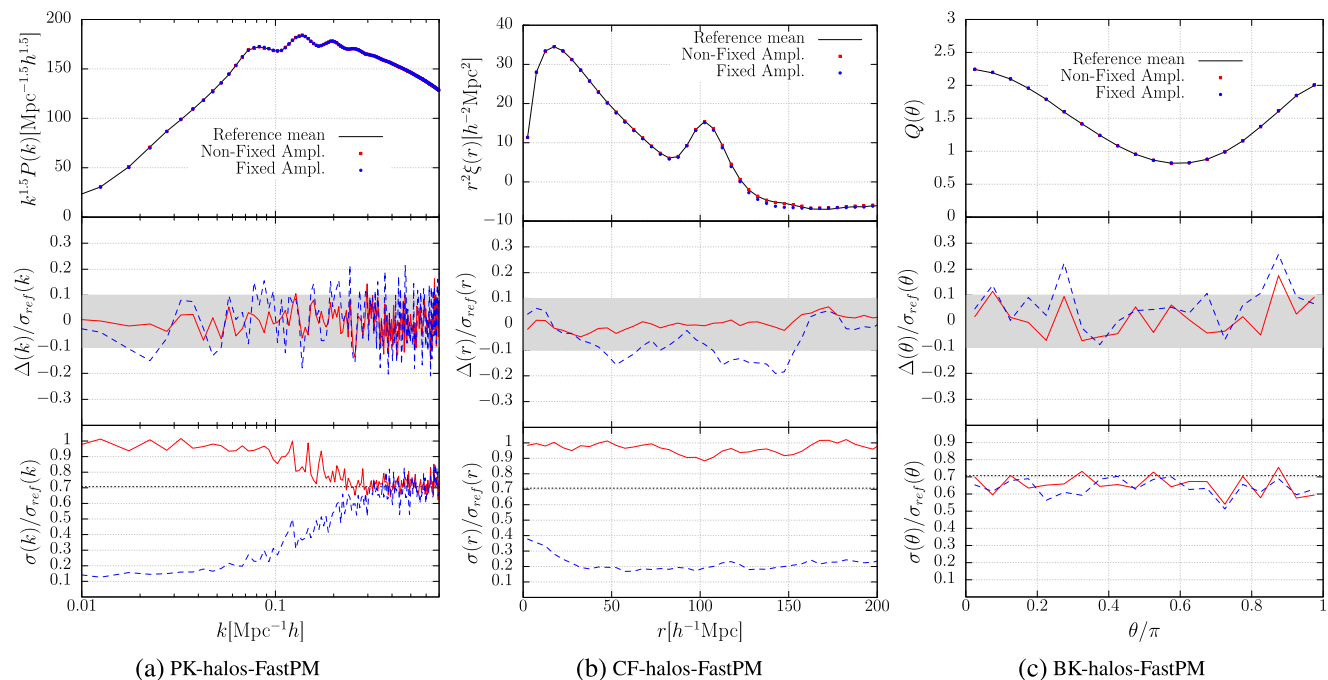


Figure 2. Performance of the SVM for dark matter haloes in real space in the two- (Fourier and configuration space) and three-point statistics. The figure replicates Fig. 1, but for haloes. We do not find any bias in these measurements (see the middle panels). The improvements in the uncertainties are weaker than the ones from dark matter clustering measurements but are still significant. There is no improvement in the power spectrum for $k > 0.3 \text{Mpc}^{-1} h$ at $z = 1$.

sets in high resolution and LR define our reference simulations (see Table 1), from which we compute the standard mean summary statistics.

When computing haloes with the Friends-of-Friends halo finder in NBODYKIT (Hand et al. 2017), we choose a minimum of 20 dark matter particles per halo and a linking length of $0.2 L_{\text{box}}/N_c$. Here, L_{box} refers to the size of one side of the simulation box and N_c to the number of cells along one axis used in the mesh computation, which was taken to correspond to the number of dark matter particles.

2.2 Results from particle mesh simulations

We perform an analysis of several different clustering statistics, including the power spectrum (PK), correlation function (CF), and bispectrum (BK, as defined for instance in Chuang et al. 2017) using dark matter particles and haloes based on the set of fast particle mesh simulations. We demonstrate below that there are no systematic biases using the SVM, and that the variance is indeed greatly reduced in two-point statistics over the scales of interest to BAO and RSD analyses.

We quantify the uncertainties and biases in these measurements through the standard deviation of the reference simulations: $\sigma_{\text{ref}}(k)$, and the deviation of the mean with respect to the reference mean: $\Delta(k)$. The top panels in Figs 1 and 2 show the original comparison of the clustering statistics; the middle panels show the comparison of the mean normalized by the uncertainty of the reference LR 100 boxes. Since the uncertainty on the mean should be inversely proportional to $\sqrt{100}$, deviations between the means should be considered as unbiased if they agree within $0.1\sigma_{\text{ref}}$. This study is performed for dark matter particles, haloes, and cosmic voids, as we discuss below.

2.2.1 Dark matter particles

The largest suppression of variance is obtained for the dark matter particle distribution. Fig. 1 shows the comparison of the dark matter particle clustering measurements from FASTPM runs with $1 h^{-1} \text{Gpc}$ side boxes and 1024^3 particles at $z = 1$, including the PK, CF, and BK. In each plot, the top panel shows the clustering measurements of the reference set, the set of paired simulations (non-fixed-amplitude), and the set of paired-fixed-amplitude simulations; the middle panel shows the difference of the mean from each paired set and the reference one divided by the standard deviation from the reference set; the bottom panel shows the ratios of the standard deviations from each paired set and the reference one. From these calculations, we confirm that the SVM does not introduce significant bias at any scale in the considered range. Since the paired simulations have twice volume of the reference simulation, there is no improvement if the uncertainty of the paired simulations is larger than or equal to $1/\sqrt{2} \sim 0.7$ of that measured from the reference simulations. We find that the improvement depends on scale. In the case of the power spectrum, we find that the improvement is significant at small k (large scales) but small at large k . If one considers for instance $k > 0.3 h \text{Mpc}^{-1}$, our results indicate that the variance at large k is dominated by higher order mode coupling terms. Interestingly, the improvement in the correlation function variance is nearly constant ($0.1\sigma_{\text{ref}}$) over the range for $r > 10 h^{-1} \text{Mpc}$. We do not find any improvement in the bispectrum with triangle configurations of $k_1 = 0.1$ and $k_2 = 0.2 h \text{Mpc}^{-1}$. Since we are interested in scales relevant to BAO and RSD analysis, this study goes further into the non-linear regime than did the study of Angulo & Pontzen (2016), which found models improvement for the bispectrum variance at scales that are significantly more linear. The data products made available at the project website will enable a deeper investigation of these effects.

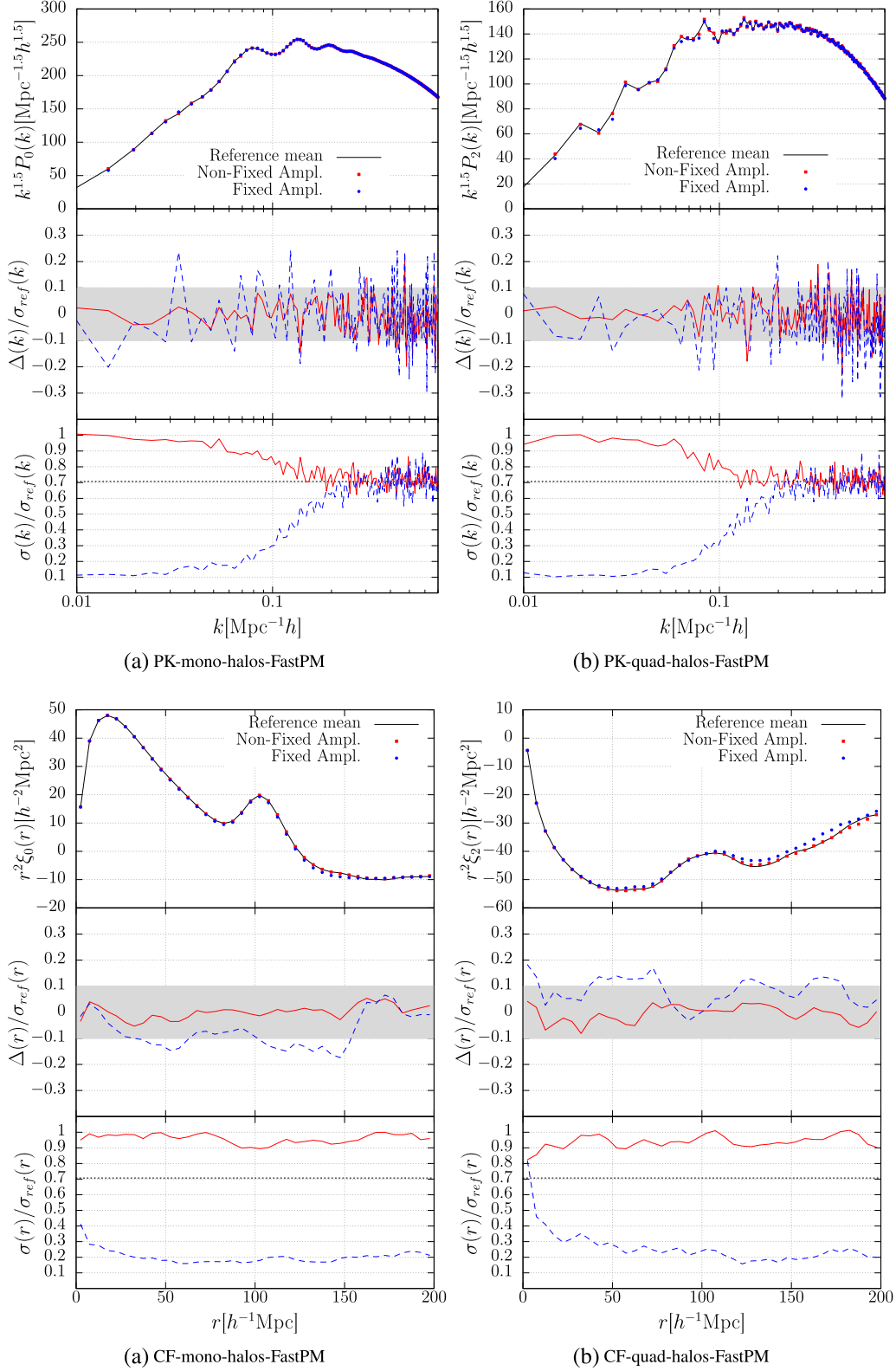


Figure 3. Performance of the SVM for dark matter haloes in redshift space in the monopole and quadrupole, for both Fourier and configuration spaces. Panels show the monopole (left) and the quadrupole (right) in Fourier space (top) and configuration space (bottom). The same conventions as in Fig. 2 are used. We do not find any bias in these measurements. The improvements are similar to those found in real space.

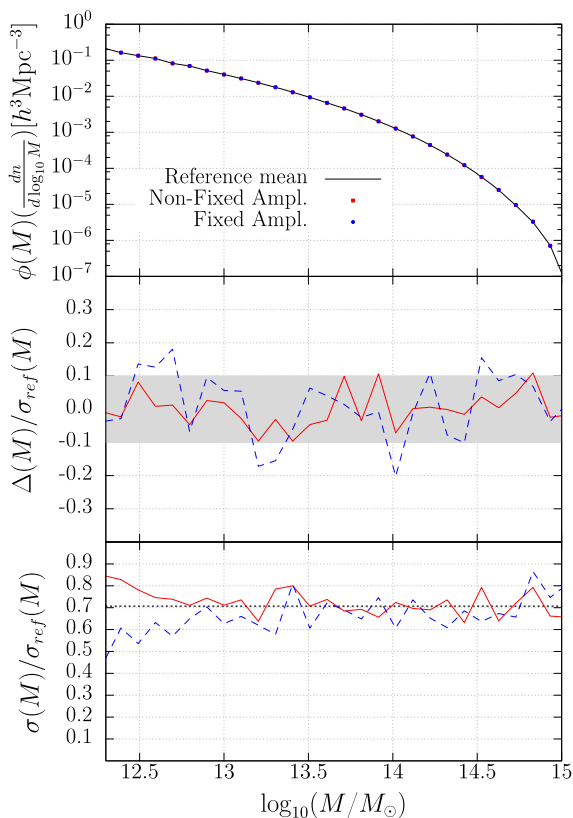


Figure 4. Halo mass functions with same conventions as in Fig. 1. We find no bias in the mean. We find a slight improvement in the variance below a mass of approximately $M < 10^{13} h^{-1} M_{\odot}$.

2.2.2 Dark matter haloes

Dark matter haloes show the same qualitative results as for the dark matter particles; quantitatively the suppression of variance is more modest. This is illustrated in Fig. 2, which compares halo clustering measurements, analogous to Fig. 1. From this, we conclude that the suppression of variance method is also not biased for dark matter haloes catalogues at any scale over the range considered. As in the case of dark matter clustering measurements, the improvement in the variance depends on the scale. In this case for the PK and CF at large scales, the ratios of the uncertainties (the bottom panels) can be as small as 0.2 or less, corresponding to an effective simulation volume of more than $25 (h^{-1} \text{Mpc})^3$. As in the dark matter particle case, we do not find improvement for the BK.

The halo population is a biased subset of all matter. Haloes’ masses and positions are sensitive to small-scale fluctuations in the initial density field, an effect often referred to as stochasticity. This stochasticity explains the difference in the results with respect to the performance of the dark matter distribution traced by particles.

The results for redshift-space halo clustering including monopoles and quadrupoles in configuration and Fourier space show a similar performance to the real-space measurements, as shown in Fig. 3. We use the same definition of the multipole expansion as in Chuang et al. (2017).

2.2.3 Halo mass function

As another relevant statistic, we investigate the halo mass function, shown for $z = 1$ in Fig. 4. We find no bias in the mean, and only a

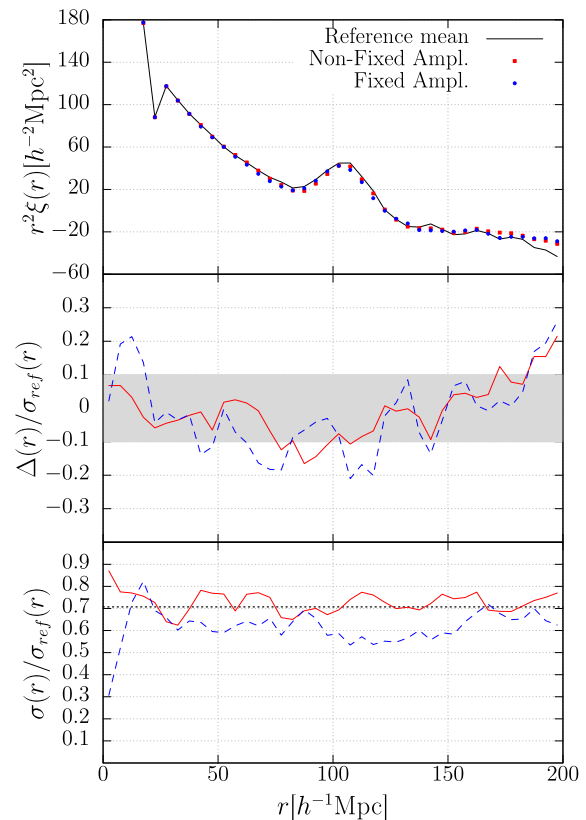


Figure 5. Void autocorrelation function selected with the radius cut $12 h^{-1} \text{Mpc}$ using the DIVE code, using the same conventions as in Fig. 1. The results show no bias and a very moderate improvement in the uncertainty.

slight improvement in the variance below a mass of approximately $M < 10^{13} h^{-1} M_{\odot}$. We further check this in the additional set of FASTPM boxes with smaller box size but higher resolution and confirm the improvement of the mass function in the lower mass bins. Further tests are shown at <http://www.unitsims.org>. In that Supporting Information, we also show that the suppression of variance is more effective (i) towards increasing redshifts, as structure formation becomes more linear; and (ii) for lower mass cuts, as the higher mass populations suffer more from stochasticity (see also Section 3 for another representation of these trends).

2.2.4 Void clustering

We now consider cosmic voids, focusing on the well-defined convention used in the void finder code DIVE (Zhao et al. 2016) that considers voids as empty spheres constrained by quartets of galaxies. This definition has proved useful to study the troughs of the density field, i.e. the clustering within cosmic voids, and to obtain improved measurements of the BAO signature (see Kitaura et al. 2016; Liang et al. 2016; Zhao et al. 2018). Cosmic voids in fact are measures of the higher order statistics of the galaxy distribution (see above-mentioned papers and references therein), and are therefore interesting to study the performance of SVMs. As expected from the BK result, the autocorrelation function of the voids, shown in Fig. 5, shows a very moderate improvement in the uncertainty and no bias. However, the cross-correlation functions between haloes and voids present significant improvements, shown in Fig. 6.

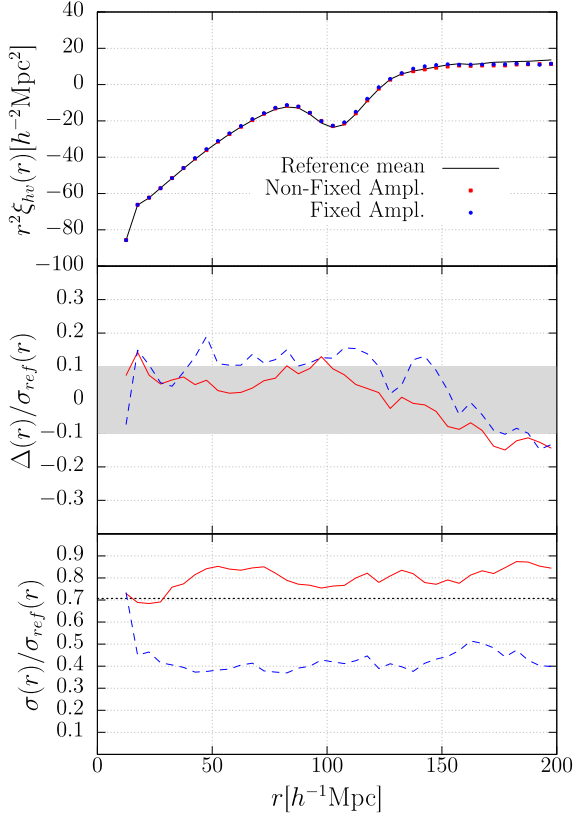


Figure 6. Same as Fig. 5, but for void–galaxy cross-correlation function. This shows significant improvement in the prediction, compared to the void autocorrelation function.

3 APPLICATION OF SVM FOR CLUSTERING ANALYSIS FROM GALAXY SURVEYS

We have demonstrated in the previous section that the SVM does not introduce any bias, and significantly reduces uncertainty in the two-point statistics. We now describe our first two pairs of high-resolution full N -body simulations aimed at the analysis of ELG and LRG data from DESI- and Euclid-like surveys.

3.1 Setup

We use the N -body code GADGET (Springel 2005), a full MPI parallel code that uses PM + Tree algorithms to compute the Newtonian forces between the dark matter particles by splitting the gravitational force into a long-range term (computed through the PM method) and a short-range term taken from the nearest neighbours, using a Tree method to categorize the particles according to their relative distances. This code makes use of the public software library FFTW for parallel Fast-Fourier transforms and the GNU Scientific Library (GSL). We are using a non-public version of the GADGET code, L-GADGET, that is highly optimized for large-volume simulations with a cubic domain decomposition and an efficient use of internal memory. This code has been extensively used to produce large-volume simulations with billions to hundreds of billions of particles, including the Multidark simulation suite (see <http://www.multidark.org>) and the Millennium series of simulations (including the Millennium XXL with more than 300 billion particles).

The paired initial conditions with fixed amplitude are generated using second-order Lagrangian perturbation theory with FASTPM (Feng et al. 2016). We use the same cosmology as the FASTPM simulations generated for this study (see Section 2). The box size is $1 h^{-1} \text{Gpc}$ and the simulation is started at $a \equiv 1/(1+z) = 0.01$ ($z = 99$). The number of particles is 4096^3 , giving a particle mass is $\sim 1.2 \times 10^9 h^{-1} M_{\odot}$. A slice of $500 \times 500 h^{-1} \text{Mpc}$ and $0.5 h^{-1} \text{Mpc}$ thickness from the density fields of a pair of GADGET simulations. We show one simulation in bright colour and one in dark colour. One can see that the overdensity regions (e.g. knots) in one simulation correspond to the underdensity regions (e.g. voids) in the other one. We use the halo finder code ROCKSTAR (Behroozi et al. 2013a) to identify haloes and compute their merging histories using the CONSISTENT TREES software (Behroozi et al. 2013b).

3.2 Results from full N -body simulations

Below we present tests based on two sets of pairs of full N -body simulations. The resolution of these Gpc scale simulations has been chosen to match the resolution of our small-volume PM simulations using the FASTPM code. The halo catalogues were generated using a minimum halo mass of $1.2 \times 10^{11} h^{-1} M_{\odot}$ (at this limit the mass function is quite complete, as shown in the rightmost panel of Fig. 8). This permits us to assess the improvement in the statistics from the FASTPM simulations with the $250 h^{-1} \text{Mpc}$ box size and a mesh of 1024^3 . We show the power spectrum, correlation function, and halo mass function measurements from our fixed-amplitude-paired GADGET N -body simulations in Fig. 8, which turn out to be remarkably smooth for the different redshift snapshots. We explore robust statistical measures in the next section to further assess the quality of the simulations.

3.2.1 Estimator quantifying the improvement in SVM

Thus far, we have shown the improvements at different scales for different clustering statistics. However, in a practical cosmological analysis (see e.g. Chuang et al. 2017), we use a specific scale range (e.g. $40 < r < 200 h^{-1} \text{Mpc}$ in configuration space or $0.02 < k < 0.2 h \text{Mpc}^{-1}$ in Fourier space), so the improvement should be determined by the whole range. To quantitatively assess the improvement, we adopt the Fisher information matrix formalism evaluating the improvement of the constraining power on a given cosmological parameter by performing the analysis within a certain scale range. In this approach, the uncertainty of a given cosmological parameter (θ) is defined as

$$\begin{aligned} \text{Var}(\theta) &= (F^{-1})_{\theta, \theta} \\ &= \left\langle \frac{-\partial^2 \ln P(\theta)}{\partial \theta^2} \right\rangle^{-1} \\ &\propto \left\langle -\frac{\partial^2 \left(\text{Tr}[\log C] + \sum_{ij} f_i \text{Tr} C_{ij}^{-1} f_j \right)}{\partial \theta^2} \right\rangle^{-1} \end{aligned} \quad (1)$$

where $P(\theta)$ is the posterior, and $\langle \dots \rangle$ denotes the expectation value. C is the covariance matrix of some measurement f (e.g. PK or CF), and i, j are the indices of the elements, i.e. $C_{ij} = \langle f_i f_j \rangle$ (Dodelson 2003). We have assumed a flat prior on the parameter θ and a Gaussian likelihood. For simplicity, we further assume that all the measured data points within the scale range of interest have the same sensitivity to the parameter θ , i.e. $\partial f_i / \partial \theta = \partial f_j / \partial \theta$, $\forall i, j$. With these assumptions, the uncertainty of the parameter θ can be related to the covariance matrix of the data vector via the following

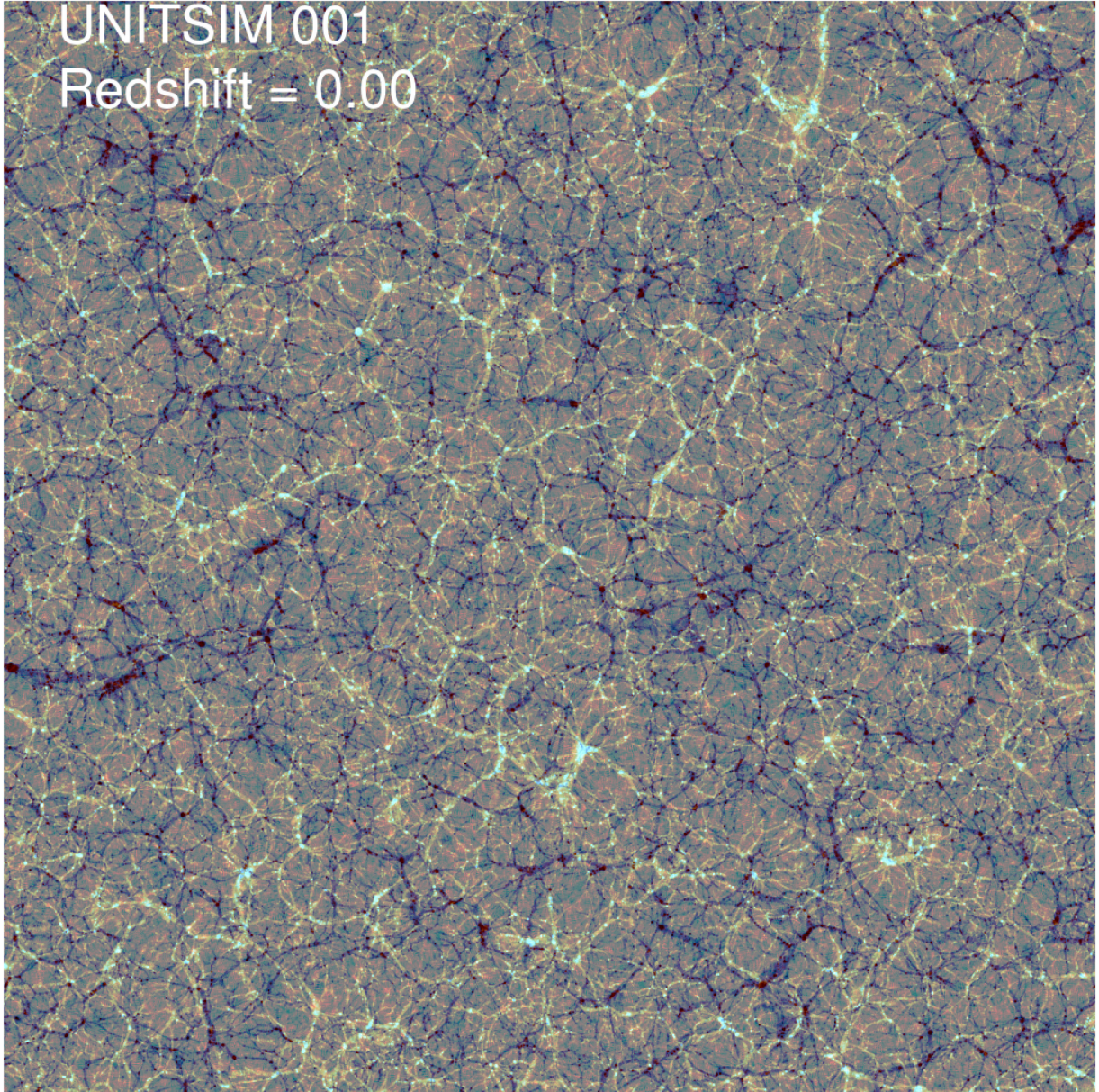


Figure 7. A slice of $500 \times 500 h^{-1} \text{Mpc}$ and $0.5 h^{-1} \text{Mpc}$ thickness from the density fields of a pair of GADGET simulations. We show one simulation in bright colour and one in dark colour. One can see that the overdensity regions (e.g. knots) in one simulation are the underdensity regions (e.g. voids) in the other one.

equation,

$$\text{Var}(\theta) = A \left(\sum_{i,j} (C^{-1})_{ij} \right)^{-1}, \quad (2)$$

where i, j go through all the data points within the scale range of interest, and A is assumed to be a constant involving terms $(\partial f / \partial \theta)^2$.

We then quantify the covariance matrix of the data vector from the simulation. We note that in order to perform the cosmological analysis, one has to account for two types of uncertainties. The first one is the theoretical uncertainty, represented by the theoretical covariance, C_{theo} , which is driven by the standard error on the statistic, f in the simulations used to validate the models. The second one is the observational uncertainties encoded in the covariance matrix C_{obs} . For a galaxy survey, for example, this would include sample variance on large scales, and stochasticity on small scales.

The total covariance matrix is given by the sum of the individual ones, i.e.

$$C = C_{\text{theo}} + C_{\text{obs}}. \quad (3)$$

The reasonable assumption here is that there are no cross-covariances between the two.

The theoretical covariance matrix, C_{theo} , can be calculated from the simulations used for validating the models with either the fixed amplitude or the regular N -body simulations. We will quantify the difference between these two choices below. We first estimate the observational covariance matrix, C_{obs} , by rescaling the covariance matrix from the regular simulations based on the expected survey volume. Consider an effective volume of $20 (h^{-1} \text{Gpc})^3$, roughly corresponding to that of the DESI and Euclid surveys. The covariance matrix including a pair of fixed-amplitude simulations can be

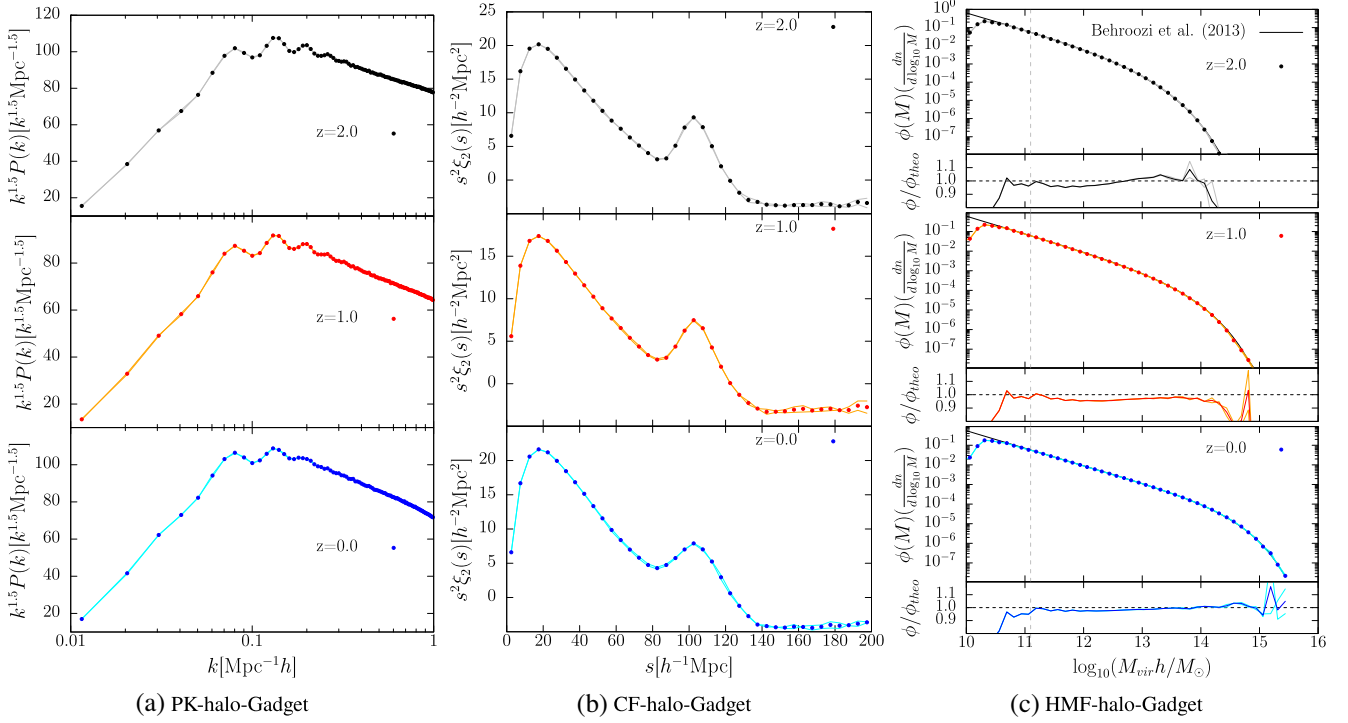


Figure 8. Power spectrum, correlation function, and halo mass function measurements from full N -body simulations with the SVM. SVM reduces the variances significantly so that the measurements are very smooth.

computed by

$$C = C_{\text{SVM}} + \frac{C_1}{V_{\text{EFFS}}}, \quad (4)$$

where C_1 is the covariance matrix of a single regular $1(h^{-1}\text{Gpc})^3$ box, V_{EFFS} is the effective survey volume ($20(h^{-1}\text{Gpc})^3$ in our study), and C_{SVM} is the covariance matrix of the SVM (paired fixed-amplitude simulation). Following equation (2), we compute the variance, Var_{SVM} .

Let us now answer the question: what is the size of the required standard simulation, that yields the equivalent variance of a pair of simulations with the SVM? Given a normal simulation with volume $V = (h^{-1}\text{Gpc})^3$, the total covariance matrix is given by

$$C_V = \frac{C_1}{V} + \frac{C_1}{V_{\text{EFFS}}}. \quad (5)$$

We now compute the variance Var_V based on equation (2). By solving $\text{Var}_{\text{SVM}} = \text{Var}_V$, we obtain the equivalent volume (V) that our paired fixed-amplitude simulations are representing. This is shown in Fig. 9; the equivalent volume vs. scale ranges used in the power spectrum and correlation function analysis are shown. Here, the maximum separation was fixed and the minimum separation was varied in the correlation function analysis; while the minimum k was fixed and the maximum k was varied in the power spectrum analysis.

We find that a pair of $1(h^{-1}\text{Gpc})^3$ boxes can potentially correspond to effective volumes of up to $100(h^{-1}\text{Gpc})^3$ considering haloes with lower masses. We also find that the equivalent volume is sensitive to the power spectrum, but not to the correlation function analysis. One might obtain very large effective volumes by ignoring the covariance matrix from observations, artificially driven by the uncertainty at large scales (e.g. small k). Thus, this additional

covariance matrix needs to be taken into account, as we do in our analysis.

Interestingly, the naive correspondence between $k \sim 0.35 h\text{Mpc}^{-1}$ and $r \sim 20 h^{-1}\text{Mpc}$ using $k = 2\pi/L$, yields completely different effective volumes: roughly 10 and $100(h^{-1}\text{Gpc})^3$, respectively (see lower panels in Fig. 9), thus emphasizing the difference in Fourier and configuration space analyses when a limited range in k or r is used.

In contrast to configuration space, Fourier space is more sensitive to large scales (low k s), which are already linear (e.g. $k \sim 0.2 h\text{Mpc}^{-1}$). Although fixing the amplitude of the power spectrum is crucial in reducing variance, as we showed in detail in Section 2.1, it does not remove variance induced by non-linear gravitational mode coupling. This is very apparent in Fourier space analysis. We can find analogous examples in the literature comparing the two-point statistics in Fourier and configuration space, such as (1) aliasing introduced by the gridding process of a set of point masses onto a mesh (Hockney & Eastwood 1988), in which a clouds-in-cells mass assignment scheme applied on dark matter particles in a cosmological simulation with cell resolutions of a few Mpc scales underestimates the true power spectrum down to $k \sim 0.2 h\text{Mpc}^{-1}$ (Jing 2005); (2) or in the clustering analysis of galaxies in redshift space, in which the virial motions (a.k.a. fingers-of-god, Jackson 1972) are present only below a certain scale (of ~ 20 Mpc), but are visible down to $k \sim 0.1 h\text{Mpc}^{-1}$ in Fourier space. This also indicates that pairing simulations with opposite phases and fixed amplitudes is not very effective in suppressing the variance at small scales, as we already saw in the three-point statistics analysis (see Section 2.1) and further improvements should be investigated.

We conclude from this analysis, that our two pairs of high-resolution N -body simulations with the SVM have an effective volume larger than 7 times that of the DESI or Euclid effective

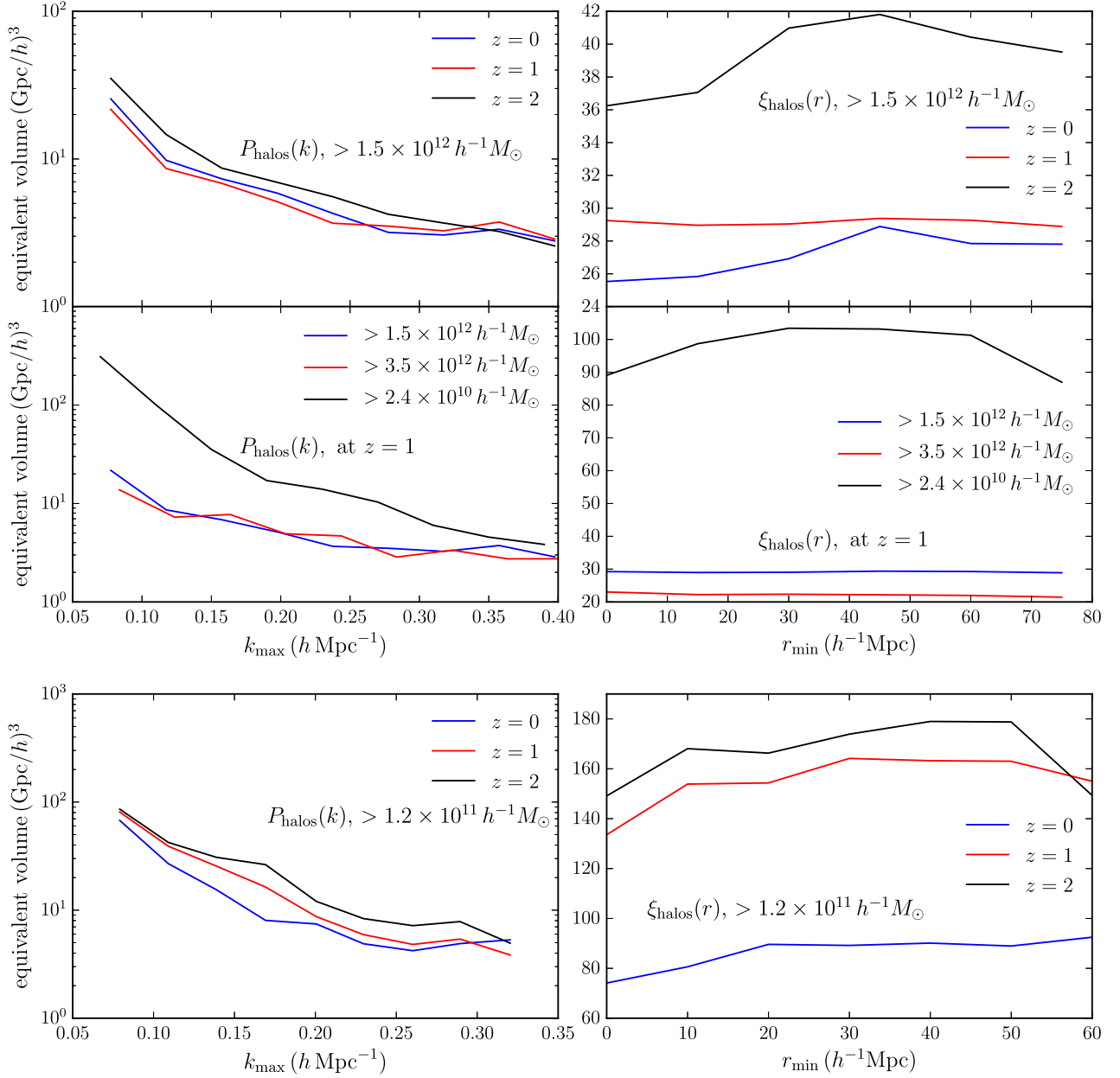


Figure 9. Equivalent volume study of the SVM for different redshifts and mass cuts. The upper four panels show the equivalent volumes of the catalogues from one pair of $1 h^{-1}\text{Gpc}$ boxes. The lower two panels show the equivalent volumes of the catalogues from two pairs of $1 h^{-1}\text{Gpc}$ boxes based on our high-resolution full N -body simulations. The results shown in the left-hand panels are in Fourier space (varying k_{\max} with fixed $k_{\min} = 0$) and those shown in the right-hand panels are in configuration space (varying r_{\min} with fixed $r_{\max} = 120$).

survey volumes when the analysis is performed in configuration space. We are currently preparing larger sets of SVM N -body simulations to ensure that this accuracy is also achieved in Fourier space.

4 SUMMARY AND CONCLUSIONS

In this work, we have presented the UNIT N -body cosmological simulation project. We present four simulations (two pairs) along with this paper. The box size is $1 h^{-1}\text{Gpc}$ and the number of particles of each box is 4096^3 , resulting in a particle mass of $\sim 1.2 \times 10^9 h^{-1}M_{\odot}$. We have made their corresponding data products

publicly available through the website <http://www.unitsims.org>, including

- (i) dark matter particles
- (ii) density fields
- (iii) halo catalogues
- (iv) dark matter clustering statistics
- (v) halo clustering statistics (real and redshift space)
- (vi) void clustering statistics

We show that the effective volume of our simulation suite is equivalent to $150 (h^{-1}\text{Mpc})^3$ (7 times of the effective survey volume of DESI or Euclid), using a mass resolution of $\sim 1.2 \times 10^9 h^{-1}M_{\odot}$,

enough to resolve the host haloes of the galaxy sample observed by DESI (ELGs) or *Euclid* ($H\alpha$ galaxies).

Our work relies on the SVM approach recently introduced by Angulo & Pontzen (2016). In order to demonstrate the practicality of the SVM for large-scale structure analyses, we investigate a number of issues including potential biases introduced by the method, and characterize the improvement in the theoretical uncertainty and effective volume in a number of different regimes.

We have performed a large number (800) of accurate PM simulations using the FASTPM code, and have demonstrated that

- (i) no significant biases are introduced that would affect BAO or RSD analysis;
- (ii) the error in two-point statistics in configuration space is significantly reduced;
- (iii) the error in two-point statistics in Fourier space in moderately reduced; and
- (iv) no significant improvement is found for the three-point statistics on scales relevant to BAO and RSD analysis.

We also performed an analysis including RSDs, and 3D halo distributions beyond the halo mass function. We found that the improvements in galaxy bispectrum and void autocorrelation function using SVM are small. However, the improvement in the void–galaxy cross-correlation is significant; this indicates that the fixed-amplitude method should also be useful for void studies.

We introduced a parameter for quantifying the improvement of the SVMs, and show that these simulations are equivalent to a typical simulation with volume of $100 (h^{-1} \text{Gpc})^3$. The exact number depends on the analysis method considered (e.g., power spectrum or correlation function analysis), redshift, scale range, and the galaxy sample used.

With current state-of-the-art techniques we found that for a galaxy survey with effective volume of $20 (h^{-1} \text{Gpc})^3$ at $z = 1$, the reduction in variance resulting from the SVM is about a factor of 40 using two-point correlation function analysis. This means that our two pairs of simulations with full N -body calculations with volumes of $(1h^{-1} \text{Gpc})^3$ and 4096^3 particles lead to the same variance as ~ 150 of such simulations. This provides optimal reference clustering measurements to validate theoretical models in configuration space. The covariance matrices for the clustering measurements using SVM can be estimated based on the approximate methods presented in this paper, since they are very different from the typical Gaussian statistics (see Angulo & Pontzen 2016). This motivates future work to compute larger sets of N -body simulations using SVM. We are pursuing this, along with further analyses to investigate mode-coupling effects from larger scales and ways of correcting them.

In the spirit of sharing scientific results with the community, we have made the full N -body simulations in addition to the FASTPM products produced in this work publicly available through the website <http://www.unitsims.org>. We hope that these data products will enable a number of studies to further unveil the nature of dark energy and structure formation with galaxy surveys.

ACKNOWLEDGEMENTS

We thank Raul E. Angulo, Elena Massara, and Andrew Pontzen for useful discussions, and Volker Springel for providing the L-GADGET code. We acknowledge PRACE for awarding us access to the MareNostrum Supercomputer hosted by the Barcelona Supercomputing Center, Spain, under project grant number 2016163937. GY and AK would like to thank MINECO/FEDER (Spain) for partial

financial support under project grant AYA2015-63810P. AK further thanks the Spanish Red Consolider MultiDark FPA2017-90566-REDC. Francisco-Shu Kitaura (FSK) thanks support from the grants RYC2015-18693, SEV-2015-0548, and AYA2017-89891-P. CZ thanks support under National Natural Science Foundation of China (NSFC) grant number 11673025. SA is supported by the European Research Council through the COSFORM Research Grant (number 670193). This research received support from the U.S. Department of Energy under contract number DE-AC02-76SF00515, Additional computing support was provided by SLAC and by National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under contract number DE-AC02-05CH11231.

REFERENCES

- Abell P. A. et al., 2009, preprint ([arXiv:0912.0201](https://arxiv.org/abs/0912.0201))
- Angulo R. E., Pontzen A., 2016, *MNRAS*, 462, L1
- Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, *MNRAS*, 426, 2046
- Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013a, *ApJ*, 762, 109
- Behroozi P. S., Wechsler R. H., Busha M. T., Klypin A. A., Primack J. R., 2013b, *ApJ*, 763, 18
- Benitez N. et al., 2014, preprint ([arXiv:1403.5237](https://arxiv.org/abs/1403.5237))
- Chuang C.-H. et al., 2017, *MNRAS*, 471, 2370
- Cochrane R. K., Best P. N., Sobral D., Smail I., Wake D. A., Stott J. P., Geach J. E., 2017, *MNRAS*, 469, 2913
- de Jong R. S. et al., 2012, Proc. SPIE, 8446, 84460T
- Dodelson S., 2003, *Modern Cosmology*. Academic Press, Amsterdam
- Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, *MNRAS*, 463, 2273
- Fosalba P., Crocce M., Gaztañaga E., Castander F. J., 2015, *MNRAS*, 448, 2987
- González-Pérez V. et al., 2018, *MNRAS*, 474, 4024
- Habib S. et al., 2016, *New Astron.*, 42, 49
- Hand N., Feng Y., Beutler F., Li Y., Modi C., Seljak U., Slepian Z., 2017, *AJ*, 156, 160,
- Hill G., et al., 2008, in Kodama T., Yamada T., Aoki K. eds, ASP Conf. Ser. Vol. 399, *Panoramic Views of Galaxy Formation and Evolution*. Astron. Soc. Pac., San Francisco, p. 115
- Hockney R. W., Eastwood J. W., 1988, *Computer Simulation using Particles*. Hilger, Bristol
- Jackson J. C., 1972, *MNRAS*, 156, 1P
- Jing Y. P., 2005, *ApJ*, 620, 559
- Kitaura F.-S. et al., 2016, *Phys. Rev. Lett.*, 116, 171301
- Klypin A., Prada F., 2018, preprint ([arXiv:1809.03637](https://arxiv.org/abs/1809.03637))
- Klypin A., Yepes G., Gottlober S., Prada F., Hess S., 2016, *MNRAS*, 457, 4340
- Laureijs R. et al., 2011, preprint ([arXiv:1110.3193](https://arxiv.org/abs/1110.3193))
- Levi M. et al., 2013, preprint ([arXiv:1308.0847](https://arxiv.org/abs/1308.0847))
- Liang Y., Zhao C., Chuang C.-H., Kitaura F.-S., Tao C., 2016, *MNRAS*, 459, 4020
- Planck Collaboration XIII et al., 2016, *A&A*, 594, A13
- Pontzen A., Slosar A., Roth N., Peiris H. V., 2016, *Phys. Rev.*, D93, 103519
- Potter D., Stadel J., Teyssier R., 2016, *Comput. Astrophys. Cosmol.*, 4, 2
- Schlegel D. et al., 2011, preprint ([arXiv:1106.1706](https://arxiv.org/abs/1106.1706))
- Skillman S. W., Warren M. S., Turk M. J., Wechsler R. H., Holz D. E., Sutter P. M., 2014
- Spergel D. et al., 2013, preprint ([arXiv:1305.5422](https://arxiv.org/abs/1305.5422))
- Springel V., 2005, *MNRAS*, 364, 1105
- Takada M. et al., 2014, *PASJ*, 66, R1
- Tassev S., Zaldarriaga M., Eisenstein D., 2013, *J. Cosmol. Astropart. Phys.*, 1306, 036
- Villaescusa-Navarro F. et al., 2018, *ApJ*, 867, 137
- White M., 2015, *MNRAS*, 450, 3822

Zel'dovich Y. B., 1970, *A&A*, 5, 84

Zhao C., Tao C., Liang Y., Kitaura F.-S., Chuang C.-H., 2016, *MNRAS*, 459, 2670

Zhao C., Chuang C.-H., Liang Y., Kitaura F.-S., Vargas-Magaña M., Tao C., Pellejero-Ibanez M., Yepes G., 2018, preprint ([arXiv:1802.03990](https://arxiv.org/abs/1802.03990))

¹*Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA*

²*Departamento de Física Teórica, Módulo 8, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain*

³*CIAFF, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain*

⁴*Instituto de Astrofísica de Canarias (IAC), C/Vía Láctea, s/n, E-38200 La Laguna, Tenerife, Spain*

⁵*Departamento Astrofísica, Universidad de La Laguna (ULL), E-38206 La Laguna, Tenerife, Spain*

⁶*Department of Physics, Berkeley Center for Cosmological Physics, University of California Berkeley, Berkeley CA 94720, USA*

⁷*Dipartimento di Fisica e Astronomia, Alma Mater Studiorum Università di Bologna, via Gobetti 93/2, I-40129 Bologna, Italy*

⁸*INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, via Gobetti 93/3, I-40129 Bologna, Italy*

⁹*Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*

¹⁰*SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*

¹¹*Laboratoire d'Astrophysique, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, CH-1290 Versoix, Switzerland*

¹²*National Astronomical Observatories, Chinese Academy of Science, Beijing 100012, P.R.China*

¹³*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*

¹⁴*Dipartimento di Fisica e Scienze della Terra, Università degli Studi di Ferrara, via Saragat 1, I-44122 Ferrara, Italy*

¹⁵*INFN – Sezione di Bologna, viale Berti Pichat 6/2, I-40127 Bologna, Italy*

¹⁶*ICRAR, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.