Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **433,** 2389–2402 (2013)
Advance Access publication 2013 June 19

doi:10.1093/mnras/stt907

An accurate tool for the fast generation of dark matter halo catalogues

P. Monaco,[1,2]⋆ E. Sefusatti,[3,4] S. Borgani,[1,2,5] M. Crocce,[6] P. Fosalba,[6] R. K. Sheth[4,7] and T. Theuns[8,9]

[1]*Dipartimento di Fisica - Sezione di Astronomia, Università di Trieste, via Tiepolo 11, I-34131 Trieste, Italy*
[2]*INAF, Osservatorio Astronomico di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy*
[3]*INAF, Osservatorio Astronomico di Brera, Via Bianchi 46, I-23807 Merate (LC), Italy*
[4]*The Abdus Salam International Center for Theoretical Physics, Strada costiera, 11, I-34151 Trieste, Italy*
[5]*INFN, Istituto Nazionale di Fisica Nucleare, Trieste, Italy*
[6]*Institut de Ciències de l'Espai, IEEC-CSIC, Campus UAB, Facultat de Ciències, Torre C5 par-2, E-08193 Barcelona, Spain*
[7]*Center for Particle Cosmology, University of Pennsylvania, 209 S. 33rd St., Philadelphia, PA 19104, USA*
[8]*Institute for Computational Cosmology, Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK*
[9]*Department of Physics, University of Antwerp, Campus Groenenborger, Groenenborgerlaan 171, B-2020 Antwerp, Belgium*

Accepted 2013 May 21. Received 2013 May 6; in original form 2013 March 27

ABSTRACT

We present a new parallel implementation of the *PINpointing Orbit Crossing-Collapsed HIerarchical Objects* (PINOCCHIO) algorithm, a quick tool, based on Lagrangian Perturbation Theory, for the hierarchical build-up of dark matter (DM) haloes in cosmological volumes. To assess its ability to predict halo correlations on large scales, we compare its results with those of an N-body simulation of a 3 h^{-1} Gpc box sampled with 2048^3 particles taken from the MICE suite, matching the same seeds for the initial conditions. Thanks to the Fastest Fourier Transforms in the West (FFTW) libraries and to the relatively simple design, the code shows very good scaling properties. The CPU time required by PINOCCHIO is a tiny fraction ($\sim 1/2000$) of that required by the MICE simulation. Varying some of PINOCCHIO numerical parameters allows one to produce a universal mass function that lies in the range allowed by published fits, although it underestimates the MICE mass function of Friends-of-Friends (FoF) haloes in the high-mass tail. We compare the matter–halo and the halo–halo power spectra with those of the MICE simulation and find that these two-point statistics are well recovered on large scales. In particular, when catalogues are matched in number density, agreement within 10 per cent is achieved for the halo power spectrum. At scales $k > 0.1\,h\,\mathrm{Mpc}^{-1}$, the inaccuracy of the Zel'dovich approximation in locating halo positions causes an underestimate of the power spectrum that can be modelled as a Gaussian factor with a damping scale of $d = 3\,h^{-1}$ Mpc at $z = 0$, decreasing at higher redshift. Finally, a remarkable match is obtained for the reduced halo bispectrum, showing a good description of non-linear halo bias. Our results demonstrate the potential of PINOCCHIO as an accurate and flexible tool for generating large ensembles of mock galaxy surveys, with interesting applications for the analysis of large galaxy redshift surveys.

Key words: surveys – cosmology: theory – dark matter.

1 INTRODUCTION

Recent measurements of the cosmic microwave background radiation (e.g. Bennett et al. 2012; Story et al. 2012; Ade et al. 2013a; Das et al. 2013) have yielded accurate measurements of the geometry of the Universe and the statistics of the linear, large-scale perturbations visible at redshift $\sim$1100, the epoch of recombination. Thanks to these experiments, uncertainties on the main cosmological parameters have been beaten down to the per cent level (Hinshaw et al. 2012; Sievers et al. 2013; Ade et al. 2013b). The advantage of studying the Universe before perturbations started to evolve beyond their linear regime is, however, counterbalanced by the limit of observing it at a single cosmic time. Performing measurements at lower redshifts is then desirable because the late-time growth of perturbations in a flat Universe is slowed down by the dominance of the elusive dark energy, so measurements of the growth of perturbations in the redshift range from $z = 0$ to $z \sim$1–2 would lead to tight constraints on the equation of state of dark energy and possibly provide evidence of physics beyond a simple cosmological constant.

⋆E-mail: monaco@oats.inaf.it

At the same time, accurate measurements of density (through the galaxy power spectrum and higher moments), potential (through galaxy weak lensing) and high-density peaks (through the mass function of galaxy clusters) can characterize the growth of perturbations to a level of detail sufficient to distinguish the predictions of General Relativity from those of some non-standard gravity models (e.g. Amendola et al. 2012, and references therein), constrain other quantities like neutrino masses (Lahav et al. 2010; Carbone et al. 2012; Costanzi Alunno Cerbolini et al. 2013) and the degree of non-Gaussianity in the primordial perturbations (Desjacques & Seljak 2010; Liguori et al. 2010).

For this reason many ongoing and future observational campaigns, such as DES,[1] Euclid,[2] PanSTARRS,[3] LSST[4] or SKA,[5] are surveying or will survey large parts of the sky to a depth that will reach $z \sim 1$ or beyond. Taking the future Euclid mission (Laureijs et al. 2011) as an example, with $\sim$15 000 deg^2 of the sky surveyed in the $0.5 < z < 2$ redshift range, uncertainties in the estimates of physical parameters from observable quantities will be significantly affected by systematics connected to sample variance and to the bias with which galaxies trace mass. This bias is ultimately determined by the complex physics of baryons and will generally depend on redshift and on the specific sample selection. An accurate assessment of these theoretical systematics requires the use of numerical simulations to generate the non-linear distribution of dark matter (hereafter DM) and models to populate the resulting DM haloes with mock galaxies. Even assuming that large-scale structures can be accurately described by the gravitational evolution of pure collisionless DM and that the generation of galaxies in DM haloes is under control, the requirements for mock catalogues (typically of Gpc^3 volumes and mass resolution below $10^{10}\,h^{-1}\,\mathrm{M}_\odot$ for on-going and future experiments) are quite demanding. Such large simulations need more than 10^{10} particles, on-the-fly group finders and nearly 100 outputs to generate merger trees and past-light-cones. In this case the hardware requirements in terms of memory and disc storage raise more problems than the computing time needed to carry out a single simulation. The problem becomes untreatable when a very large number of realizations (of the order of 1000 or more) are needed to estimate the covariance matrix of the galaxy power spectrum (e.g. Manera et al. 2013). This is even more so for higher-order statistics (Sefusatti et al. 2006).

This has prompted a number of recent works which use approximations to the mildly non-linear evolution of perturbations (e.g. Kitaura & Heß 2012; Manera et al. 2013; Tassev, Zaldarriaga & Eisenstein 2013). Several of these are based on Lagrangian Perturbation Theory (hereafter LPT; Moutarde et al. 1991; Buchert & Ehlers 1993; Catelan 1995), a perturbative solution of a set of equations for the displacements of mass elements from their initial position. With LPT it is possible to accurately recover the large-scale density field of matter, but a fair reconstruction of DM haloes requires a different approach.

A decade ago, Monaco, Theuns & Taffoni (2002, hereafter Paper I) presented a code, called *PINpointing Orbit Crossing-Collapsed HIerarchical Objects* (PINOCCHIO), which was able to generate, with very limited computing resources, a catalogue of DM haloes with known mass, position and velocity from a realization of a Gaussian density contrast field on a cubic grid, i.e. the same initial conditions that are used by most simulations. In that paper and in Taffoni, Monaco & Theuns (2002) the code was thoroughly tested against two simulations that were state-of-the-art at that time. It was shown not only to reproduce (to within $\sim$5–20 per cent) statistics such as the mass function and two-point correlation function of DM haloes, but also to generate DM haloes that agreed with the simulated ones at the object-by-object level. The code was tested by other groups, who confirmed its accuracy in reproducing merger histories (Li et al. 2007; Zhao et al. 2009) and velocities of DM haloes (Heisenberg, Schäfer & Bartelmann 2011). It was also used by several groups to study, e.g., DM halo density profiles (Lu et al. 2006) and concentrations (Zhao et al. 2003), the Sunyaev–Zeldovich effect in clusters (Peel, Battye & Kay 2009), the properties of X-ray-selected clusters (Pierre et al. 2011), galaxy clustering (Zheng, Coil & Zehavi 2007), the formation of the first stars (Schneider et al. 2006) and of supermassive black holes (Jahnke & Macciò 2011).

In this paper we present a new version of the PINOCCHIO code, designed to perform large runs (in our tests we use up to 2160^3 particles) on hundreds of computing cores of a parallel computer. With respect to Paper I, this version implements the same algorithm but is fully parallel. We test the accuracy of PINOCCHIO on a much larger range of masses and scales by comparing its results with a large simulation kindly made available by the MICE collaboration (Fosalba et al. 2008; Crocce et al. 2010). We address clustering in Fourier space, and demonstrate that the accuracy with which power spectrum and bispectrum of DM haloes is reconstructed can be easily pushed below the $\sim$10 per cent level. We also show CPU time requirements and scaling properties to demonstrate that this code can easily scale up to hundreds of cores, and identify the improvements that are needed to run it on thousands of cores. As an example, a 2160^3 realization requires 38 min of wall-clock time on 324 2.4 GHz cores of a linux machine, for a cost of 206 CPU hours, so that running 10 000 such realizations would require just 2×10^6 CPU-hours. This code provides fine time sampling of merger histories, necessary to reconstruct halo positions along the past-light cone (Manera et al. 2013; Merson et al. 2013) and to run semi-analytic models of galaxy formation (Benson et al. 2012). This makes it invaluable for addressing a range of problems such as sample variance, the estimation of covariance matrices or sampling of parameter space, where many very large realizations are needed.

The paper is organized as follows. Section 2 presents the algorithm, its latest parallel implementation and its performance. Section 3 presents the simulations used for a comparison. In Section 4 we quantify the accuracy with which power spectrum and bispectrum of DM haloes are recovered. Finally, Section 5 discusses the results and the main conclusions. The code is available under GNU/GPL license on http://adlibitum.oats.inaf.it/monaco/Homepage/Pinocchio/index.html.

2 PINOCCHIO

The code is fully described in Paper I, so here we only give a brief account of how it works.

2.1 The algorithm

The algorithm behind the PINOCCHIO code has roots in the extended Press & Schechter approach (Bond et al. 1991) and in its extension to non-spherical collapse by Monaco (1995) and Monaco (1997), but it does not use the Fokker–Planck approach based on sharp k-space filtering. The calculation starts from the generation of a

[1] http://www.darkenergysurvey.org/
[2] http://www.euclid-ec.org/
[3] http://pan-starrs.ifa.hawaii.edu/public/science-goals/galaxies-cosmology.html
[4] http://www.lsst.org/lsst/science
[5] http://www.skatelescope.org/

linear density field on a regular grid, as done when generating the initial conditions of an N-body simulation, and is divided in two parts: (i) the computation of collapse times of individual particles, performed by smoothing the density field on several scales and using an ellipsoidal model based on LPT to compute individual times of collapse; (ii) the fragmentation of the collapsed medium into distinct objects, performed with an algorithm that mimics the hierarchical build-up of DM haloes.

2.1.1 *Collapse times*

We start from a realization of a Gaussian field on a cubic grid[6] of N^3 vertices, assumed to have a physical size L. This Gaussian field is assumed to represent a linear density contrast field, defined as the density contrast at a very early time t_i, linearly extrapolated to the present:

$$\delta_l(\boldsymbol{q}) = \frac{\delta(\boldsymbol{q}, t_i)}{D(t_i)}\,, \tag{1}$$

where $\boldsymbol{q}$ is the *Lagrangian* coordinate of the mass element, i.e. its initial position at $t = 0$, and the growing mode $D(t)$ is normalized to unity at $z = 0$. The power spectrum of δ_l is given by the cosmological model and $\langle\delta_l^2\rangle = \sigma_8^2$ when the field is top-hat smoothed on a scale of 8 Mpc h^{-1}. Following the EPS approach, the density field is smoothed on a set of smoothing radii R. This is done with a Gaussian filter so the resulting trajectories are not random walks but are highly correlated in σ^2. Smoothing radii are chosen so that the corresponding mass variances are logarithmically spaced in intervals of 0.15 dex; typically from 10 to 20 smoothing radii are needed to sample the trajectories.

As described in Monaco (1995), at early times the evolution of a mass element can be described as the evolution of an ellipsoid, whose principal axes are given by the deformation tensor, i.e. the Hessian of the (peculiar) gravitational potential. This is true at least until the ellipsoid collapses on its shortest axis. The dependence of ellipsoid evolution on the background cosmology can be approximately factorized out by using the linear growing mode as a time coordinate. In this case a very good approximation of this evolution can be obtained by using third-order LPT (Monaco 1997). The argument can be reversed, so that ellipsoidal collapse can be considered as a truncation of LPT where all non-locality is given by the deformation tensor. This allows one to treat the collapse of the first axis as an event of 'orbit crossing', after which the LPT approach breaks down. LPT is slow to converge in the case of a sphere, and this leads to an overestimate of collapse times for spherical peaks; to fix it Monaco (1997) found an empirical correction that reproduces the numerical solution of ellipsoidal collapse for quasi-spherical cases.

For each smoothing radius the code performs a series of Fast Fourier Transforms (FFTs) to compute the deformation tensor. Then, for each grid point, and using the ellipsoidal truncation of third-order LPT and its correction for quasi-spherical cases, the code computes the time $t_{\rm coll}(\boldsymbol{q})$ at which the mass element at $\boldsymbol{q}$ is expected to collapse. Using the growing mode as a time coordinate, the relevant quantity is the inverse of the collapse time of each mass element $\boldsymbol{q}$:

$$F(\boldsymbol{q}) = 1/D(t_{\rm coll}(\boldsymbol{q}))\,. \tag{2}$$

In the EPS language, for each grid point $\boldsymbol{q}$ we construct a trajectory in the plane defined by mass variance of the smoothed field $\sigma^2(R)$ and inverse collapse time $F(\boldsymbol{q}; R)$. If we used spherical collapse we would have $F = \delta/\delta_c$, so the *absorbing barrier* at δ_c, the linear density contrast at which collapse is expected to take place, is replaced by a barrier placed at the inverse of the time (the growing mode) at which DM haloes are requested. When a mass element is predicted to collapse at the smoothing radius R, it is interpreted as belonging to a halo of mass at least $M = 4\pi R^3/3\,\bar{\rho}$ ($\bar{\rho}$ being the average matter density). (The absorbing barrier construction prevents the same mass element from being assigned to haloes with mass smaller than M.) In the same spirit, for each grid point $\boldsymbol{q}$ the code records the highest value of F along the trajectory, called $F_{\rm max}$, the associated smoothing scale $R_{\rm max}$, and the velocity $v_{\rm max}$ at that position when smoothed on scale $R_{\rm max}$. $F_{\rm max}$ is interpreted as the time at which, given the mass resolution of the realization, the grid point is expected to collapse, and $v_{\rm max}$ is computed from the first derivative of the peculiar potential each time the $F_{\rm max}$ value is updated.

2.1.2 *Fragmentation*

The first part of the algorithm provides, for each grid point, an inverse collapse time $F_{\rm max}$ and a velocity $v_{\rm max}$, plus the smoothing radius $R_{\rm max}$ at which these have been computed. With these it is possible to predict, at any time, which regions of Lagrangian space have gone into orbit crossing collapse. The fragmentation of the collapsed medium into distinct DM haloes is done with a code that mimics the hierarchical clustering of DM haloes (see also the description in Heisenberg et al. 2011).

It is convenient to describe grid points as 'particles' in the following. One thing worth stressing is that orbit crossing collapse does not imply that the particle belongs to a DM halo, because the filamentary network lying outside the haloes may have undergone such a collapse and yet be far from a fully relaxed state; the code makes a distinction between collapsed particles in haloes and those in the filamentary network. Particles are first sorted in descending $F_{\rm max}$, and considered in this (effectively chronological) order. For each collapsing particle the six neighbouring particles are considered, and the different haloes to which the neighbouring particles belong are counted. The following cases are possible.

(i) All six neighbours have not yet collapsed. The particle is then a peak of the $F_{\rm max}$ field and is treated as a new DM halo with one particle.

(ii) The particle touches only one halo. To decide whether the particle is to be accreted on it, both the particle and the halo are displaced using the Zel'dovich approximation (ZA; first-order LPT). If d is their distance after displacement, accretion takes place if the particle gets 'near enough' to the group:

$$d < f_a \times R + f_{ra} + f_s(R\sigma(R))^{1.7}, \tag{3}$$

where $R = \sqrt{N_{\rm h}} \times 3/4\pi$ is the halo Lagrangian radius in grid units ($N_{\rm h}$ being the number of particles belonging to it), f_a, f_{ra} and f_s are free parameters and the factor $(R\sigma(R))^{1.7}$ [with $\sigma(R)$ computed at the collapse time], discussed in appendix A of Paper I, is a correction for the increasing inaccuracy of Zel'dovich displacements as the density fields become more non-linear. If the particle does not accrete on to any halo, it is tagged as a 'filament' particle. After each accretion event all neighbouring particles that have been previously tagged as filaments are accreted on to the halo as well.

(iii) The particle touches more than one halo. First the code checks whether the particle should be accreted on to one halo (the one with minimal value of d/R). Then it checks each pair of haloes to determine whether they should be merged together. This happens

[6] The code can run on non-cubic grids as well.

Table 1. Adopted values of the best-fitting parameters. The right-hand column gives the effect that a change in that parameter has on the mass function.

Parameter	Eq.	Value	Effect on mass function
f_a	3	0.285	Normalization and slope
f_{ra}	3	0.180	Normalization
f_s	3	0.060	Dependence on mass resolution and z
f_m	4	0.350	Slope
f_{rm}	4	0.700	Abundance of poorly sampled haloes

if they get 'near enough' after Zel'dovich displacements have been applied:

$$d < f_m \times \max(R_1, R_2) + f_{ma} \, . \tag{4}$$

In case the particle was not supposed to accrete on to any halo, accretion is checked again after the merger(s).

(iv) The particle touches only filament particles. Then it is tagged as filament as well.

This fragmentation code allows a very accurate time sampling of the merger trees, because haloes are updated each time a collapsing particle touches them. The full catalogue of DM haloes is output each time it is requested, with masses, centres of mass in the Lagrangian space, displacements obtained with the ZA and peculiar velocities. Merger histories are output only at the final time, giving a complete time sampling by reporting the masses of merging haloes at each merger. The values of the free parameters are chosen by fitting to the desired halo mass function. Table 1 lists the values used in this paper, and the effect that each parameter has on the mass function.

2.2 The code

The original scalar code (Version 1) was written in FORTRAN 77 and designed to work on a simple PC. It allowed us to perform runs of 256^3 particles on a 450 MHz PentiumIII machine with 512 Mbyte of RAM in nearly 6 h, a remarkable achievement that allowed us to obtain reasonable statistics of merger histories with no access to a supercomputer. Because memory is the limiting factor in this case, the code has an out-of-core design: it keeps in memory only one component of the derivatives of the potential at a time, while the other components are saved on the disc. The most time- and memory-consuming part is the computation of collapse times; fragmentation takes less than 10 per cent of time.

In 2005, P. Monaco and T. Theuns wrote a parallel (MPI[7]) version of PINOCCHIO (Version 2), that was publicized among interested researchers and used in several of the papers mentioned in the Introduction. It is written in FORTRAN 90 and uses the Fastest Fourier Transforms in the West (FFTW) package (Frigo & Johnson 2012) to compute FFTs. While parallelizing the computation of $F_{\rm max}$ is straightforward (FFTW takes care of most communications), the fragmentation code was parallelized rather inefficiently, with one task performing the fragmentation and other tasks acting as storage; fragmentation is so quick that even this parallelization gives reasonable running times. Memory requirements were still minimized with an out-of-core strategy. This code is suitable to run on tens of cores, and requires fast access to the disc; when the number of cores increases, reading and writing on the disc become the limiting factor.

[7] Message Passing Interface.

The version we use in this paper (Version 3) has been designed to run on hundreds if not thousands of cores of a massively parallel super-computer. The two separate codes have been merged and no out-of-core strategy is adopted, so the amount of needed memory rises by a factor of 3 with respect to the previous version. The computation of collapse times is performed as in version 2. Fragmentation is performed by dividing the box into sub-volumes and distributing one sub-volume to each MPI task. The tasks do not communicate during this process, and each sub-volume needs to extend the calculation to a boundary layer, where reconstruction is affected by boundaries. From our tests and for a standard cosmology, the reconstruction of the largest objects is convergent when the boundary layer is larger than about 30 Mpc. This strategy minimizes the number of communications among tasks, and the boundary layer requires an overhead that is typically of the order of some tens per cent for large cosmological boxes. For small boxes at very high resolution this overhead would become dominant, in which case the serial code of Version 1 (on a large shared-memory machine) or the parallel code of Version 2 would be preferable.

Because FFTW distributes memory to tasks in planes, while the fragmentation code works with sub-boxes, a communication round is needed between the two codes to redistribute $F_{\rm max}$ and velocities. In the version we use here we have implemented a naive distribution scheme where tasks always communicate in pairs.

To generate the initial linear density field in the Fourier space, we have merged PINOCCHIO with a part of the code taken from N-GENIC by V. Springel.[8] Besides a few technical improvements with respect to the original PINOCCHIO code, this has the advantage to be able to faithfully reproduce a simulation run from initial conditions generated with N-GENIC, or with the second-order LPT (hereafter 2LPT) version by R. Scoccimarro[9] (Crocce, Pueblas & Scoccimarro 2006), just from the knowledge of the assumed cosmology and the random number seed.

The code has also been extended to consider a wider range of cosmologies including a generic, redshift-dependent equation of state of the quintessence, but the computation of collapse times based on ellipsoidal collapse still relies on the assumption that the dependence on cosmology is factorized out of dynamical evolution when the growing mode $D(t)$ is used as a clock, an approximation that should be tested before using the code for more general cosmologies. Displacements of groups from their final position are still computed with the ZA.

2.3 Performance and scaling

To test its performance and its strong and weak scaling properties, we ran the code on the PLX machine at the Centro Interuniversitario del Nord Est per il CAlcolo (CINECA), a linux infiniband cluster with each node equipped by two six-core 2.4 GHz processors and 48 Gb of RAM.

The left-hand panel of Fig. 1 shows a strong scaling test obtained by distributing a 720^3 particles box of $720\,h^{-1}$ Mpc of comoving length on one to 27 nodes, using 12 MPI tasks per node (one task per core). We do not use multi-threading in this version of the code. The red continuous curve gives the total time needed to complete the run while the dashed blue, dot–dashed green and dotted orange curves show, respectively, the time needed to compute inverse collapse

[8] http://www.mpa-garching.mpg.de/gadget/
[9] http://cosmo.nyu.edu/roman/2LPT/

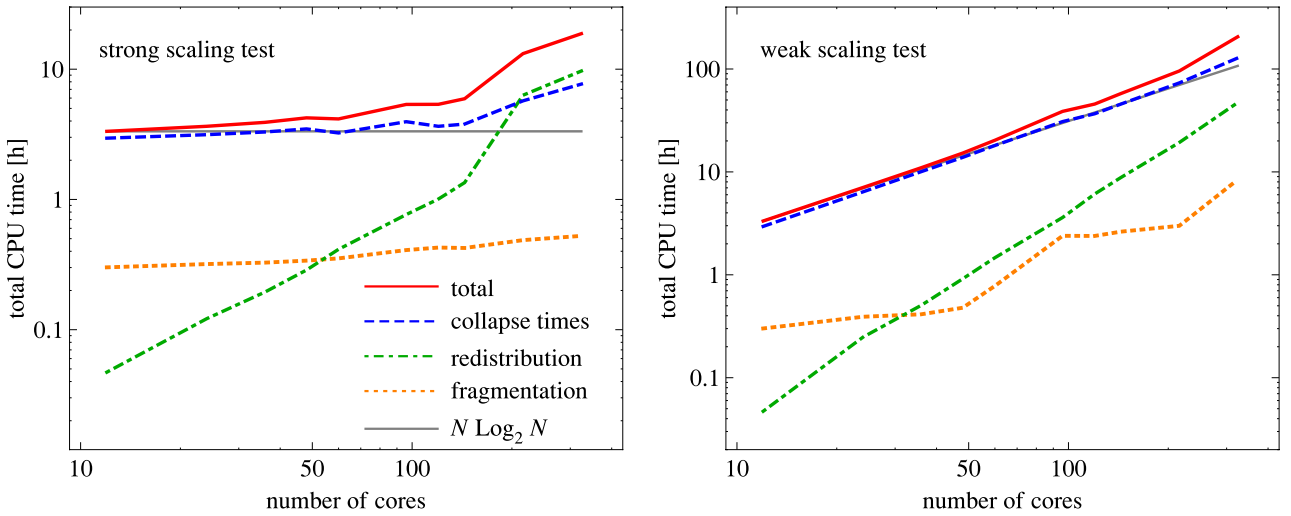

Figure 1. Left- and right-hand panels show a strong and a weak scaling test, respectively. In the strong test a $720\,h^{-1}$ Mpc box of 720^3 particles is distributed on one to 27 nodes (12 to 324 cores, one task per core), while in the weak test the number of particles is increased proportionally to the number of cores, starting from the same simulation on a single node. In each panel we show the total CPU time needed to complete the run (red, continuous curve), to compute the $F_{\rm max}$ inverse collapse times (blue, dashed, curve), to redistribute the memory from planes to sub-volumes (green, dot–dashed, curve) and to fragment the collapsed medium into haloes (orange, dotted curve). The black line gives the ideal $N\,\log_2 N$ scaling (in the strong scaling test the number of particles N is constant).

times $F_{\rm max}$, to perform the redistribution of memory from planes to sub-boxes and to fragment the collapsed medium. The horizontal black line gives the ideal $N\,\log_2 N$ scaling expected in this case (it is constant due to the fixed number of grid points). Thanks to the FFTW libraries, the computation of collapse times scales very near the ideal case up to 144 cores, with some increase of CPU time likely due to the increased overhead of communications. When more cores are used, FFTW starts to distribute planes to tasks in an uneven way, so that only some of the allocated cores are actually working (180 over 216 on 18 nodes, 240 over 324 on 27 nodes), while the others remain idle. This problem can be clearly avoided with a careful choice of the number of tasks. Fragmentation scales relatively well, with a modest increase of CPU time related to the increasing overhead of boundary layers. Redistribution is negligible for a small number of tasks but does not scale; in this test it becomes dominant at the same time when collapse times get far from the ideal scaling.

The right-hand panel of the same figure shows a weak scaling test obtained by increasing, at fixed mass resolution, the number of particles proportionally to the number of cores used, up to 2160^3 on 27 computing nodes (the same number of particles as the Millennium simulation; Springel et al. 2005). In this test we use rectangular boxes. The black line denotes the ideal $N\,\log_2 N$ scaling. Again, computation of collapse times and fragmentation scale very near the ideal case, while the redistribution becomes more and more significant as the number of tasks increases, though it is not dominant even for the largest simulation.

The 2160^3 particles box run on 324 cores of the PLX machine takes $\sim$38 min (for a cost of 206 CPU hours), with computation of collapse times taking 62 per cent of time (37 per cent needed by FFTs), redistribution 23 per cent and fragmentation 13 per cent. Just to give an example, a numerical project of 10 000 Millennium-sized simulations on the same machine would require only $\sim 2 \times 10^6$ CPU hours and would be over in less than 9 months on only 324 core, or a month on about 3000 cores. An improvement of the redistribution code would lower requirements by 20 per cent and would allow the code to be applied to larger box sizes.

3 SIMULATIONS

To test the accuracy of PINOCCHIO for the clustering of DM haloes we compare to a simulation taken from the MICE suite of cosmological N-body simulations (Crocce et al. 2010).[10] MICE is a large set of Λcold dark matter (ΛCDM) simulations performed with the GADGET-2 code described in Springel (2005). The MICE catalogues provide Friend-of-Friends (FoF) haloes with linking length $b = 0.2$ in units of the mean inter-particle distance. The assumed cosmology is that of a flat, ΛCDM universe with $\Omega_{\rm m} = 0.25$, $\Omega_{\rm b} = 0.044$, $n_s = 0.95$, $\sigma_8 = 0.8$ and $h = 0.7$ ($\Omega_{\rm b}$ is used to generate the initial conditions but all particles are collisionless). In the rest of this paper we will use the term 'haloes' for both N-body FoF haloes *and* for PINOCCHIO 'groups' of particles since this choice should not lead to any ambiguity.

We focus specifically on one run, MICE3072-HR (following the denomination of Crocce et al. 2010), consisting of a box of sides $3072\,h^{-1}$ Mpc sampled by 2048^3 particles, each of mass $2.3 \times 10^{11}\,h^{-1}\,{\rm M}_\odot$. Note that, unlike other simulations in the MICE suite, the MICE3072-HR run does not use 2LPT initial conditions. We use here for the mass function a tabulated correction provided by the MICE collaboration for those runs with ZA initial conditions, but we do not attempt to correct halo masses when applying mass cuts to compute correlation statistics. The PINOCCHIO run required 31 min on 25 computing nodes (300 cores), so the total cost was 155 CPU hours, a tiny fraction ($\sim$1/2000) of the 370 000 h needed by the N-body simulation on the Marenostrum supercomputer. For testing purposes, we have also used the smaller MICE768 simulation, a $768\,h^{-1}$ Mpc box sampled by 1024^3 particles with mass $2.9 \times 10^{10}\,h^{-1}\,{\rm M}_\odot$. This has a higher mass resolution, but its volume is not large enough to be used for large-scale clustering statistics. For clarity, we use similar names for the PINOCCHIO run, replacing the 'MICE' prefix with 'P'.

The comparison with MICE simulations allows us to test PINOCCHIO on much larger volumes than previously done. The halo catalogues

[10] Selected halo catalogues and other data are available for download at http://maia.ice.cat/mice/

are public and the mass function analysis has been performed by the MICE collaboration (Crocce et al. 2010). As mentioned above, the current version of PINOCCHIO makes use of the N-GENIC code for the initial conditions, so it is simple to set up the same initial conditions used in the simulations, removing differences due to sample variance, and allowing a comparison at the object-by-object level. Extensions of NGENIC to 2LPT initial conditions are available and are crucial for accurately simulating the halo mass function (Crocce et al. 2006). Finally, the N-GENIC code has been further extended to include non-Gaussian initial conditions (Scoccimarro et al. 2012). Although we will not consider this possibility here, the extension of PINOCCHIO to this specific departure from the Standard Cosmological Model is, in principle, straightforward.

Fig. 2 provides a first, qualitative comparison of PINOCCHIO with an N-body run at $z = 0$. For this comparison we have used the smaller MICE768 simulation that has a better mass resolution. In the large, top panel, corresponding to a $500 \times 400 \times 20\,h^{-1}\,\mathrm{Mpc}^3$ volume, blue dots represent individual haloes from P768 plotted on top of the corresponding haloes from MICE768, shown as red dots. The size of each dot is proportional to the halo virial radius. It has been enlarged for clarity, leading to unrealistic overlaps between haloes in each realization. All haloes with $\log_{10}M/(\,h^{-1}\,\mathrm{M}_\odot) \geq 12.5$ are shown. The lower panels show in detail a sub-volume of $90 \times 90\,h^{-1}$ Mpc area and the same thickness of $20\,h^{-1}$ Mpc, with the left and right ones corresponding respectively to the individual N-body and PINOCCHIO outputs and with the central one showing again the two together by means of open circles to provide a clearer comparison of sizes and positions.

While large-scale structure is well reproduced, it can be seen that the most massive haloes in PINOCCHIO tend to be more isolated than their MICE counterparts, which have more numerous smaller haloes in their vicinity. We know from Paper I that PINOCCHIO provides a good match at the object-by-object level, so this mismatch is related to the limitations of the ZA to properly reproduce the displacement field and reconstruct large-scale high-density peaks. The relatively thin slicing causes some matching halo pairs to be in or out the slice, thus artificially increasing the number of apparently unmatched haloes. To make full sense of this comparison, one should take into account that a 'PINOCCHIO halo' does not exactly coincide with an FoF halo, though parameters can be tuned to maximize their similarity. These issues will be explored in detail elsewhere.

3.1 Mass function and parameters

As explained in Section 2.1.2, the construction of haloes in the PINOCCHIO code depends on five free parameters whose values were determined in Paper I by fitting the mass function to the one obtained from simulations available at that time.[11] The mass function was shown to be accurate at the 5 per cent level when compared to FoF haloes with linking length $b = 0.2$ times the inter-particle distance, though a 10–20 per cent underestimate at large masses and high redshifts were reported.

For a proper comparison with much bigger simulations the parameters can be retuned to improve the agreement to a higher level of accuracy. To fully and properly complete such a task we must perform a large number of realizations and a detailed study at the object-by-object level of PINOCCHIO haloes in comparison with FoF and Spherical Overdensity (SO) haloes, paying specific attention to the high-mass tail. This will be presented in a forthcoming paper. The present paper aims at presenting a first test of version 3 of the PINOCCHIO code on cosmological volumes characterized by box sizes and particle numbers that are more than two orders of magnitude larger than those previously addressed.

At $z = 0$, numerical convergence among mass functions of simulated DM haloes has been reached at the $\lesssim$5 per cent level for masses $\lesssim 10^{14}\,\mathrm{M}_\odot$. For larger masses, differences among simulations can amount to several 10s of per cent. At fixed mass, the disagreement worsens at higher redshift, where objects correspond to rarer peaks of the linear density field. This is also an effect of the steepness of the high-mass tail of the mass function because of which small differences in mass result in large differences in number density. Moreover, the mass function is approximately 'universal', i.e. mass functions at all redshifts lie on the same relation when the adimensional quantity $(M^2/\bar{\rho})(\mathrm{d}n(M)/\mathrm{d}M)$ [$\mathrm{d}n(M)$ being the number density of haloes of mass between M and $M + \mathrm{d}M$] is shown as a function of $\nu = \delta_c/\sigma(M)$, with $\sigma(M)$ being the mass variance at the scale M. However, recent determinations have reported small but significant violations of universality (Tinker et al. 2008; Crocce et al. 2010). Fig. 3 compares, in terms of ratios to the Sheth & Tormen (1999) fitting formula, several analytic fits from the literature, obtained both for FoF (Jenkins et al. 2001; Warren et al. 2006; Reed et al. 2007; Crocce et al. 2010; Courtin et al. 2011; Angulo et al. 2012) and SO (Tinker et al. 2008) haloes.

Using PINOCCHIO with the parameter values given in Paper I, we confirmed the tendency, already noticed in Paper I and in Peel, Battye & Kay (2009), of PINOCCHIO to underestimate the number density of rare objects. To improve this trend we performed some parameter tuning. We first dropped the dependence, described in appendix A of Paper I, of f_a and f_{ra} on resolution. Both f_a and f_{ra} influence the normalization of the mass function, but the first one also steepens it. We checked that increasing f_a to 0.285 while lowering f_{ra} to 0.180 provides a number density of rare objects compatible with simulations though at the lower end of the allowed range. The other parameters were left as in Paper I; Table 1 reports the parameter values used in this paper. Finer tuning can be achieved by using several large nested boxes and sampling a wider parameter space; because the number of constraints is much larger than the number of parameters, degeneracies in parameter values can be broken with this approach. In Fig. 3 we report the mass function measured in the P768 (blue data points) and P3072-HR (red data points) realizations; the error bars on the data points represent simply the expected Poisson error. We notice in the first place good agreement between the results for the two boxes with different resolution, particularly at large redshift. We find good agreement over quite a large range of masses with the Warren et al. (2006) fit (and to a lesser extent to the Angulo et al. 2012, one) However, we were unable to reproduce the MICE FoF counts of Crocce et al. (2010) at high redshift. Since the agreement with the SO mass function of Tinker et al. (2008) is better, the disagreement with MICE may be related to the tendency of FOF to overlink haloes and to include filaments that surround the rarest haloes.

To assist in the interpretation of the correlation results presented in the next section, Fig. 4 shows a more direct comparison of both the MICE fit and the MICE3072-HR mass function with the one from P3072-HR. In particular, the upper panel of Fig. 4 shows the PINOCCHIO adimensional mass function (red curve with error bars) with the results of MICE3072-HR (blue curve) and the analytic fits of Crocce et al. (2010) (black curve) and Warren et al. (2006) (*black, dashed curve*). The lower panels give the residuals for both PINOCCHIO and MICE results w.r.t. the Warren et al. (2006) fitting

[11] A standard CDM simulation with 360^3 particles in a box $500\,h^{-1}$ Mpc on a side, and a smaller ΛCDM simulation with 256^3 particles in a $100\,h^{-1}$ Mpc box.

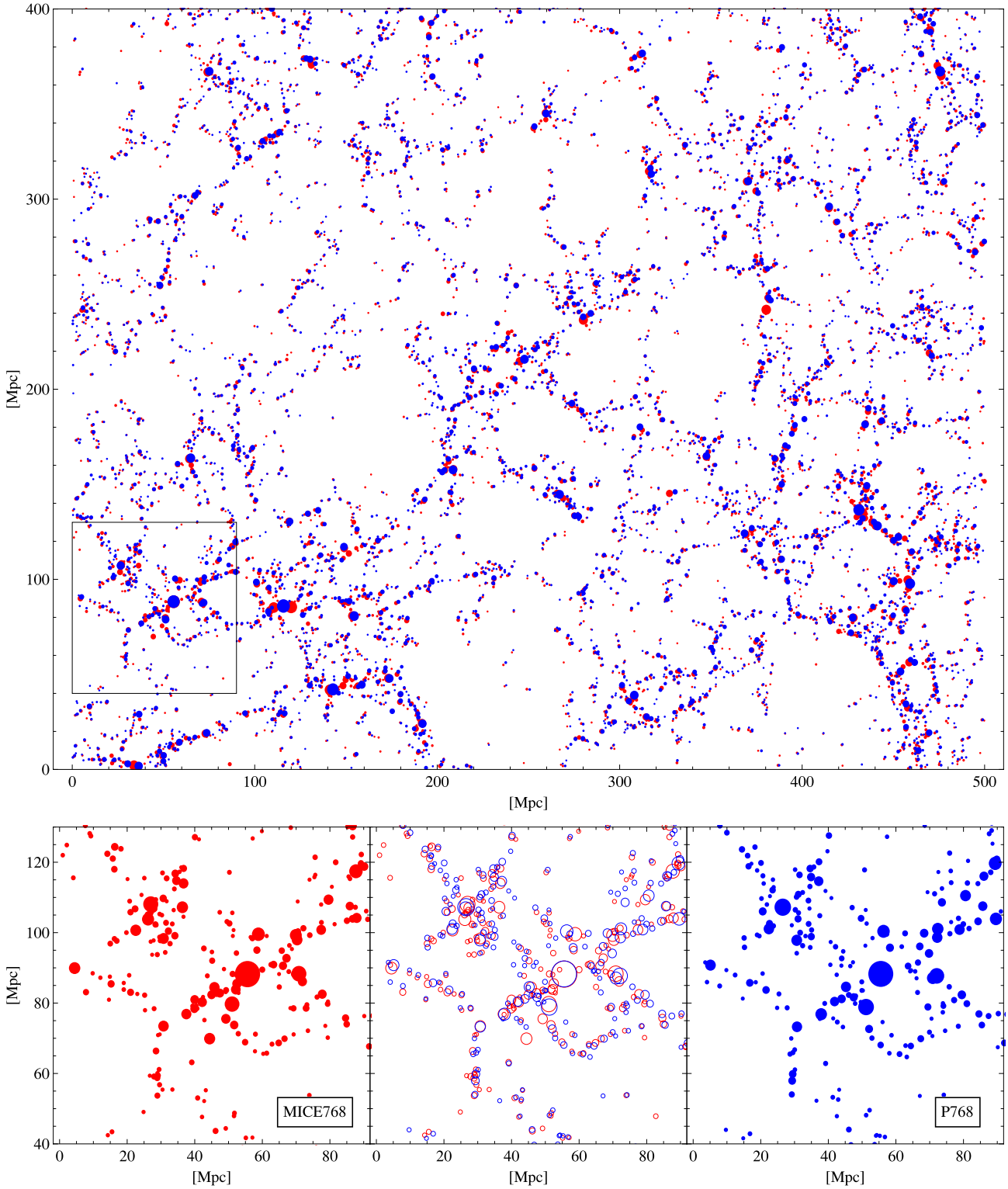

Figure 2. Comparison between the halo distributions predicted by the PINOCCHIO P768 realization (blue) and the MICE768 *N*-body simulation (red) on a $500\,h^{-1}$ Mpc $\times$ $400\,h^{-1}$ Mpc field, $20\,h^{-1}$ Mpc slice. The upper panel shows the entire field. The lower panels show a $70\,h^{-1}$ Mpc $\times$ $70\,h^{-1}$ Mpc zoom with the two separate distributions and overlapping as circles. All haloes with $\log_{10}M/(\,h^{-1}\,\mathrm{M}_{\odot}) \geq 12.5$ are shown by discs and circles having 1.7 $\times$ the virial radius.

formula. Error bars are shown only for the PINOCCHIO output and account for Poisson noise.

The lower mass bin used in this plot corresponds to haloes of a minimum of 200 particles, both for PINOCCHIO and for MICE, i.e. the same cut-off assumed in Crocce et al. (2010) for the fitting procedure. FoF masses from the MICE catalogues account for the mass correction suggested by Warren et al. (2006) in order to avoid the statistical noise effects due to the estimate of the halo density

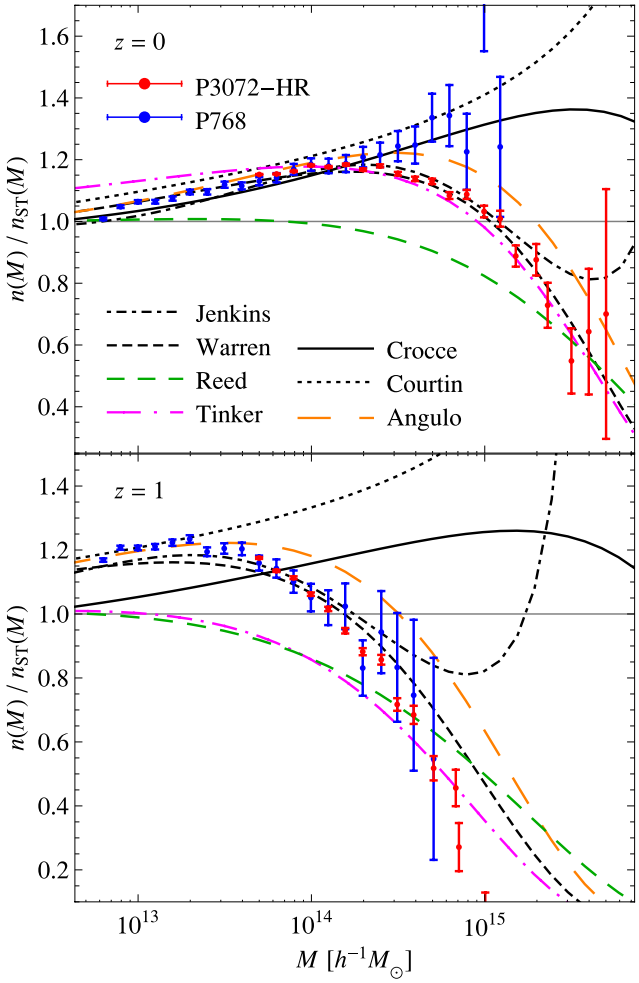

Figure 3. Ratio of the mass function measured in the P768 (blue data points) and P3072-HR (red data points) realizations to the Sheth & Tormen (1999) fitting formula. For comparison, the same ratio is shown for several other analytic fits taken from the literature: Jenkins et al. (2001) (black, dot–dashed curve), Warren et al. (2006) (black, short dashed), Reed et al. (2007) (green, medium dashed), Tinker et al. (2008) (magenta, dot-long dashed), Crocce et al. (2010) (black, continuous), Courtin et al. (2011) (black, dotted) and Angulo et al. (2012) (orange, long dashed).

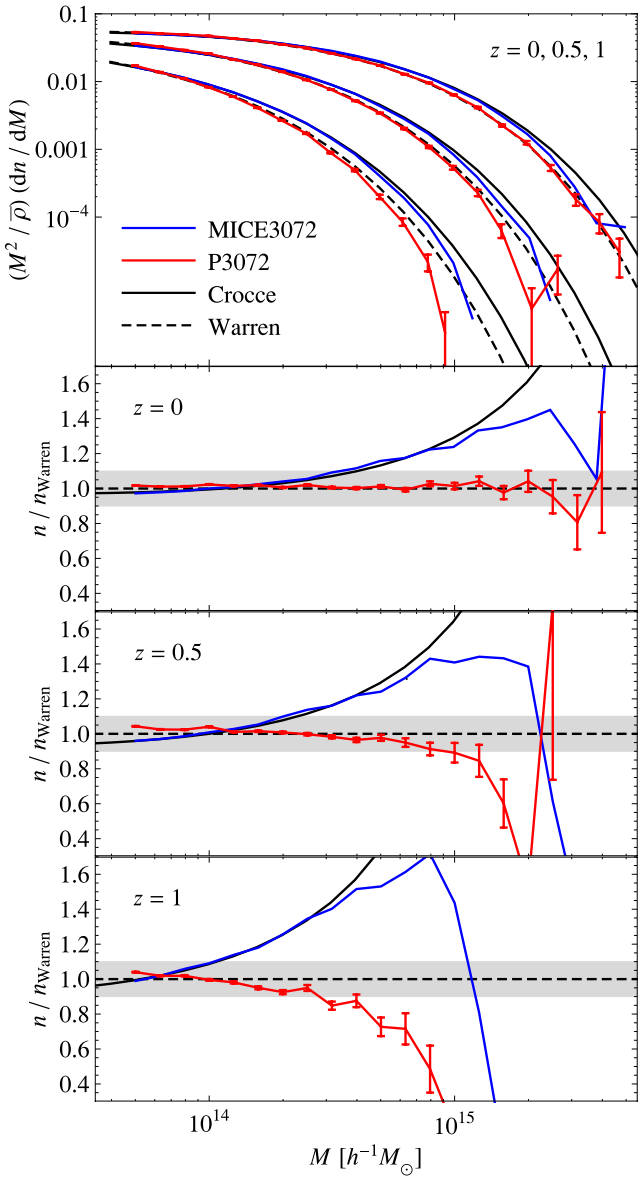

Figure 4. The top panel shows the adimensional mass function predicted by PINOCCHIO (red, continuous curve with error bars) compared with the Warren et al. (2006) fit (black, dashed curve) and the MICE fit (black, continuous curve) and data (blue, continuous curve) for MICE3072-HR at $z = 0$, 0.5 and 1. The lower panels show, for each redshift, the residuals w.r.t. the Warren et al. (2006) fitted with, in addition to the MICE results. The shaded grey region corresponds to deviations within 10 per cent w.r.t. the Warren et al. (2006) prediction.

field with a finite number of particles. This correction corresponds to defining a 'corrected' number of particles per halo given by $n_p^{\rm corr} = n_p\,(1 - n_p^{-0.6})$. Also, the MICE mass function is corrected as suggested in Crocce et al. (2010) to reproduce the result of 2LPT initial conditions. We do not consider such corrections for PINOCCHIO masses, which are instead given simply by the number of particles belonging to a given halo.

As already noted, PINOCCHIO reproduces the mass function of MICE3072-HR within 10 per cent at $M \sim 10^{14}\,h^{-1}\,{\rm M}_\odot$. At larger masses, however, it increasingly underestimates the MICE halo number density, especially when compared with the MICE analytic fit that takes advantage of the results from the full MICE set, including two boxes of larger size with respect to MICE3072-HR. Thanks to the parameter tuning, the P3072-HR mass function reproduces the Warren fit to within a few per cent in the range where the MICE3072-HR mass function closely follows the MICE analytic fit. For $z > 0$ the PINOCCHIO mass function goes below the Warren et al. (2006) fit, but this happens in the same mass range where the MICE mass function for the same MICE3072-HR box starts to underestimate the analytical fit obtained using larger boxes. So this discrepancy may be related to the smallness of the box.

4 ACCURACY TESTS FOR CLUSTERING STATISTICS

In this section we present a direct comparison of the halo power spectrum and bispectrum predicted by the current version of PINOCCHIO with their counterparts measured in the MICE3072-HR simulations.

4.1 Power spectrum

Paper I, using the two-point correlation function, showed that on the relatively small scales ($<30\,h^{-1}$ Mpc) testable with those simulations and using Zel'dovich displacements, the clustering of haloes is recovered by PINOCCHIO at the 10–20 per cent level. Here we examine instead measurements of the halo–halo power spectrum and the halo–mass cross-power spectrum on significantly larger scales, encompassing the acoustic oscillation range at low cosmic variance.

A first comparison between PINOCCHIO and MICE is performed in mass thresholds, taking each halo mass at face value, as predicted by PINOCCHIO, without any rescaling to match the two mass functions. Therefore, any mass function discrepancy will affect the normalization of the power spectrum. Fig. 5 shows the ratios between the matter–halo (top panels) and halo–halo (bottom panels) power spectra from P3072-HR and the corresponding ones from the MICE3072-HR. The first quantity, in particular, is obtained by correlating the halo number density, computed on a grid with a cloud-in-cell algorithm, with the non-linear mass density field measured from the simulation output for both PINOCCHIO and MICE. We consider as an example three distinct populations defined by $\log_{10}(M/\,h^{-1}\,\mathrm{M}_\odot) > 13.7$, 14 and 14.5. The lowest mass threshold corresponds to haloes of 200 particles. Finally, for the halo power spectrum comparison we keep the shot-noise contribution, since the halo power spectrum including shot-noise is more relevant for covariance estimation purposes.

Differences in the mass functions will result in differences in the power spectrum (mostly its normalization) that will be larger for the halo–halo case. For a fixed mass threshold, in fact, PINOCCHIO objects are relatively rarer, and therefore more biased, than their MICE counterparts. The results of Fig. 5 show that for the lowest mass thresholds ($\log_{10}(M/Ms) > 13.7$) and lower redshift $z \leq 0.5$, where the mass function is well-matched, the discrepancy between the PINOCCHIO and MICE cross-power is below the 5 per cent level at scales $k < 0.1\,h\,\mathrm{Mpc}^{-1}$. The agreement worsens in the case of the halo–halo power spectrum but stays well within the 10 per cent level. Larger discrepancies in the normalization, of the order of 10–20 per cent, occur for the largest mass threshold and at $z = 1$.

To confirm our interpretation of the impact of the mass function discrepancy on the power spectrum, we consider as well power spectrum measurements performed on halo populations defined directly in terms of halo number density. The corresponding ratios with the MICE results are shown in Fig. 6. In this case we consider populations defined by a total number of most massive haloes N taking the values $\log_{10}N = 4$, 4.5, 5, 5.5 and 6, corresponding to masses roughly ranging from $\log_{10}(M/\,h^{-1}\,\mathrm{M}_\odot) \simeq 13.9$ to 14.9 at redshift zero. Notice the remarkably low scatter among different density populations. The overall departure from the MICE results at large scales is about 4 and 8 per cent for the cross- and halo–halo power spectra, respectively.

In the mildly non-linear regime an increasing suppression of PINOCCHIO power spectra is also evident. It is stronger for smaller haloes, and is likely related to PINOCCHIO's use of the ZA for particle displacements and consequent halo positions. To demonstrate this, Fig. 7 shows the ratio between the matter power spectrum obtained from the Zel'dovich displacements for all particles and the same quantity computed using the simulation output. We consider, in this case, the smaller MICE768 run and the corresponding P3072-HR one. Similarly to Fig. 5, this ratio shows a damping of the ZA power spectrum by 15 per cent at $k = 0.1\,h\,\mathrm{Mpc}^{-1}$. Clearly PINOCCHIO haloes cannot show better clustering properties as long as ZA is used for the displacements.

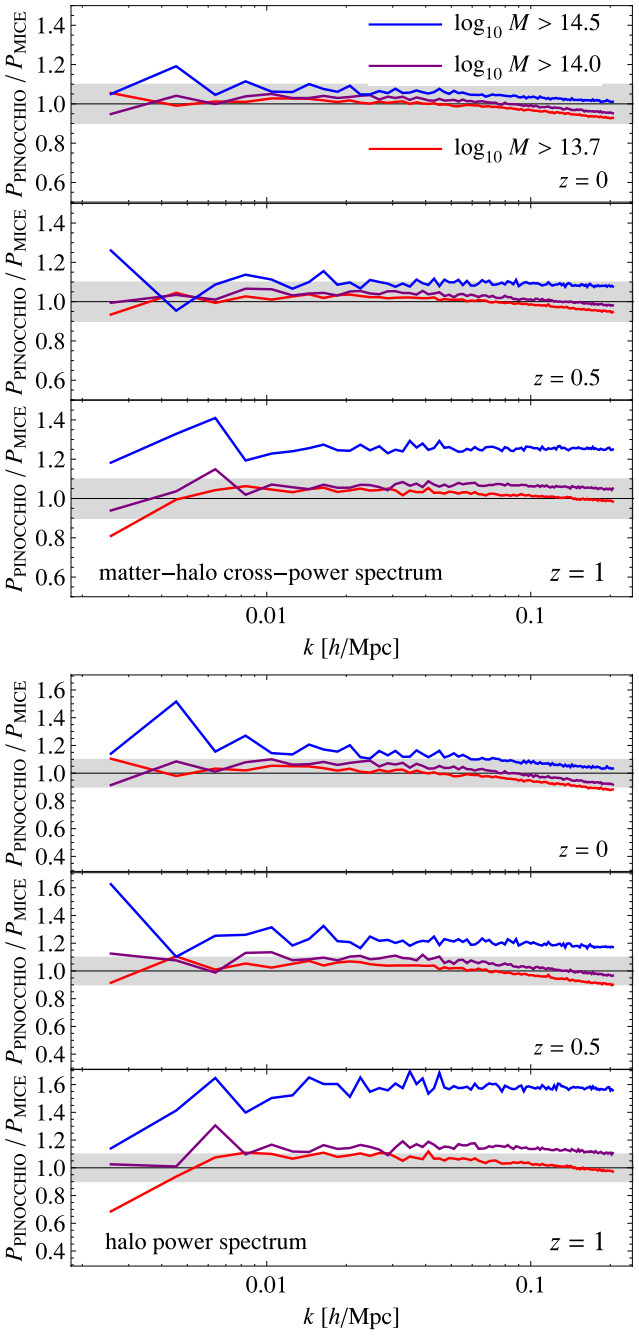

Figure 5. Ratios between power spectra computed using the P3072-HR and MICE3072-HR catalogues. Top panels show the matter–halo cross-power spectrum, bottom panels the halo–halo power spectrum. We show results at redshifts $z = 0$, 0.5 and 1 for different thresholds in mass defined by $\log_{10}(M\,h^{-1}\,\mathrm{M}_\odot) > 13.7$, 14 and 14.5. The shaded grey area corresponds to discrepancies below 10 per cent.

The inaccuracy of Zel'dovich displacements can be approximately described as a Gaussian scatter of halo positions about the true ones. Considering this uncertainty, the power spectrum of PINOCCHIO haloes, in turn, can be crudely modelled as the 'true' one obtained from the simulation times a Gaussian, random scatter term:

$$P_{\rm PIN}(k) = P_{\rm MICE}(k)\,\mathrm{e}^{-k^2 d^2}. \qquad (5)$$

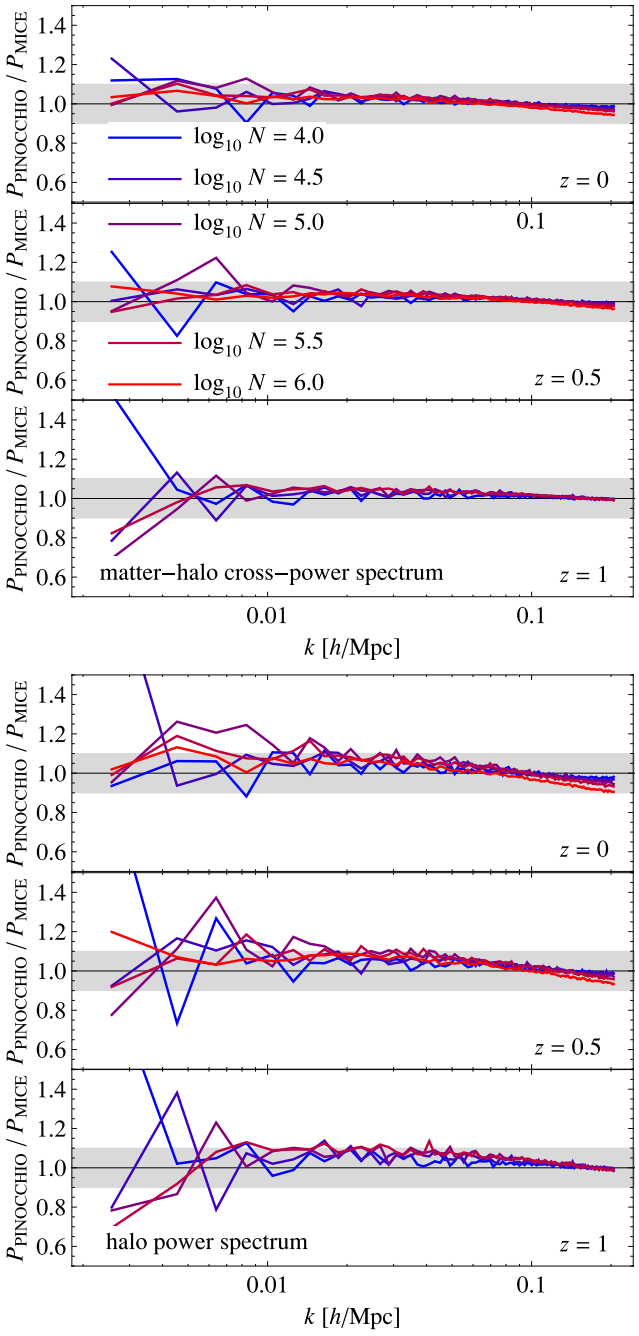

Figure 6. Same as Fig. 6 but for halo populations defined by a fixed number N of the most massive haloes with $\log_{10}N = 4$, 4.5, 5, 5.5 and 6. The shaded grey area corresponds to discrepancies below 10 per cent.

Fig. 8 shows the ratio of the MICE (blue curves) and PINOCCHIO (red curves) power spectra (with shot-noise subtracted) w.r.t. the linear, matter power spectrum without acoustic oscillations (Eisenstein & Hu 1998). The normalization of the PINOCCHIO power spectrum is rescaled to match the MICE one at large scales in order to highlight the different scale dependence at small scales. In addition, the figure shows a *corrected* PINOCCHIO power spectrum obtained as $P_{\rm PIN}\, {\rm e}^{k^2 d^2}$, with $d = 3$ and $2.7\,h^{-1}$ Mpc, respectively, at $z = 0$ and 0.5. The two panels show two different populations defined by different thresholds in mass, as indicated. After this Gaussian damping correction is applied, the residual difference is roughly a constant bias, whose value depends on mass, as already shown in Fig. 5.

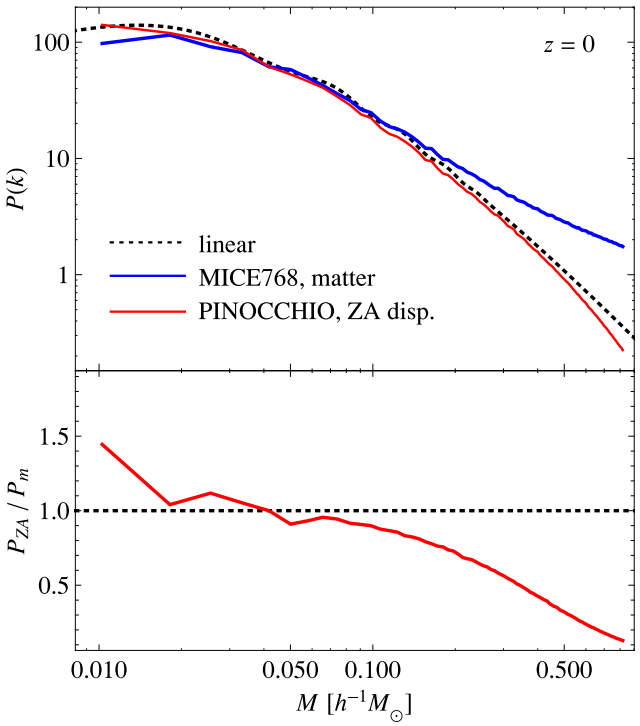

Figure 7. Top panel: comparison of the non-linear matter power spectrum from MICE768 (blue curve) with that measured using ZA displacements for all particles (red, continuous curve), and in linear theory (dotted curve). Bottom panel: ratio between the PINOCCHIO ZA power spectrum and the fully non-linear matter power spectrum in MICE.

Note that the PINOCCHIO halo power spectrum reproduces quite accurately the sampling noise of the N-body power spectrum over the whole Baryonic Acoustic Oscillations (BAOs) range. This is particularly evident from the middle panel of Fig. 8, corresponding to $\log_{10}M > 13.7$, where the corrected PINOCCHIO values practically coincide with the MICE measurements, including the constant bias term. However, additional corrections are required at smaller scales ($k > 0.17\,h\,{\rm Mpc}^{-1}$).

4.2 Bispectrum

To provide a complete statistical description of the halo distribution it is necessary to consider its non-Gaussian properties. In this respect, the lowest order correlation statistic that measures this in Fourier space is the bispectrum. Here we compare measurements from MICE3072-HR and P3072-HR of the *reduced* halo bispectrum, $Q_{\rm h}(k_1, k_2, k_3)$ (Fry 1984). This quantity is defined as the ratio between the halo bispectrum, $B_{\rm h}(k_1, k_2, k_3)$, i.e. the three-point function of the halo density field in Fourier space, and a suitable combination of quadratic terms of the halo power spectrum, that is

$$Q_{\rm h}(k_1, k_2, k_3) = \frac{B_{\rm h}(k_1, k_2, k_3)}{P_{\rm h}(k_1)\, P_{\rm h}(k_2) + 2{\rm perm.}}\,. \qquad (6)$$

The denominator removes the overall scale dependence of the halo bispectrum to highlight its dependence on the shape of the triangular configuration considered.

We measure the halo bispectrum $B(k_1, k_2, k_3)$ for all triangular configurations defined by wavenumbers k_i in bins of $\Delta k = 0.004\,h\,{\rm Mpc}^{-1}$ up to a maximum value of $k_{\rm max} = 0.13\,h\,{\rm Mpc}^{-1}$,

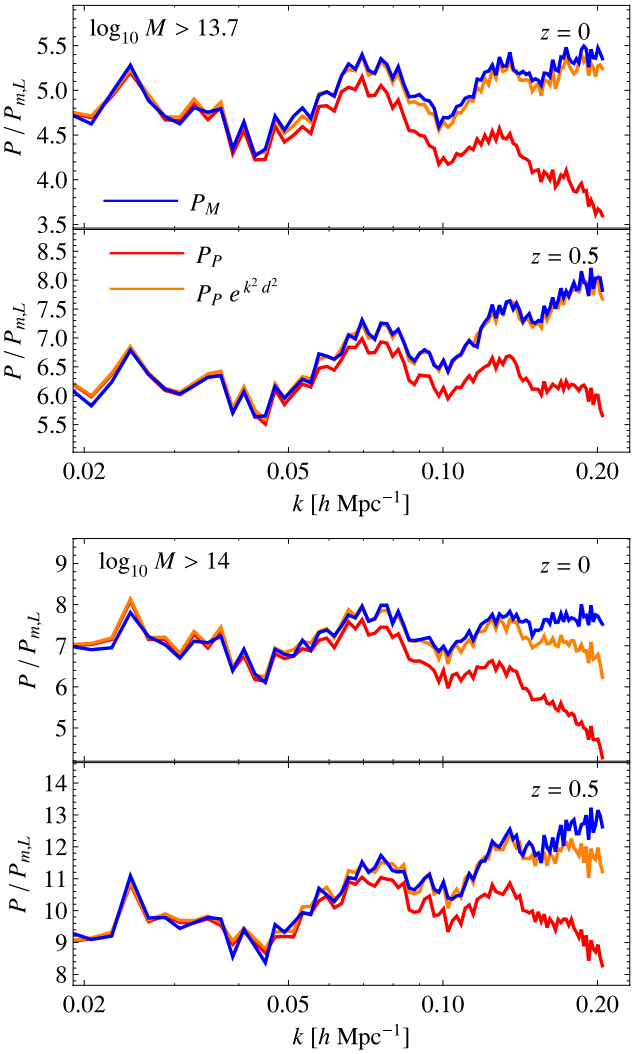

Figure 8. Comparison of the PINOCCHIO and MICE halo power spectra over the acoustic oscillations range. We show the ratio of the MICE3072-HR (blue curves) and P3072-HR (red curves) power spectra w.r.t. the $\Omega_b = 0$ linear matter power spectrum of Eisenstein & Hu (1998). In addition, the orange curves show the P3072-HR power spectrum corrected according to equation (5) with $d = 3$ and $2.7\,h^{-1}$ Mpc, respectively, at $z = 0$ and 0.5. The normalization of the PINOCCHIO power spectrum is rescaled to match the MICE one at large scales to highlight the different scale dependence at small scales. Different panels show different thresholds in mass as indicated. In all cases, shot noise has been removed. PINOCCHIO reproduces the sampling noise induced by the random initial conditions correctly over the whole range shown.

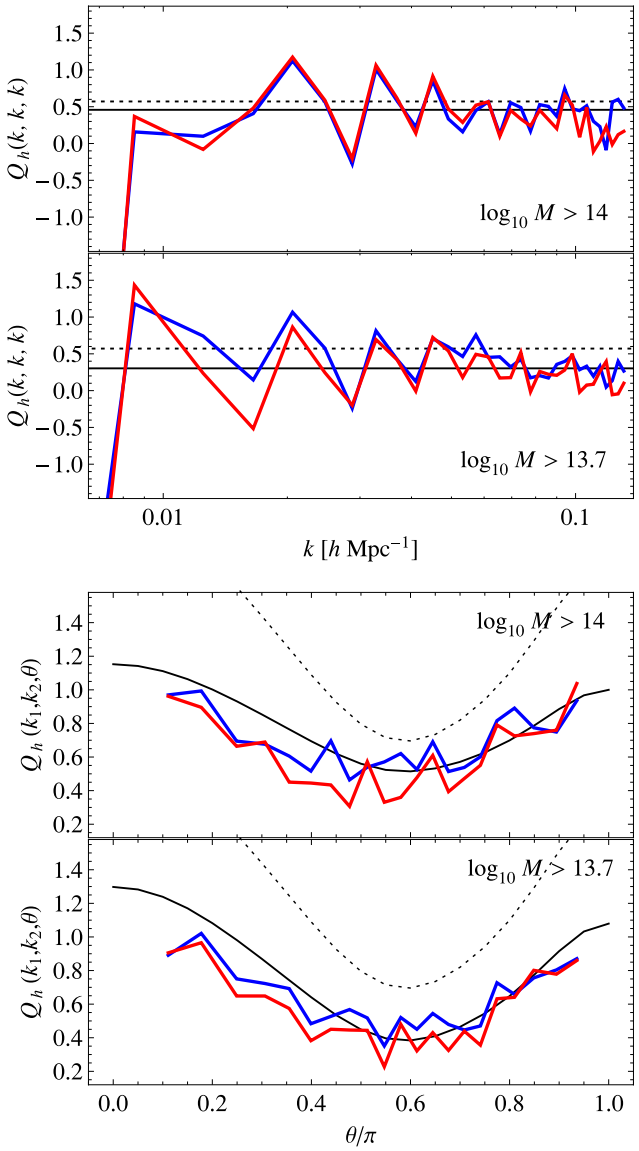

Figure 9. Comparison of the reduced halo bispectrum. Top two panels show, for two different mass thresholds, measurements of the reduced halo bispectrum for equilateral configurations, $Q(k, k, k)$ as function of k. Bottom two panels show the same quantity, $Q(k_1, k_2, \theta)$ with two wavenumbers are fixed to the values $k_1 = 0.04\,h\,\mathrm{Mpc}^{-1}$ and $k_2 = 2k_1$, as a function of the angle theta between them. In all panels blue curves show measurements from MICE3072-HR and red curves measurements from P3072-HR. The dotted, black curve provides the prediction for the reduced *matter* bispectrum at tree-level in perturbation theory.

focusing therefore on large scales. On such scales, it is possible to approximate the halo bispectrum by (Fry & Gaztañaga 1993)

$$Q_{\rm h}(k_1, k_2, k_3) = \frac{1}{b_1}\, Q(k_1, k_2, k_3) + \frac{b_2}{b_1^2}\,, \qquad (7)$$

where $Q(k_1, k_2, k_3)$ denotes the reduced bispectrum of the matter distribution, defined similarly to equation (6) in terms of matter correlators while b_1 and b_2 represent, respectively, the linear and quadratic bias coefficient, assumed to be constant for a given halo population (for a recent assessment of the validity of this approximation, see, for instance, Chan, Scoccimarro & Sheth 2012; Pollack, Smith & Porciani 2012; Sefusatti, Crocce & Desjacques 2010, 2012).

From the above expression it is therefore evident that the reduced halo bispectrum is equally sensitive to linear and non-linear bias, providing a test for the ability of PINOCCHIO to correctly reproduce halo bias beyond its linear approximation relevant for the large-scale power spectrum. With this in mind, the upper panel of Fig. 9 shows the equilateral configurations of the reduced halo bispectrum, $Q_{\rm h}(k, k, k)$, as a function of k for two different mass thresholds at $z = 0$. The bottom panel shows instead the quantity $Q(k_1, k_2, \theta)$ with two wavenumbers fixed to the values $k_1 = 0.04\,h\,\mathrm{Mpc}^{-1}$ and $k_2 = 2k_1$, as a function of the angle θ between them. In all panels blue curves show measurements from MICE3072-HR and red curves measurements from P3072-HR. The dotted, black curve

represents the prediction for the reduced *matter* bispectrum at tree-level in perturbation theory, while the continuous black curve shows equation (7), with the values for b_1 and b_2 determined from applying the peak-background split argument to the Crocce et al. (2010) mass function: $b_1 \simeq 2.2$ and 2.6 and $b_2 \simeq 1$ and 2.2 for the two mass thresholds of $\log_{10}M = 13.7$ and 14 at $z = 0$. These theoretical predictions turn out to be rather accurate, despite the crudeness of the tree-level expression in equation (7).

These measurements show that PINOCCHIO accurately reproduces the hierarchical relation between halo power spectrum and bispectrum and to a certain extent the non-linear properties of halo bias. In fact, although the bispectrum itself suffers from the same Gaussian damping effect which we saw for the halo–halo power spectrum, this approximately cancels out in the ratio defined in equation (6). This is remarkable also because the first-order, i.e. Zel'dovich, displacements do not guarantee good recovery of the non-linear bispectrum, failing to reproduce even its tree-level expression in Eulerian PT.

5 DISCUSSION AND CONCLUSIONS

We have presented a new parallel (MPI) version of the PINOCCHIO code, optimized to run on hundreds of cores of a super-computer. We showed that the PINOCCHIO code can quickly produce simulated catalogues of DM haloes that closely reproduce the results of very large simulations like those of the MICE project (Crocce et al. 2010): halo abundances are almost universal, and within a few per cent of the fit proposed by Warren et al. (2006). On the other hand PINOCCHIO underproduces the abundance of the rarest FoF objects found in some recent simulations, including those of the MICE group. This may indicate problems with the fragmentation part of the PINOCCHIO code, but it may also be related to the tendency of the FOF algorithm to overlink haloes.

In addition to the abundances, we showed that PINOCCHIO can also reproduce the spatial distribution of the haloes. Despite the fact that it uses Zel'dovich displacements to compute the final positions of haloes, PINOCCHIO can reproduce the matter–halo and the halo–halo power spectra at $k < 0.1\,h\,\mathrm{Mpc}^{-1}$. The linear bias factor is well recovered as long as the number density of objects is well matched.[12] Using mass thresholds so as to match the same number density of haloes in the PINOCCHIO and MICE catalogues, the matter–halo and the halo–halo power spectra of the simulation are recovered to within 4 and 8 per cent, respectively. At smaller scales the PINOCCHIO power spectrum shows a damping that is due to the inaccuracy of Zel'dovich displacements in predicting the final positions of haloes. This can be roughly modelled as a Gaussian noise term with a damping scale of $3\,h^{-1}\,\mathrm{Mpc}$ at $z = 0$. Good agreement is also obtained for the reduced bispectrum of the haloes, with the effects of the damping term approximately cancelling out in the ratio. For both two- and three-point statistics the noise in the quantities computed from PINOCCHIO catalogues closely follows that computed from simulations. We did not address in this paper velocity fields. This is an important step in the analysis that must be performed before extending the comparison to the redshift space. Monaco et al. (2005) and Heisenberg et al. (2011) already presented studies of the behaviour of peculiar velocities of PINOCCHIO haloes. A detailed study of velocity power spectrum and redshift-space clustering will be the subject of a forthcoming paper.

This version of the PINOCCHIO code is particularly suited for addressing the massive production of catalogues of DM haloes, the first step in the generation of mock galaxy catalogues using Halo Occupation Distribution, abundance matching or semi-analytic models. Its scaling properties demonstrate the feasibility of running many (∼10 000) massive simulations in a reasonable amount of time: we estimate 2×10^6 CPU hours to produce 10 000 2160^3 boxes, though some minor parts of the code scale poorly and must be improved. The typical result of a PINOCCHIO run is a catalogue of haloes with known mass, position, velocity and merger history that requires orders of magnitude less disc space than needed by a typical simulation, not to mention the complicated and ill-defined post-processing needed by a standard simulation to produce well-behaved halo merger histories, which are a natural outcome of PINOCCHIO. The speed-up comes at the cost of information about the internal structures of haloes. But with refined models of the evolution of DM haloes after mergers, such as are commonly used in semi-analytic models to predict the merging time of galaxies, it is possible to approximately reconstruct the abundance of halo substructures from halo merging histories (see e.g. Giocoli et al. 2010).

Clearly, PINOCCHIO is not meant to be a substitute for *N*-body simulations. Rather, a natural application of our code is the determination of the covariance properties of large-scale structure observables (e.g. the galaxy power spectrum), as well as the study of systematic effects (e.g. the selection function) and possible correlation between the two (see e.g. Ross et al. 2012). Indeed, the mocks produced by Manera et al. (2013) with a version of the PTHALOS code for the BOSS survey (Eisenstein et al. 2011) were an essential ingredient for many analyses beyond error estimation, like power spectrum at large scales, BAOs, redshift distorsions. PINOCCHIO itself has been used by de la Torre et al. (2013) for computing the covariance matrix of the redshift–space galaxy correlation function in the range of scales ∼1 to ∼30 Mpc. This is a difficult range to reproduce, because Zel'dovich displacements are inaccurate at these scales. Nevertheless, the authors could take advantage of a relatively large number of PINOCCHIO realizations by applying the shrinkage method of Pope & Szapudi (2008), using only a few mocks from the Multi-Dark simulation by Prada et al. (2012) to subtract out the bias in the determination of the correlation function. While the determination of uncertainties does not require per cent accuracy, a very large number of mock catalogues is crucial for the proper estimation of large covariance matrices.

PINOCCHIO shows several advantages compared to other simplified tools for the quick production of large-scale structure. Algorithms based on LPT to reproduce the non-linear matter density field may be quicker, but they are not as precise in determining where the DM haloes are, especially at small masses (Manera et al. 2013). Methods which use Particle-Mesh integrations in a few time-steps can be very accurate in the generating the non-linear mass field (Tassev et al. 2013), but the price paid is poor time sampling of halo merger histories, as well as the post-processing needed to produce halo catalogues in the first place. The sparse time-sampling also complicates the generation of halo catalogues along the past light cone which is much simpler in PINOCCHIO because all displacements are always done in one single time-step, so any level of time sampling can be easily achieved. In particular, masses are updated every time a particle is added to the group, and merger histories report masses for each merging pair of haloes, so with the minimal output given by PINOCCHIO, halo mass accretion histories are available at

[12] In a forthcoming paper (Paranjape et al. 2013) we will compare PINOCCHIO predictions with simulations and theoretical expectations based on excursion set theory, and will show that PINOCCHIO can reproduce the bias of DM haloes well beyond the linear bias approximation.

each halo merger even without outputting the halo catalogues many times.

The present version of the code works for a range of ΛCDM cosmologies that includes arbitrary redshift-dependent equation of state of the quintessence, and can be easily extended to non-Gaussian cosmologies simply by changing the initial conditions generator. We are currently developing PINOCCHIO in two further directions. Positions of haloes can be computed with second and third-order LPT with associated overheads in memory and CPU time amounting to ∼30 and ∼100 per cent. We expect that 2LPT would improve the accuracy with which halo positions (and hence masses) are predicted, thus improving the halo power spectra and bispectra (i.e. reducing the corrections currently needed at large wavenumbers). It also could help in recovering the right number density of very rare haloes. Full 3LPT would allow one to predict the collapse times without using the ellipsoidal truncation of LPT proposed by Monaco (1997) that works under the approximation that using the growing mode as a time coordinate factorizes the dependence of cosmology out of the dynamics of a mass element. This would allow one to quickly produce simulations in any cosmology where an LPT expansion can be formulated.

The other direction of development is the on-the-fly production of the output on the past light cone of an observer randomly placed in the simulation volume, taking advantage of the periodic boundary conditions to simulate a very large volume. The fine time sampling of PINOCCHIO eliminates the need to output the full catalogues many times as must be done if one wishes to reconstruct the past-light cone at the post-processing level in the conventional way.

Our final aim is to propose a quick, flexible, scalable and open-source tool to generate, with minimal resources, large catalogues of DM haloes that reproduce the statistics of simulations to an accuracy which justifies the use of this tool for high-precision cosmology.

ACKNOWLEDGEMENTS

We thank Stefano Anselmi for discussions. PM and SB acknowledge financial contributions from the European Commissions FP7 Marie Curie Initial Training Network CosmoComp (PITN-GA-2009-238356), from PRIN MIUR 2010-2011 J91J12000450001 'The dark Universe and the cosmic evolution of baryons: from current surveys to Euclid', from PRIN-INAF 2009 'Towards an Italian Network for Computational Cosmology', from ASI/INAF agreement I/023/12/0, from PRIN-MIUR09 'Tracing the growth of structures in the Universe" and from a FRA2012 grant of the Trieste University. ES and RS were supported in part by NSF-AST 0908241.

REFERENCES

Ade P. A. R. et al., 2013a, preprint (arXiv e-prints, 1303.5062)
Ade P. A. R. et al., 2013b, preprint (arXiv e-prints, 1303.5076)
Amendola L. et al., 2012, preprint (arXiv e-prints, 1206.1225)
Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, MNRAS, 426, 2046
Bennett C. L. et al., 2012, preprint (arXiv e-prints, 1212.5225)
Benson A. J., Borgani S., De Lucia G., Boylan-Kolchin M., Monaco P., 2012, MNRAS, 419, 3590
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Buchert T., Ehlers J., 1993, MNRAS, 264, 375
Carbone C., Fedeli C., Moscardini L., Cimatti A., 2012, J. Cosmol. Astropart. Phys., 3, 23
Catelan P., 1995, MNRAS, 276, 115
Chan K. C., Scoccimarro R., Sheth R. K., 2012, Phys. Rev. D, 85, 083509
Costanzi Alunno Cerbolini M., Sartoris B., Xia J.-Q., Biviano A., Borgani S., Viel M., 2013, preprint (arXiv e-prints, 1303.4550)
Courtin J., Rasera Y., Alimi J.-M., Corasaniti P., Boucher V., Füzfa A., 2011, MNRAS, 410, 1911
Crocce M., Pueblas S., Scoccimarro R., 2006, MNRAS, 373, 369
Crocce M., Fosalba P., Castander F. J., Gaztañaga E., 2010, MNRAS, 403, 1353
Das S. et al., 2013, preprint (arXiv e-prints, 1301.1037)
de la Torre S. et al., 2013, preprint (arXiv e-prints, 1303.2622)
Desjacques V., Seljak U., 2010, Classical Quant. Gravity, 27, 124011
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Eisenstein D. J. et al., 2011, AJ, 142, 72
Fosalba P., Gaztañaga E., Castander F. J., Manera M., 2008, MNRAS, 391, 435
Frigo M., Johnson S. G., 2012, FFTW: Fastest Fourier Transform in the West. Astrophysics Source Code Library, 1201.015
Fry J. N., 1984, ApJ, 279, 499
Fry J. N., Gaztañaga E., 1993, ApJ, 413, 447
Giocoli C., Tormen G., Sheth R. K., van den Bosch F. C., 2010, MNRAS, 404, 502
Heisenberg L., Schäfer B. M., Bartelmann M., 2011, MNRAS, 416, 3057
Hinshaw G. et al., 2012, preprint (arXiv e-prints, 1212.5226)
Jahnke K., Macciò A. V., 2011, ApJ, 734, 92
Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372
Kitaura F.-S., Heß S., 2012, preprint (arXiv e-prints, 1212.3514)
Lahav O., Kiakotou A., Abdalla F. B., Blake C., 2010, MNRAS, 405, 168
Laureijs R. et al., 2011, preprint (arXiv e-prints, 1110.3193)
Li Y., Mo H. J., van den Bosch F. C., Lin W. P., 2007, MNRAS, 379, 689
Liguori M., Sefusatti E., Fergusson J. R., Shellard E. P. S., 2010, Adv. Astron., 2010, 1001.4707
Lu Y., Mo H. J., Katz N., Weinberg M. D., 2006, MNRAS, 368, 1931
Manera M. et al., 2013, MNRAS, 428, 1036
Merson A. I. et al., 2013, MNRAS, 429, 556
Monaco P., 1995, ApJ, 447, 23
Monaco P., 1997, MNRAS, 287, 753
Monaco P., Theuns T., Taffoni G., 2002, MNRAS, 331, 587 (Paper I)
Monaco P., Møller P., Fynbo J. P. U., Weidinger M., Ledoux C., Theuns T., 2005, A&A, 440, 799
Moutarde F., Alimi J.-M., Bouchet F. R., Pellat R., Ramani A., 1991, ApJ, 382, 377
Peel M. W., Battye R. A., Kay S. T., 2009, MNRAS, 397, 2189
Pierre M., Pacaud F., Juin J. B., Melin J. B., Valageas P., Clerc N., Corasaniti P., 2011, MNRAS, 414, 1732
Pollack J. E., Smith R. E., Porciani C., 2012, MNRAS, 420, 3469
Pope A. C., Szapudi I., 2008, MNRAS, 389, 766
Prada F., Klypin A. A., Cuesta A. J., Betancort-Rijo J. E., Primack J. R., 2012, MNRAS, 423, 3018
Paranjape A., Sefusatti E., Chan K. C., Desjacques V., Monaco P., Sheth R. K., 2013, MNRAS, submitted (arXiv:1305.5830)
Reed D. S., Bower R., Frenk C. S., Jenkins A., Theuns T., 2007, MNRAS, 374, 2
Ross A. J. et al., 2012, MNRAS, 424, 564
Schneider R., Salvaterra R., Ferrara A., Ciardi B., 2006, MNRAS, 369, 825
Scoccimarro R., Hui L., Manera M., Chan K. C., 2012, Phys. Rev. D, 85, 083002
Sefusatti E., Crocce M., Pueblas S., Scoccimarro R., 2006, Phys. Rev. D, 74, 023522
Sefusatti E., Crocce M., Desjacques V., 2010, MNRAS, 406, 1014
Sefusatti E., Crocce M., Desjacques V., 2012, MNRAS, 425, 2903
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
Sievers J. L. et al., 2013, preprint (arXiv e-prints, 1301.0824)
Springel V., 2005, MNRAS, 364, 1105
Springel V. et al., 2005, Nat, 435, 629
Story K. T. et al., 2012, preprint (arXiv e-prints, 1210.7231)
Taffoni G., Monaco P., Theuns T., 2002, MNRAS, 333, 623

Tassev S., Zaldarriaga M., Eisenstein D., 2013, preprint (arXiv e-prints, 1301.0322)
Tinker J., Kravtsov A. V., Klypin A. A., Abazajian K., Warren M. S., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709
Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, ApJ, 646, 881
Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2003, ApJ, 597, L9
Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2009, ApJ, 707, 354
Zheng Z., Coil A. L., Zehavi I., 2007, ApJ, 667, 760

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