Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **473,** 5098–5112 (2018)
Advance Access publication 2017 October 9

doi:10.1093/mnras/stx2616

A question of separation: disentangling tracer bias and gravitational non-linearity with counts-in-cells statistics

C. Uhlemann,[1]★ M. Feix,[2] S. Codis,[2,3,4]★ C. Pichon,[2,5] F. Bernardeau,[2,4] B. L'Huillier,[6] J. Kim,[5] S. E. Hong,[6] C. Laigle,[7] C. Park,[5] J. Shin[5] and D. Pogosyan[8]

[1]*Institute for Theoretical Physics, Utrecht University, Leuvenlaan 4, NL-3584 CE, Utrecht, the Netherlands*
[2]*CNRS, UMR 7095 & UPMC, Institut d'Astrophysique de Paris, 98 bis Boulevard Arago, F-75014, Paris, France*
[3]*Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St. George Street, Toronto, ON M5S 3H8, Canada*
[4]*CNRS & CEA, UMR 3681, Institut de Physique Théorique, F-91191, Gif-sur-Yvette, France*
[5]*Korea Institute of Advanced Study (KIAS) 85 Hoegiro, Dongdaemun-gu, Seoul, 02455, Republic of Korea*
[6]*Korea Astronomy and Space Science Institute (KASI), 776 Daedeokdae-ro, Yuseong-gu, Daejeon 34055, Republic of Korea*
[7]*Department of Physics, sub-department of Astrophysics, University of Oxford, Keble Road, Oxford OX1 3RH, UK*
[8]*Department of Physics, University of Alberta, 412 Avadh Bhatia Physics Laboratory, Edmonton, Alberta, T6G 2J1, Canada*

Accepted 2017 October 2. Received 2017 October 2; in original form 2017 May 26

ABSTRACT

Starting from a very accurate model for density-in-cells statistics of dark matter based on large deviation theory, a bias model for the tracer density in spheres is formulated. It adopts a mean bias relation based on a quadratic bias model to relate the log-densities of dark matter to those of mass-weighted dark haloes in real and redshift space. The validity of the parametrized bias model is established using a parametrization-independent extraction of the bias function. This average bias model is then combined with the dark matter PDF, neglecting any scatter around it: it nevertheless yields an excellent model for densities-in-cells statistics of mass tracers that is parametrized in terms of the underlying dark matter variance and three bias parameters. The procedure is validated on measurements of both the one- and two-point statistics of subhalo densities in the state-of-the-art Horizon Run 4 simulation showing excellent agreement for measured dark matter variance and bias parameters. Finally, it is demonstrated that this formalism allows for a joint estimation of the non-linear dark matter variance and the bias parameters using solely the statistics of subhaloes. Having verified that galaxy counts in hydrodynamical simulations sampled on a scale of 10 Mpc h^{-1} closely resemble those of subhaloes, this work provides important steps towards making theoretical predictions for density-in-cells statistics applicable to upcoming galaxy surveys like *Euclid* or *WFIRST*.

Key words: large-scale structure of Universe – cosmology: theory.

1 INTRODUCTION

Counts-in-cells statistics of galaxies have been extracted from observations in numerous works (Sheth, Mo & Saslaw 1994; Szapudi, Meiksin & Nichol 1996; Adelberger et al. 1998; Yang & Saslaw 2011; Wolk et al. 2013; Bel et al. 2016; Clerkin et al. 2017; Hurtado-Gil et al. 2017) spanning data sets from *IRAS* over SDSS to VIPERS and DES science verification. Conversely, significant theoretical progress has been made in analytically predicting the statistics of dark matter densities-in-spheres based on perturbation theory and local collapse models (Fry 1985; Balian & Schaeffer 1989; Bernardeau 1992, 1994a; Juszkiewicz, Bouchet & Colombi 1993; Munshi, Sahni & Starobinsky 1994; Bernardeau & Kofman 1995; Juszkiewicz et al. 1995; Scoccimarro & Frieman 1996; Fosalba & Gaztanaga 1998; Gaztañaga, Fosalba & Elizalde 2000; Valageas 2002a; Ohta, Kayo & Taruya 2003, providing only a non-exhaustive list of previous work), which has been recently reformulated in terms of the theory of rare events (Bernardeau 1994b; Valageas 2002b; Bernardeau, Pichon & Codis 2014; Bernardeau, Codis & Pichon 2015; Bernardeau & Reimberg 2016) with Uhlemann et al. (2016) achieving percent accuracy on the dark matter density PDF compared to state-of-the-art numerical simulations on scales of $\gtrsim$10 Mpc h^{-1}.

Such joint progress should now allow us to extract information from the mildly non-linear regime so as to efficiently improve the estimation of cosmological parameters as this formalism allows for analytical predictions in this regime. Achieving this goal

★ E-mail: cu226@cam.ac.uk (CU); codis@iap.fr (SC)

requires to relate the predictions for dark matter densities in spheres to galaxy counts which constitute biased tracers of the underlying matter field. Indeed, in addition to non-linear gravitational dynamics and the effect of redshift-space distortions, clustering analyses of large-scale structure (LSS) are hampered by the fact that astronomical objects such as galaxies do not trivially trace the underlying dark matter distribution (see Desjacques, Jeong & Schmidt 2016, for a recent review). This problem has been known for a long time (e.g. Abell 1958; Dressler 1980; Bahcall & Soneira 1983; Kaiser 1984; Coles 1986), and was subsequently confirmed in cosmological simulations demonstrating that haloes and galaxies are biased with respect to dark matter (e.g. Cen & Ostriker 1992; Kauffmann, Nusser & Steinmetz 1997; Blanton et al. 1999; Somerville et al. 2000). Since then several approaches have been pursued to accurately model these biasing relations. One main complication is that galaxy bias is generally a non-local and stochastic function of the dark matter field due to the varied physical processes partaking in galaxy formation (Dekel & Lahav 1999; Scoccimarro 2000). Yet, smoothing the matter density fields over sufficiently large scales mitigates the effects of non-locality and allows a sound description in terms of local bias expansions (e.g. Fry & Gaztanaga 1993) which aim at absorbing the underlying physics into a finite set of parameters. Later work has put such perturbative approaches on to firmer grounds by including non-local contributions and providing a consistent theoretical framework for the statistics of biased LSS tracers (e.g. Baldauf et al. 2011; Matsubara 2011; Schmidt, Jeong & Desjacques 2013; Senatore 2015; Porto 2016). Galaxies are believed to form inside the potential wells of dark matter haloes whose biasing properties can be systematically studied in numerical simulations or by means of analytic methods. Assuming that dark matter haloes are associated with peaks of the initial density field, the peak approach (Kaiser 1984; Bardeen et al. 1986) provides a non-perturbative model for biased populations and reasonably agrees with the abundance and the linear bias of virialized haloes. Concerning non-linearity as well as its dependence on other parameters like halo mass and scale, the bias of dark matter haloes is well approximated within the halo model (e.g. Mo & White 1996; Sheth & Tormen 1999; Cooray & Sheth 2002) based on the excursion set approach (Bond et al. 1991). Its relation to galaxies is typically quantified by combining cosmological N-body simulations with semianalytic models of galaxy formation (Kauffmann et al. 1999; Berlind & Weinberg 2002; Baugh 2006; Mo, van den Bosch & White 2010).

This paper will start from the dark matter side and make one crucial step towards reality by considering subhaloes, as the host of and proxies for galaxies and dark matter tracers. Such subhaloes can be extracted reliably from large cosmological simulations such as Horizon Run 4 (HR4; Kim et al. 2015) that contain enough statistics to extract continuous PDFs. Note that the focus is on the issue of biasing for the PDF, such that it is not so essential which tracers are chosen. However, the link between subhaloes and galaxies will also be discussed based on recent results from Horizon-AGN (Dubois et al. 2014a), a cosmological hydrodynamical simulation that captures the evolutionary trends of observed galaxies over the lifetime of the Universe. The relation between continuous PDFs and discrete galaxy counts is briefly addressed, see Bel et al. (2016) for a recent and more exhaustive consideration of observational effects such as masking in galaxy surveys.

In general, biasing is a notoriously challenging problem that requires the formulation of non-local and stochastic relationships between dark matter and tracer densities. This paper will however show that for the purpose of obtaining the one- and two-point statistics of tracer densities in $\sim 10\,\mathrm{Mpc}\,h^{-1}$ spheres, a mean local relationship (hence neglecting the scatter altogether) is enough to obtain predictions that are as accurate as the underlying statistics of dark matter densities. It will also show that the joint analysis of one- and two-cells counts allows us to lift the degeneracy between bias and dark matter variance, providing a key step towards making count-in-cells statistics applicable to upcoming galaxy surveys like *Euclid* or LSST, for the purpose of extracting cosmological parameters in the mildly non-linear regime.

This paper is organized as follows. Section 2 recaps the results presented in Uhlemann et al. (2016) for the dark matter density PDF. Section 3 turns to the bias between dark matter and tracer densities. After describing the HR4 simulation and the halo identification scheme, an analytic bias model is formulated and compared to measurements from the simulation using scatter plots and a parametrization-independent bias extraction. Based on Horizon-AGN, the similarity of the mean bias relations for galaxies and haloes is established and the influence of the scatter is assessed. Section 4 combines the bias model with the one-point dark matter PDF and two-point sphere bias to obtain the one-point halo PDF and two-point halo bias and establishes its accuracy against simulations. Section 5 implements this formalism to estimate simultaneously variance and biasing, and discusses applications and extensions. Finally, Section 6 concludes. Appendix B compares the large deviation statistics (LDS) prediction to the lognormal models. Appendix C shows perturbatively why the joint analysis of the one- and two-point statistics breaks the degeneracy on tracer bias and dark matter variance. Appendix D describes the hydrodynamical simulation Horizon-AGN.

2 THE DARK MATTER DENSITY PDF

As shown in Uhlemann et al. (2016), the PDF for dark matter densities ρ_m within a sphere of radius R at redshift z, valid in the mildly non-linear regime, can be obtained from LDS and is expressed as

$$\mathcal{P}_R(\rho_\mathrm{m}) = \sqrt{\frac{\Psi_R''(\rho_\mathrm{m}) + \Psi_R'(\rho_\mathrm{m})/\rho_\mathrm{m}}{2\pi\sigma_\mu^2}}\,\exp\left(-\frac{\Psi_R(\rho_\mathrm{m})}{\sigma_\mu^2}\right), \tag{1}$$

where the prime denotes a derivative with respect to ρ_m and

$$\Psi_R(\rho_\mathrm{m}) = \frac{\tau_\mathrm{SC}^2(\rho_\mathrm{m})\sigma_\mathrm{L}^2(R)}{2\sigma_\mathrm{L}^2(R\rho_\mathrm{m}^{1/3})}\,. \tag{2}$$

Here, $\sigma_\mu \equiv \sigma_\mu(R, z)$ is the non-linear variance of the log-density (because the formula has been derived from an analytic approximation based on the log-density $\mu_\mathrm{m} = \log\rho_\mathrm{m}$) while σ_L is the linear variance determined from the initial power spectrum P_L using the Fourier transform of the spherical top-hat filter W

$$\sigma_\mathrm{L}(r) = \int \mathrm{d}^3k\,(2\pi)^{-3}P_\mathrm{L}(k)W(kr)^2\,. \tag{3}$$

$\tau_\mathrm{SC}(\rho_\mathrm{m})$ is the linear density contrast averaged within the Lagrangian radius $r = R\rho_\mathrm{m}^{1/3}$ which can be mapped to the non-linearly evolved density ρ_m within radius R using the spherical collapse model. For this, an accurate approximation has been introduced by Bernardeau (1992) according to

$$\rho_\mathrm{SC}(\tau) = (1-\tau/\nu)^{-\nu} \;\Leftrightarrow\; \tau_\mathrm{SC}(\rho) = \nu(1-\rho^{-1/\nu})\,, \tag{4}$$

where the parameter ν characterises the dynamics of spherical collapse. Here, we choose $\nu = 21/13$ to exactly match the high-redshift skewness obtained from perturbation theory (Bernardeau

et al. 2014). To ensure a unit mean density and the correct normalization of the PDF, one can simply evaluate the PDF obtained from equation (1) according to

$$\hat{\mathcal{P}}_R(\rho_\mathrm{m}) = \mathcal{P}_R\left(\rho_\mathrm{m}\frac{\langle\rho_\mathrm{m}\rangle}{\langle 1\rangle}\right)\cdot\frac{\langle\rho_\mathrm{m}\rangle}{\langle 1\rangle^2}\,, \tag{5}$$

with the shorthand notation $\langle f(\rho_\mathrm{m})\rangle = \int_0^\infty \mathrm{d}\rho_\mathrm{m}\, f(\rho_\mathrm{m})\mathcal{P}_R(\rho_\mathrm{m})$. This step is necessary as equation (1) ensures the correct tree-level cumulants of order 3 and above, the right non-linear variance of μ_m and zero mean for μ_m. If instead, one wants ρ_m to have unit mean, it is necessary to correct for the non-zero value of the mean of μ_m using equation (5).

Following Codis, Bernardeau & Pichon (2016b), Uhlemann et al. (2017b), the two-point PDF of the matter density reads in the large-separation limit

$$\mathcal{P}_R\left(\rho_\mathrm{m},\rho_\mathrm{m}'\right) = \mathcal{P}_R\left(\rho_\mathrm{m}\right)\mathcal{P}_R\left(\rho_\mathrm{m}'\right)\left[1+\xi_{\circ,\mathrm{m}}(r)b_\circ\left(\rho_\mathrm{m}\right)b_\circ\left(\rho_\mathrm{m}'\right)\right], \tag{6}$$

where $r > 2R$ is the separation between two spheres of radius R and densities ρ_m and ρ_m'. The validity of this approximation in the large-separation regime has been demonstrated in fig. 10 of Uhlemann et al. (2017b). The sphere bias encodes the excess correlation (with respect to the average sphere correlation $\xi_{\circ,\mathrm{m}}$) induced by a density ρ_m at separation r and is defined as

$$b_\circ(\rho_\mathrm{m}) = \frac{\langle\rho_\mathrm{m}'|\rho_\mathrm{m};r\rangle - 1}{\xi_{\circ,\mathrm{m}}(r)}\,,\ \xi_{\circ,\mathrm{m}}(r) = \langle\rho_\mathrm{m}\rho_\mathrm{m}';r\rangle - 1\,. \tag{7}$$

At large separation, the sphere bias becomes independent of separation r and is obtained with a much higher accuracy than the approximation for the full two-point PDF (6). It can be computed using the large-deviation principle and is well approximated by (Bernardeau 1996; Abbas & Sheth 2007; Codis et al. 2016b; Uhlemann et al. 2017b)

$$b_\circ(\rho_\mathrm{m}) = \frac{\tau_\mathrm{SC}(\rho_\mathrm{m})\sigma_\mathrm{L}^2(R)}{\sigma_\mathrm{L}^2(R\rho_\mathrm{m}^{1/3})\sigma_\mu^2}\,, \tag{8}$$

with once again a normalization according to

$$\hat{b}_\circ(\rho_\mathrm{m}) = \frac{b_\circ(\rho_\mathrm{m}) - \langle b_\circ(\rho_\mathrm{m})\rangle}{\langle(\rho_\mathrm{m}-1)b_\circ(\rho_\mathrm{m})\rangle}\,. \tag{9}$$

3 BIAS BETWEEN MATTER AND TRACER DENSITIES

Let us now turn to biased tracers. Section 3.1 will first introduce the HR4 simulation while Section 3.2 describes the theoretical models for tracer (galaxy and halo) bias.

3.1 Biased tracers in Horizon Run 4 simulation

3.1.1 Halo identification

The HR4 simulation (Kim et al. 2015) is a massive N-body simulation, evolving 6300^3 particles in a $3.15\,h^{-1}$ Gpc box using the GOTPM TREEPM code (Dubinski et al. 2004). It assumes a WMAP-5 cosmology, with $(\Omega_\mathrm{m}, \Omega_\Lambda, \Omega_\mathrm{b}, h, \sigma_8, n_\mathrm{s}) = (0.26, 0.74, 0.044, 0.72, 0.79, 0.96)$, yielding a particle mass of $9\times10^9\,h^{-1}\,\mathrm{M}_\odot$. The initial conditions were generated at $z = 100$ using the second-order Lagrangian perturbation theory, which ensures accurate power spectrum and halo mass function at redshift 0 (L'Huillier, Park & Kim 2014). The haloes were detected using ordinary parallel friends-of-friends (OPFOF, Kim & Park 2006), a massively parallel implementation of the friends-of-friends (FoF) algorithm, using a canonical linking length of 0.2 mean particle separations. Subhaloes were detected by the physically self-bound (PSB) algorithm (Kim & Park 2006), which finds the density peaks within each FoF halo, removes unbound particles, similarly to the SUBFIND halo finder, and additionally truncates the subhaloes to their tidal radius. All subhaloes with more than 30 particles were considered, yielding a masses from $2.7\times10^{11}\,h^{-1}\,\mathrm{M}_\odot$ to $4.2\times10^{15}\,h^{-1}\,\mathrm{M}_\odot$.

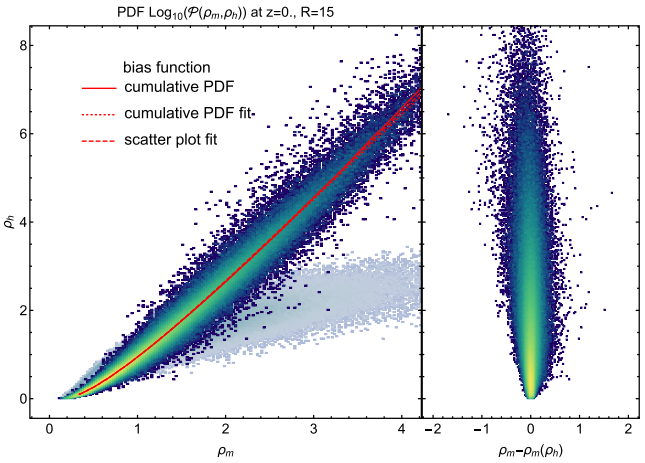

Figure 1. (Left-hand panel) Density scatter plot of the halo density ρ_h with mass-weighting (blue-green region) and number-weighting (grey region) versus the dark matter density ρ_m for radius $R = 15\,\mathrm{Mpc}\,h^{-1}$ at redshift $z = 0$. The figure also shows the the best-fitting quadratic bias model for the log-density obtained from a fit to the CDF bias function and the scatter plot (dotted and dashed line, respectively) which almost perfectly agrees with the parametrization-independent bias obtained from the CDF (red line). Note that for mass-weighted halo densities, the scatter is reduced significantly compared to number-weighted halo densities. (Right-hand panel) Residual scatter around the quadratic fit to the CDF bias function which is uniform and symmetric.

3.1.2 Weighting of halo densities

Following the observations made in Seljak, Hamaus & Desjacques (2009), Hamaus et al. (2010) and Jee et al. (2012) (Jee12 hereafter), let us consider a halo density with mass-weighting (instead of number-weighting) because this makes the bias relation much tighter and considerably reduces the scatter which is illustrated in Fig. 1. For a discussion of a local polynomial bias relation for number-weighted halo densities and their running with the smoothing scale we refer to Manera & Gaztañaga (2011), Angulo, Baugh & Lacey (2008), Chan & Scoccimarro (2012) and Paranjape et al. (2013). The reduction in scatter can be understood by the intuition that mass-weighted halo densities resemble the overall dark matter density much more closely than halo number does. Note, however, that the mass-weighted densities of subhaloes are expected to be very similar to the mass weighted density of haloes (with no substructure) as the mass is almost preserved from haloes to subhaloes. This paper considers subhaloes as defined in Section 3.1 because they can be related to galaxies using abundance matching (Kravtsov et al. 2004; Vale & Ostriker 2004), see Section 3.2.5.

3.2 Bias models: mean bias relations and their scatter

Uhlemann et al. (2016) showed that the model for the PDF of the dark matter density field $\hat{P}_R(\rho_\mathrm{m}|\sigma_\mu)$ with the variance of the log-density $\sigma_\mu(R)$ as a driving parameter was accurate at the percent level for variances $\sigma \lesssim 0.5$. Hence, the question of how to obtain

a similarly accurate model for the PDF of the density field of a biased mass tracer boils down to successfully describing the effective bias relation between dark matter densities in spheres and the corresponding densities in spheres of their tracers. For simplicity, this bias model is formulated between dark matter and halo (or galaxy) densities for spheres of identical radii, so from now on on $\rho_{\rm m}(\rho_{\rm h})$ stands for $\rho_{{\rm m},R}(\rho_{{\rm h},R})$. While in general one would expect that the full joint PDF of dark matter and tracer densities is needed, including the scatter, it is shown in what follows that an accurate mean bias relation is enough to obtain an excellent model for the biased tracer PDF. This is in the spirit of LDS, that has been previously applied to argue that the mean local gravitational evolution given by spherical collapse is good enough to predict the dark matter PDF at fixed radius at percent accuracy.[1]

3.2.1 Polynomial bias model in log-densities

In order to map the dark matter PDF to the halo PDF, let us rely on an 'inverse' bias model $\rho_{\rm m}(\rho_{\rm h})$ writing the dark matter density as a function of the halo density which, according to Jee12, has a better performance than the 'forward' bias model $\rho_{\rm h}(\rho_{\rm m})$. These bias parameters characterize the inverse relation and in particular our linear bias will typically have values around 1/2 signalling positive linear forward bias around 2. Again, following Jee12, let us use a quadratic model for the log-densities $\mu = \log\rho$ (rather than for the densities) which reads

$$\mu_{\rm m} = \sum_{n=0}^{n_{\rm max}} b_n \mu_{\rm h}^n\,,\ n_{\rm max} = 2\,. \tag{10}$$

It was checked that the higher order bias parameters are negligible, $|b_3| < 0.002$ for all redshifts and radii considered here, and lead to very minor improvements of the quality of fit that do not warrant the use of this additional parameter. Note that, since the offset b_0 is additive in the log-densities, it ensures a multiplicative renormalization for the density,[2] which is preferable according to an analytical result of Frusciante & Sheth (2012) that has been obtained from a lognormal mapping. Jee12 emphasize that the reason why equation (10) can be approximated by a linear bias model for the density fluctuations $\delta_{\rm h} = \hat{b}_1\delta_{\rm m}$ on large scales is that the ranges of log-densities $\mu_{\rm h}$ and $\mu_{\rm m}$ become small and not because the bias relation itself becomes linear. This is particular relevant here when focusing on the tails of the distribution of densities and hence the regime where linear bias is not sufficient.

3.2.2 Parametrisation-independent inference of bias

Following the idea of Sigad, Branchini & Dekel (2000), Szapudi & Pan (2004), a direct way to obtain the mean bias relation is to use the properties of the cumulative distribution functions (CDFs), defined as $\mathcal{C}(\rho) = \int_0^\rho {\rm d}\rho'\mathcal{P}(\rho')$, so that

$$\mathcal{C}_{\rm m}(\rho_{\rm m}) = \mathcal{C}_{\rm h}(\rho_{\rm h}) \ \Rightarrow\ \rho_{\rm m}(\rho_{\rm h}) = \mathcal{C}_{\rm m}^{-1}(\mathcal{C}_{\rm h}(\rho_{\rm h}))\,. \tag{11}$$

This parametrization-independent bias extraction is used to verify the accuracy of the polynomial log-bias model, equation (10), as described below.

Table 1. Collection of simulation results for different radii R (Mpc h^{-1}) and redshifts z. The measured non-linear variances σ of the log-density $\mu = \log\rho$ and the correlation ξ of the density ρ at separation $r = 30$ Mpc h^{-1} of both dark matter (m) and haloes (h) in real space (upper part) and redshift space (lower part) along with the bias parameters obtained from fitting the quadratic model from equation (10) to the bias function obtained from the CDF according to equation (11).

Param		Variance		Correlation			Bias	
z	R	$\sigma_{\mu,{\rm m}}$	$\sigma_{\mu,{\rm h}}$	$\xi_{\rho,{\rm m}}$	$\xi_{\rho,{\rm h}}$	b_0	b_1	b_2
0	10	0.613	1.276	0.041	0.093	0.068	0.604	0.058
0	15	0.475	0.855	0.043	0.099	0.036	0.618	0.058
1	10	0.411	1.006	0.015	0.067	0.054	0.460	0.055
1	15	0.310	0.692	0.016	0.071	0.028	0.473	0.055
z	R	$\sigma^z_{\mu,{\rm m}}$	$\sigma^z_{\mu,{\rm h}}$	$\xi^z_{\rho,{\rm m}}$	$\xi^z_{\rho,{\rm h}}$	b^z_0	b^z_1	b^z_2
0	10	0.614	1.286	0.041	0.115	0.086	0.566	0.052
0	15	0.476	0.911	0.043	0.122	0.048	0.574	0.052

3.2.3 Density scatter plots from numerical simulation

Fig. 1 presents a scatter plot showing $\rho_{\rm h}$ as a function of $\rho_{\rm m}$ for redshift $z = 0$ and radius $R = 15$ Mpc h^{-1} in order to assess how well bias models characterize the halo density bias. The lines correspond to the mean bias obtained in a parametrization-independent way from the CDF method (red line) and fits based on a quadratic bias model for the log-densities (dotted and dashed red line) according to equation (10). The corresponding values of the best-fitting bias parameters are given in Table 1 for different redshifts and radii. The second-order bias model for the logarithmic densities based on equation (10) agrees almost perfectly with the parametrization-independent way of inferring bias using CDFs as in equation (11) and matches simulation results very well, as has been observed in Jee12 for a wide range of mass cuts, smoothing lengths and redshifts. Indeed, differences in the fits are almost imperceptible to the eye and at the sub-percent level throughout, except for the extreme low- and high-density tail, and the residual scatter around the mean polynomial log-bias model is very symmetric and uniform. This has to be contrasted with a quadratic model in the mass-weighted halo densities that can be shown to have a clear residual skewness and to be significantly less accurate (residuals of about 2 per cent between $\rho \in [0.2, 3]$, increasing more steeply in the tails). Since the mean bias relation is used to map the PDFs, having an even scatter around the mean relation is advantageous to mitigate possible effects of the scatter. Hence, the polynomial bias model for the log-densities shall be used. Given the smallness of the prefactor b_2 for the quadratic term compared to b_1, one might wonder whether a linear model in log-densities is sufficient. Unfortunately, the linear log-density model fit shows residuals of order 5 per cent for almost all densities when compared to the parametrization-independent bias function and hence would degrade the accuracy of the halo PDF substantially. This is why the quadratic bias model for the log-densities will be used for the remainder of the text.

Furthermore, Fig. 2 presents a scatter plot for the halo density determined in redshift space $\rho_{{\rm h},z}$. This was done by converting the

[1] The large-deviation principle states that the statistics is dominated by the path that minimises the 'action' – or in our case the rate function – in order to maximize the probability. This most likely path or dynamics can be decomposed into a gravitational part, given by the spherical collapse, and an astrophysical part, given by the mean bias relation.

[2] When expanding the quadratic bias model for log-densities in the halo density contrast $\rho_{\rm h} = 1 + \delta_{\rm h}$ one obtains

$$\rho_{\rm m} = \exp(b_0)\left(1 + b_1\delta_{\rm h} + \left[\frac{1}{2}(b_1-1)b_1 + b_2\right]\delta_{\rm h}^2 + \mathcal{O}(\delta_{\rm h}^3)\right)\,.$$

Interestingly, for the similar radii $R_1 = 10$, $R_2 = 15\,{\rm Mpc}\,h^{-1}$ one finds identical b_2 and $\exp(b_0)b_1$ while b_0 and b_1 differ.

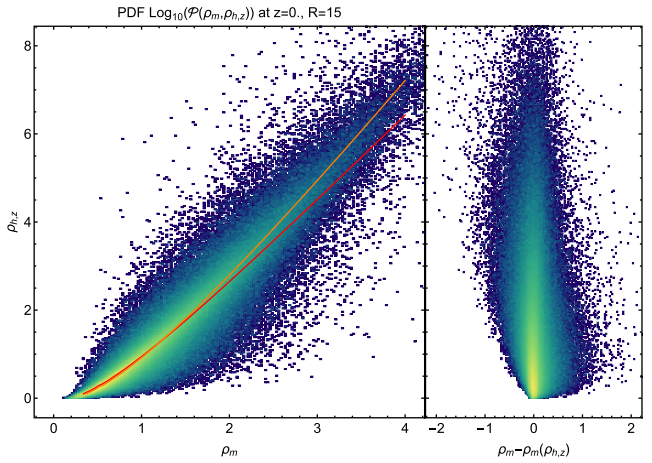

Figure 2. (Left-hand panel) Density scatter plot of the halo density $\rho_{\mathrm{h},z}$ in redshift space with mass-weighting versus the dark matter density ρ_{m} for radius $R = 15\,\mathrm{Mpc}\,h^{-1}$ at redshift $z = 0$. The figure also shows the the best-fitting quadratic bias model for the log-density obtained from a fit to the CDF bias function in real space (red line) and redshift space (orange line). (Right-hand panel) Residual scatter around the quadratic fit to the CDF bias function which is uniform and symmetric (though the dispersion is somewhat larger than that of Fig. 1).

comoving halo-positions $\boldsymbol{r}$ to the redshift-space ones $\boldsymbol{s}$ by shifting them along the fictitious line of sight (chosen in x-direction here) according to their peculiar velocity along that direction

$$\boldsymbol{s} = \boldsymbol{r} + \frac{1+z}{H(z)} v_x \hat{\boldsymbol{x}} \,. \tag{12}$$

As was done in real space, a parametrization-independent extraction of the mean bias relation was used as a complement to the polynomial bias model in the log-densities (10) for mass-weighted halo densities in redshift space, thereby extending the results of Jee12. When comparing the scatter plot from redshift space to its real space analogue (shown in Fig. 1), one can clearly see a enhanced scatter around the mean bias relation. Yet, this extra scatter does not directly translate into inaccuracies of the PDF, as shown in Fig. 5.

3.2.4 *From densities to counts in cells*

When considering the statistics of discrete tracers of the density field, such as effective dark matter particles in simulations or haloes identified therein, one needs to account for finite sampling effects in the cells. The sampling process determines how a discrete cell-count N in a sphere of radius R arises from the value of the underlying average density ρ and can be written as a convolution

$$\mathcal{P}(N) = \int \mathcal{P}(N|\rho)\mathcal{P}(\rho)\mathrm{d}\rho \,, \tag{13}$$

where $\mathcal{P}(N|\rho)$ is the sampling conditional probability of finding a cell-count N given a density field value ρ. In order to predict the PDF of the random variable N from the PDF of the underlying field ρ, one needs an expression for the sampling conditional probability. The most widespread of such schemes is local Poisson sampling (see e.g. Bel et al. 2016; Repp & Szapudi 2017a), where

$$\mathcal{P}(N|\rho) = \frac{(\bar{N}\rho)^N}{N!} \exp(-\bar{N}\rho) \,, \tag{14}$$

with $\bar{N}$ as the mean number of objects per cell and $\bar{N}\rho$ the expectation value of the number count given average cell density ρ. For a discussion of how to determine model parameters from data that constitutes a discrete realization of an underlying continuous density, we refer to Fry et al. (2011a) and Bel et al. (2016). Roughly, one can expect that for smoothing scales which ensure that there are enough tracers per cell, discreteness effects are mitigated. For densities around the mean density this is fulfilled for sphere radii larger than the mean separation of tracers. In the HR4 simulation, the mean separation of subhaloes is around $d \simeq 4.5\,\mathrm{Mpc}\,h^{-1}$ between redshift $z = 1$ and $z = 0$ and hence smaller than the sphere radii considered. For large densities where one has more tracers per cell, discreteness effects are expected to be smallest while they will be most relevant for very low densities. Typical galaxy densities for current and forthcoming galaxy surveys such as BOSS, DESI and *Euclid* are of order a few $10^{-4}(\mathrm{Mpc}\,h^{-1})^{-3}$ which yields an order 1–10 tracers per sphere of radius $15\,\mathrm{Mpc}\,h^{-1}$ at average density. Note that, for tracer densities that are not number-weighted, but weighted by mass or luminosity, one technically does not simply count objects, but the formalism remains similar.

Besides finite sampling effects in the cells, there are also sampling effects due to the finite number of cells available in the cosmic volume, see appendix C in Codis et al. (2016b).

3.2.5 *Applicability to galaxies*

In order to check to which extent our formalism developed for haloes will be applicable to galaxies, mass-weighted densities of haloes, galaxies and luminosity-weighted densities of galaxies were extracted from the Horizon-AGN simulation. Horizon-AGN is a full-physics hydrodynamical simulation in a cosmological volume with side length $L_{\mathrm{box}} = 100\,\mathrm{Mpc}\,h^{-1}$ (Dubois et al. 2014a) and in the latest generation of state-of-the-art simulations. Dark matter and mass-weighted subhalo densities in 125 non-overlapping spheres of radius $R = 10\,\mathrm{Mpc}\,h^{-1}$ are extracted from the simulated box at $z = 1$. In order to mimic observational measurements, mass- and luminosity-weighted (in the K_{s} band) galaxy densities are extracted from the simulated light-cone in a redshift range around $z = 1$. All the galaxies with a mass $M > 10^{9.5}\,\mathrm{M}_{\odot}$ were included. While photometric surveys can quite easily reach such limit at $z \approx 1$, spectroscopic surveys are generally a sparser. Realistic galaxy luminosities have been computed in post-processing using spectral synthesis, and galaxy stellar masses have been computed from photometry using SED-fitting, as usually done in observational data sets, which naturally allows us to incorporate realistic errors (Laigle et al. in preparation; see Appendix D for more details). We didn't find any qualitative difference between the mean bias relations for galaxies and haloes. Indeed, Fig. 3 displays the CDF of dark matter, mass-weighted subhaloes as well as mass- and luminosity-weighted galaxies together with the corresponding scatter plot. The blue, green and orange lines and points correspond to mass-weighted subhaloes, galaxies and luminosity-weighted galaxies, respectively, and are practically undistinguishable given the statistics we have,[3] although the scatter of the galaxies is significantly increased compared to haloes. This is a very promising result that motivates the use of mass-weighted halo density fields in this work. A thorough study of galaxy and halo bias in Horizon-AGN will be the topic of a forthcoming paper (Chisari et al., in preparation). Note that, if one weights the

[3] Note that the box size is rather small, so the error bars are likely underestimated due to missing large-scale modes. Even if the small differences between the three curves were statistically significant, their agreement is impressively good.

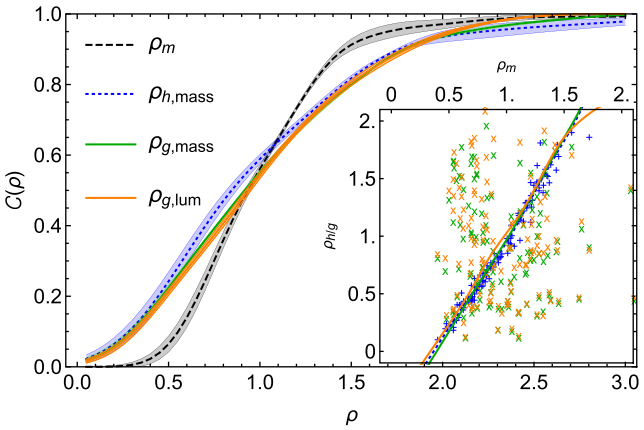

Figure 3. The CDF for densities in spheres of radius $R = 10\,\mathrm{Mpc}\,h^{-1}$ as measured from Horizon-AGN at redshift $z = 1$ for dark matter (dashed black line), mass-weighted subhaloes (dotted blue line), mass-weighted galaxies (solid green line) and luminosity-weighted galaxies (solid orange line). The shaded areas show an estimation of the error based on 5 subsamples. The inset shows the scatter plot comparing mass- and luminosity-weighted galaxies (green and orange crosses) to mass-weighted subhaloes (blue plusses) including the bias function extracted from the CDF method from equation (11).

galaxy densities with the mass of the host subhalo, the resemblance is even closer and the scatter reduced. But in practice this would require both measuring the stellar masses (or luminosities) of the galaxies and relating them to the masses of the host subhaloes. The accuracy of the former is limited by the error on galaxy mass which is expected to be a function of the mass and redshift. At low redshift ($z < 1$), the observed galaxy mass is generally underestimated compared to the intrinsic one and in general one can have a discrepancy up to $\Delta(\log M_\mathrm{g}) \simeq 15$ per cent depending on the quality of the spectroscopy or photometry available to estimate the stellar mass (see e.g. Pforr, Maraston & Tonini 2012; Mobasher et al. 2015, Laigle et al. in preparation). When adding a Gaussian noise of this size to the measured halo masses, as explicitly checked at $z = 0$ for the radii $R = 10$, $15\,\mathrm{Mpc}\,h^{-1}$, the corresponding PDFs of the mass-weighted halo densities remain almost unchanged except for their deep tails. The best-fitting bias parameters change only marginally, with the linear and largest bias parameter b_1 being most robust (sub-percent difference) and larger effects on the relatively small bias-renormalization b_0 (5–7 per cent difference) and the quadratic bias b_2 (2–4 per cent difference). For relating galaxy mass to halo mass, one can then use techniques based on subhalo abundance matching (SHAM; Behroozi, Conroy & Wechsler 2010) or its extensions (see e.g. Yang et al. 2012; Kulier & Ostriker 2015), which are very close in spirit to the modelling of bias used here and typically give an error of a similar size than the mass determination, at least for large halo masses. The same idea can be applied to galaxy luminosities (see e.g. Vale & Ostriker 2004, 2006; Cooray & Milosavljević 2005) which can be measured much more reliably than galaxy masses. Very recently, Moster, Naab & White (2017) presented an empirical model for galaxy formation finding that average star formation and accretion rates are in good agreement with models following an abundance matching strategy. One can also determine the galaxy-halo connection, in particular the stellar-to-halo mass ratio, from a joint lensing and clustering analysis of observations (as done in Coupon et al. 2015; Zu & Mandelbaum 2015) when using the halo occupation distribution (HOD) framework that assumes that the number of galaxies per halo is solely a function of halo mass, split into central and satellite contributions.

4 THE BIASED TRACER DENSITY PDF

Having established the accuracy of the bias model, let us now combine it with the one-point dark matter PDF and two-point sphere bias to obtain the one-point halo PDF and two-point halo bias. The accuracy of the analytical predictions for one- and two-point statistics will be checked against the simulation. In Appendix B, the analytical model for the halo PDF is compared to phenomenological reconstructions based on lognormal distributions and their extensions through cumulant expansions.

4.1 Mapping to the tracer PDF with the mean bias relation

The halo density PDF, $\mathcal{P}_\mathrm{h}$, can be generally written as a convolution of the dark matter PDF $\mathcal{P}_\mathrm{m}$ and the conditional PDF of finding a certain halo density given a dark matter density

$$\mathcal{P}_\mathrm{h}(\rho_\mathrm{h}) = \int \mathrm{d}\rho_\mathrm{m}\, \mathcal{P}_\mathrm{bias}(\rho_\mathrm{h}|\rho_\mathrm{m})\, \mathcal{P}_\mathrm{m}(\rho_\mathrm{m}), \tag{15}$$

where the conditional PDF $\mathcal{P}_\mathrm{bias}(\rho_\mathrm{h}|\rho_\mathrm{m})$ depends on the details of halo formation and its associated parameters such as, e.g. halo mass, smoothing scales and redshift, but also includes stochasticity which results from an incomplete understanding of the formation process (e.g. Dekel & Lahav 1999). One could attempt to model the joint PDF with the help of simulated and observed data sets in the spirit of the halo model of galaxy clustering (e.g. Berlind & Weinberg 2002; Cooray & Sheth 2002). Here, the scatter around the mean relation between ρ_m and ρ_h will be neglected: this none the less leads to an excellent model for the halo PDF provided the underlying bias model is appropriate. Equipped with a bias model for the mean relation $\rho_\mathrm{m}(\rho_\mathrm{h})$, the halo PDF $\mathcal{P}_\mathrm{h}$ is now obtained from the dark matter PDF $\mathcal{P}_\mathrm{m}$ in equation (1) by conservation of probability

$$\mathcal{P}_\mathrm{h}(\rho_\mathrm{h}) = \mathcal{P}_\mathrm{m}(\rho_\mathrm{m}(\rho_\mathrm{h}))\, |\mathrm{d}\rho_\mathrm{m}/\mathrm{d}\rho_\mathrm{h}|\ , \tag{16}$$

where it is assumed that $\rho_\mathrm{m}(\rho_\mathrm{h})$ is a strictly monotonic function. Using equation (6), the halo two-point PDF can eventually be written down as

$$\mathcal{P}_\mathrm{h}(\rho_\mathrm{h}, \rho_\mathrm{h}'; r) = \mathcal{P}_\mathrm{h}(\rho_\mathrm{h})\mathcal{P}_\mathrm{h}(\rho_\mathrm{h}') \times \left[1 + \xi_{\circ,\mathrm{m}}(r) b_{\circ,\mathrm{m}}(\rho_\mathrm{m}(\rho_\mathrm{h})) b_{\circ,\mathrm{m}}(\rho_\mathrm{m}'(\rho_\mathrm{h}'))\right]\ , \tag{17}$$

where $\xi_\circ(r)$ denotes the correlation function of spheres at separation r. In the remainder of the text, we will drop the separation r as an argument, later it will be fixed to $r = 30\,\mathrm{Mpc}\,h^{-1}$. We expect that the accuracy of the large-separation approximation for haloes (17) is similar to the one for dark matter considered in Uhlemann et al. (2017b). One can then define the modulation of the two-point correlation function, the sphere bias $b_\circ$ for haloes from the result for dark matter given in equation (8)

$$b_{\circ,\mathrm{h}}(\rho_\mathrm{h}) = b_{\circ,\mathrm{m}}(\rho_\mathrm{m}(\rho_\mathrm{h}))\, \sqrt{\xi_{\circ,\mathrm{m}}/\xi_{\circ,\mathrm{h}}}\,, \tag{18}$$

where the ratio of correlation functions is given by

$$\sqrt{\xi_{\circ,\mathrm{h}}/\xi_{\circ,\mathrm{m}}} = \left\langle \rho_\mathrm{h}(\rho_\mathrm{m}) b_{\circ,\mathrm{m}}(\rho_\mathrm{m}) \right\rangle, \tag{19}$$

and can be approximated by expanding the log-bias relation to first order to obtain $\sqrt{\xi_{\circ,\mathrm{m}}/\xi_{\circ,\mathrm{h}}} \simeq \exp(b_0) b_1$.

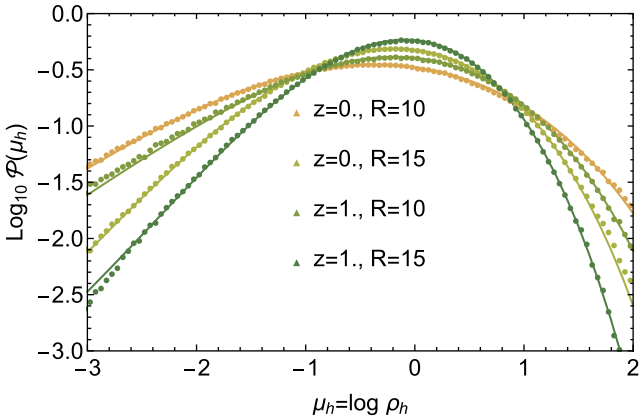

Figure 4. Logarithmic view on the halo PDF as measured from the simulation in real space (data points) at redshifts $z = 0$, 1 for radii $R = 10$, 15 Mpc h^{-1} and analytically predicted (lines) using the measured dark matter variance and bias parameters given in the upper part of Table 1.

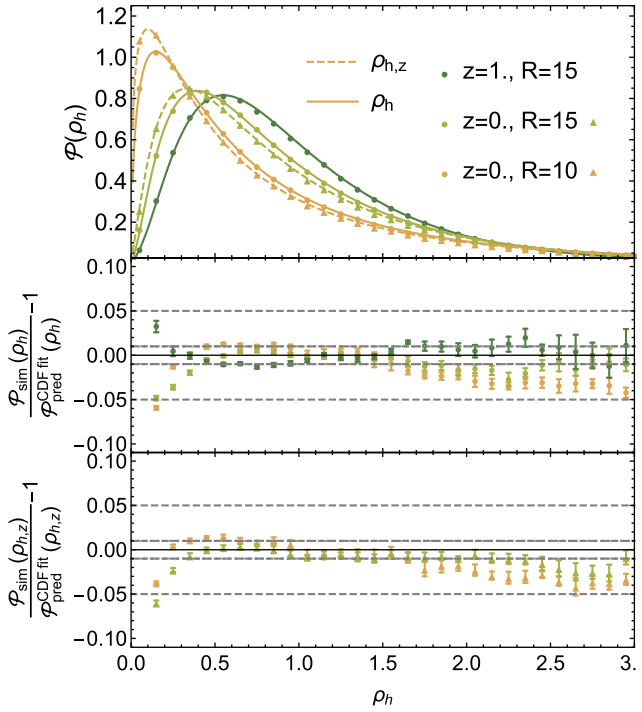

Figure 5. (Top panel) Halo mass-density PDFs $\mathcal{P}_\mathrm{h}$ for measurements based on halo catalogues in real space (points) and redshift space (triangles). Shown are results for the quadratic bias for the log-densities models in real space (solid lines) and redshift space (dashed lines) with fit values according to Table 1. (Middle and bottom panel) The corresponding residuals in real space (middle) and redshift space (bottom).

4.2 Checking the accuracy of halo PDF against simulations

Figs 4 and 5 show the result of the halo-PDF obtained from (16) using the measured variance of the dark matter log-density and the best-fitting bias parameters for the bias model for the log-densities up to second order reported in Table 1. The prediction for the halo PDF clearly matches the data, presenting residuals at the percent level in a wide range of halo densities from 0.2 to 3, in both real and redshift space. This should be contrasted to the log-normal PDF family discussed in Appendix B. This is very encouraging given the level of non-linearities involved in halo formation. The

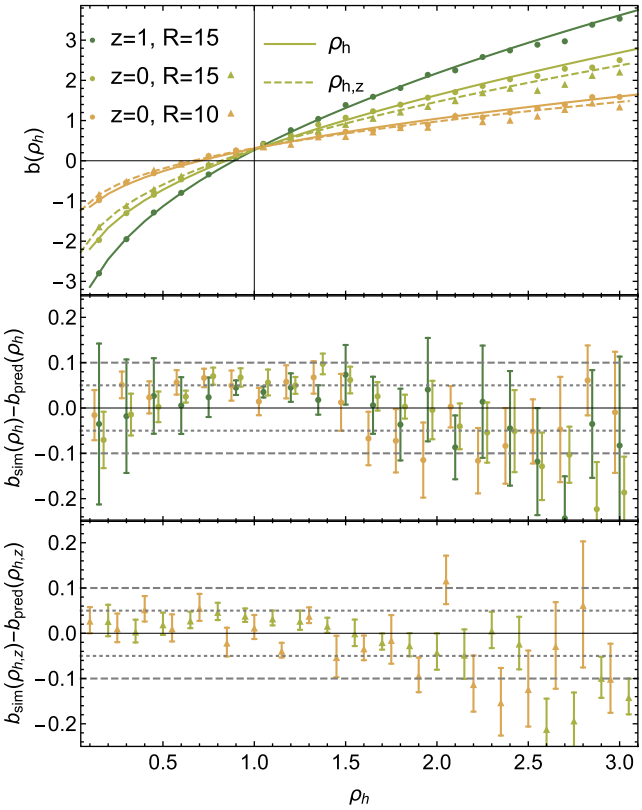

Figure 6. Halo sphere bias function $b_\circ$ describing the modulation of the two-point halo correlation function with the density as measured for a separation $r = 30$ Mpc h^{-1} in real space (circles) and redshift space (triangles) in comparison to the analytical prediction based on a measured dark matter variance σ_μ and the fitted bias parameters in real space b_n (solid lines) and redshift space b_n^z (dashed lines) as given in Table 1.

scatter of the bias relation could in principle have degraded the accuracy of the PDF, but Fig. 5 shows that it turns out to be a small effect. This remains true for counts of haloes in redshift space, even though the redshift space scatter plot displayed significantly larger scatter than its real space counterpart. Fig. 6 compares the prediction for the sphere bias function in both real and redshift space, based on the same inputs as used for the halo PDF, with the measurements from the simulation and is also displaying excellent agreement. Note that, for the redshift-space correlation, which has an angular dependence, only the monopole is effectively probed. To measure the sphere bias function, encoding the excess correlation between densities in spheres according to equation (7), a separation of $r = 30$ Mpc h^{-1} is chosen, giving a grid of non-overlapping spheres. The densities of the six neighbouring spheres are collected in bins of width $\Delta\rho = 0.15$; precise formulas are given by equations (19) and (20) in Uhlemann et al. (2017b).

5 APPLICATION: PARAMETER ESTIMATION

One of the main goals of constructing tracer statistics is to extract cosmological parameters from counts in cells. Let us now make use of the one-point halo PDF (16) alone or combine it with the density-dependent sphere bias (18) to estimate either the bias parameters, the underlying dark matter variance, or both.

Due to the strong (although not complete) degeneracy between the dark matter variance and linear bias [that can be shown to hold exactly for a linearly biased lognormal PDF (see Appendix B), and

is given as $\sigma_{\mu,\mathrm{h}} \simeq \sigma_{\mu,\mathrm{m}}/b_1$ at leading order in perturbation theory according to equation (C4)], it turns out one cannot use the one-point statistics alone to jointly determine the dark matter variance and bias parameters. This is not at all surprising, given the well-known degeneracy between linear bias and the clustering amplitude, caused by the fact that a low matter fluctuation amplitude can be masked out by a high galaxy bias or vice versa (see e.g. Seljak et al. 2005). In principle, if (i) all the statistics could be measured exactly, (ii) the truncation in the bias model was fully justified and (iii) the dark matter PDF was exactly given by the LDS model and in particular different from log-normal, then it should be possible to measure jointly the dark matter variance and the three bias parameters. We found that in practice, when considering limited noisy samples (with a number of spheres corresponding to the available cosmic volume at low redshifts), only the first three cumulants (mean, variance, skewness) carry significant information in a statistical sense. Indeed, the next order contribution to the PDF is coming from the fourth-order cumulant which scales like σ^6 and is therefore negligible with respect to the mean, variance and third-order cumulant. The information coming from this term is therefore much lower than that of the lower order terms.[4] Hence, measuring the one-point PDF can only put three constraints on the parameters of the model. For a quadratic bias model, this means that one effectively ends up with a degeneracy line (i.e a one-dimensional manifold) in the four-dimensional parameter space. Because the information coming from higher order cumulants is not exactly zero, we expect this degeneracy line to be a very elongated ellipsoid instead which is indeed what we find. Indeed, Section 5.1 shows how in practice the one-point model does not yield enough information to measure both on realistic surveys and discusses complementary strategies when relying on one-point statistics only, while Section 5.2 explains why one- and two-points halo counts does break this degeneracy in principle. Finally, Section 5.3 shows how a joint fit of both counts from the HR4 simulation yields an estimate of all four parameters plus the dark matter correlation function.

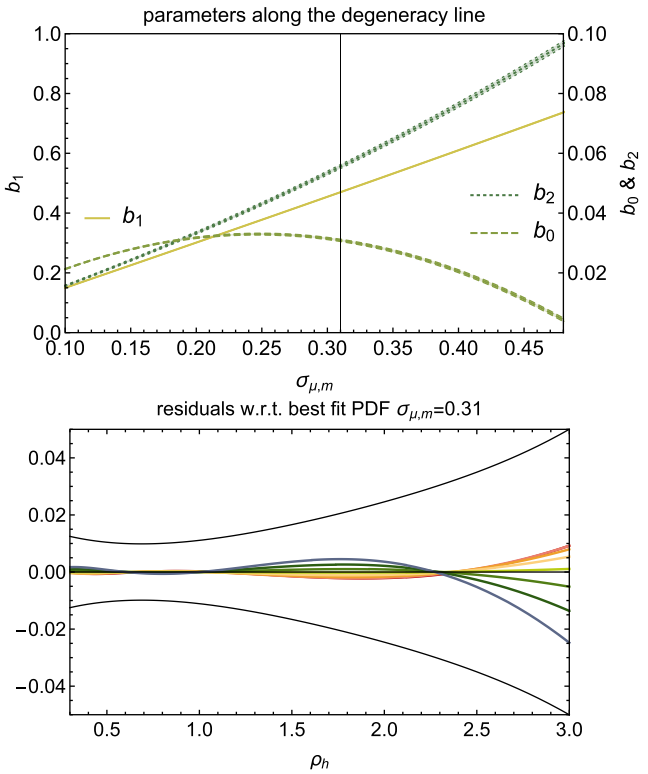

Figure 7. Top: Parameters along the degeneracy line obtained from a fit to the measured density PDF at $z = 1$ and for a radius $R = 15\,\mathrm{Mpc}\,h^{-1}$ in the HR4 simulation when determining the bias parameters b_n given a fixed dark matter variance $\sigma_{\mu,\mathrm{m}}$. The thin shaded area corresponds to the one-sigma confidence interval for different values of the $\sigma_{\mu,\mathrm{m}}$. Bottom: predicted density PDF along the degeneracy line from $\sigma_{\mu,\mathrm{m}} = 0.1$ (red) to 0.45 (blue). Only residuals compared to the true value $\sigma_{\mu,\mathrm{m}} = 0.31$ are displayed. The black lines show the one-sigma error on the measured PDF, obtained by fitting with a polynomial the binned one-sigma error bars.

5.1 Bias-variance degeneracy in one-point statistics

In order to quantify the bias-variance degeneracy in one-point statistics, let us measure the density PDF at $z = 1$ in the HR4 simulation covered by spheres of radius $R = 15\,\mathrm{Mpc}\,h^{-1}$, and get 1σ error bars as the error on the mean estimated from eight subcubes. Let us describe the degeneracy with $\sigma_{\mu,\mathrm{m}}$ as the curvilinear coordinate and for each value of $\sigma_{\mu,\mathrm{m}}$ between 0.1 and 0.5, and fit the measured non-linear PDF from $\rho_\mathrm{h} = 0.3$ to 3 with bins $\Delta\rho_{\mathrm{h},\mathcal{P}} = 0.01$ as this is the regime where the model is expected to work well. The 1σ confidence intervals of the bias parameters as a function of $\sigma_{\mu,\mathrm{m}}$ are displayed in the top panel of Fig. 7. As expected from the perturbative argument, the degeneracy line is dominated by a linear relationship between b_1 and $\sigma_{\mu,\mathrm{m}}$ (with slope $\sigma_{\mu,\mathrm{h}} \approx 0.7$) with higher order correction leading to non-zero (but small) values of b_0 and b_2. The parabolic shape of b_0 and the linear growth of b_2 with σ, as well as their smallness, can in fact be understood perturbatively, as shown in equations (C3) and (C7) in Appendix C2. The bottom panel of Fig. 7 shows that the predicted PDFs along the degeneracy line are all within the 1σ error bars of the simulation and therefore cannot be distinguished. Combining this observable with other probes or using a model for the dark matter variance should in principle break this degeneracy.

If the non-linear dark matter variance was known, for example from empirical relations found in simulations (such as Repp & Szapudi 2017b) or higher order perturbation theory (see e.g. Scoccimarro & Frieman 1996), one could use the analytic dark matter PDF (1) to obtain the CDF of dark matter $\mathcal{C}_\mathrm{m}$ and then the bias relation using equation (11) by measuring the halo CDF $\hat{\mathcal{C}}_\mathrm{h}$. Note that this procedure essentially looks for a non-linear transformation of halo densities such that the result is distributed according to the dark matter PDF equation (1), and hence similar in spirit to the idea of Gaussianising the field (see e.g. McCullagh et al. 2016).

Conversely, if the bias parameters (including their time evolution) were known from either theory or measured from an independent probe, one could use the analytic halo PDF (16) to determine the dark matter variance and use this to constrain, for example the dark energy equation of state as demonstrated for dark matter in Codis et al. (2016a). Analytical attempts to predict cumulants of the halo density have been based on bias models starting from Press–Schechter (Casas-Miranda et al. 2002; Casas-Miranda, Mo & Boerner 2003), its extensions like excursion sets or peak theory, or the halo model (Fry et al. 2011b). Note that to take advantage of this idea one needs access to the bias that relates averaged halo and matter densities rather than the bias based on n-point functions. While there is a mapping between the two in the large-scale limit,

[4] In addition, it is of the same order as the next order perturbative contribution to the third-order cumulant which is not well-captured by the LDS being exact only at tree order.

for $R \gtrsim 50\,\mathrm{Mpc}\,h^{-1}$, they are not equal and their relation depends on the shape of the power spectrum as well as the smoothing radius and filter shape, as pointed out in Desjacques et al. (2016). For particular observational signatures that are not degenerate with bias, such as local primordial non-Gaussianity (Uhlemann et. al. in preparation), the present formalism allows us to take the nature of tracers into account and hence to obtain more realistic constraints. In principle, future peculiar velocity surveys could also gain us qualitative insights into biasing following the idea described in Uhlemann et al. (2017a), although their statistical power is unlikely to yield accurate enough constraints.

5.2 Joint one- and two-point statistics: the basic idea

In order to break the degeneracy between bias parameters and the dark matter variance, one can make use of the two-point statistics from equation (17) to jointly constrain the dark matter variance and biases. The two-point halo PDF is built from the one-point halo PDFs (16) and the density-dependent sphere bias (18) that modulates the two-point correlation function which were successfully compared to numerical simulations in Section 4.

Let us present here the basic idea behind the degeneracy lift. The leading-order mixed cumulant depends on the two-point sphere bias function via

$$C_{12,\mathrm{h}} = \left\langle \delta_\mathrm{h}^2 \delta_\mathrm{h}' \right\rangle = \xi_{\circ,\mathrm{h}} \int (\rho_\mathrm{h} - 1)^2 b_{\circ,\mathrm{h}}(\rho_\mathrm{h}) P(\rho_\mathrm{h})\, \mathrm{d}\rho_\mathrm{h}\,. \tag{20}$$

Since the sphere bias function is not linear $b_{\circ,\mathrm{h}}(\rho) \not\propto \rho_\mathrm{h} - 1$, especially in the tails that are sensitive to b_2, equation (20) differs from the one-point cumulant given by the skewness

$$C_{3,\mathrm{h}} = \langle \delta_\mathrm{h}^3 \rangle = \int (\rho_\mathrm{h} - 1)^3 P(\rho_\mathrm{h})\, \mathrm{d}\rho_\mathrm{h}\,. \tag{21}$$

The leading order expressions[5] relating the corresponding dark matter and halo reduced cumulants defined as $S_3 = C_3/\sigma^4$ and $S_{12} = C_{12}/(\xi\sigma^2)$ for the adopted (inverse quadratic in the log-densities) biasing model are consistently given by

$$S_3^{\mu,\mathrm{m}} = S_3 - 3 = b_1^{-1}\left(S_3^{\mu,\mathrm{h}} + 6b_2/b_1\right)\,, \tag{22}$$

$$S_{12}^{\mu,\mathrm{m}} = S_{12} - 2 = b_1^{-1}\left(S_{12}^{\mu,\mathrm{h}} + 4b_2/b_1\right)\,. \tag{23}$$

Combining equations (22) and (23) allows us in principle to solve for the bias parameters, by relying on theoretical predictions for the dark matter cumulants on the one hand, and measurements for the halo cumulants on the other hand.[6]

This paper extends this cumulant-based strategy by taking advantage of the full two-point information (Bernardeau & Schaeffer 1992; Munshi, Melott & Coles 2000) which consistently include higher order cumulants leading to improved accuracy, as demonstrated in Codis et al. (2016a) and Uhlemann et al. (2017b). In effect, instead of being restricted to the lowest order cumulants, it makes simultaneous use of the one-point PDF and the two-point sphere bias function. Indeed, it can be shown that the two-point sphere bias' slope with respect to the density is sensitive to bias alone, hence the joint analysis of both counts breaks the degeneracy. Appendix C sketches a proof at the perturbative level.

[5] Note that, at that order, the reduced cumulants of the density ρ and log-density μ only differ by a constant, see Uhlemann et al. (2016).

[6] These expressions closely resemble those given in Bel & Marinoni (2012) which use a forward biasing model in the densities. This paper relies on the lowest order cumulants predicted by tree-order perturbation theory and combines them in a difference that is suspected to be more robust than the individual cumulants.

Table 2. Collection of the results (best fits and 1σ confidence intervals) of the joint fitting procedure for $\mathcal{P}_\mathrm{h}$ and $b_{\circ,\mathrm{h}}$ for radius $R\,(\mathrm{Mpc}\,h^{-1})$ and redshift z at separation $r = 30\,\mathrm{Mpc}\,h^{-1}$. The expected values, as given in Table 1, are $\sigma_{\mu,\mathrm{m}} = 0.310$, $\xi_{\circ,\mathrm{m}} = 0.016$, $b_0 = 0.028$, $b_1 = 0.473$, $b_2 = 0.055$ and lie well within the confidence intervals.

Param		Dark matter		Tracer bias		
z	R	$\sigma_{\mu,\mathrm{m}}$	$\xi_{\circ,\mathrm{m}}(r)$	b_0	b_1	b_2
1	15	0.306 $\pm$0.015	0.0154 $\pm$0.0016	0.0309 $\pm$0.0016	0.463 $\pm$0.024	0.0534 $\pm$0.0032

5.3 Joint one- and two-point statistics: a worked example

Let us finally present a worked-out fiducial experiment that allows us to simultaneously obtain the dark matter variance, correlation function as well as the bias parameters from measurements of one-point halo PDF $P_\mathrm{h}(\rho_\mathrm{h}|\sigma_{\mu,\mathrm{m}}, b_0, b_1, b_2)$ given a redshift z and sphere radius R and the two-point halo sphere bias $b_\mathrm{h}(\rho_\mathrm{h}|\xi_\mathrm{m}(r), \sigma_{\mu,\mathrm{m}}, b_0, b_1, b_2)$ at a separation $r \geq 2R$. In practice, sampling the joint likelihood for five parameters is computationally expensive and tricky because the joint PDF is noisy and the signal coming from the sphere bias rather small.[7] Let us therefore resort here to a simpler fitting procedure to illustrate the capability of the one- and two-point halo statistics for jointly constraining the dark matter variance and correlation along with the bias parameters. A data sample is derived from the simulation by binning the halo densities and measuring a histogram for the PDF P_h in the range $\rho_\mathrm{h} \in [0.1, 3]$ with bin width $\Delta\rho_{\mathrm{h},\mathcal{P}} = 0.01$ and the scaled halo sphere bias $\tilde{b}_{\circ,\mathrm{h}}$

$$\tilde{b}_{\circ,\mathrm{h}}(\rho_\mathrm{h}) \equiv \left\langle \rho_\mathrm{h}'(r) | \rho_\mathrm{h} \right\rangle - 1 = \xi_{\circ,\mathrm{h}} b_{\circ,\mathrm{h}}(\rho_\mathrm{h}), \tag{24}$$

in the range $\rho_\mathrm{h} \in [0.07, 2.5]$ with bin width $\Delta\rho_{\mathrm{h},b} = 3/21$. The scaled halo sphere bias is used instead of the halo sphere bias as this is the direct observable. The LDS prediction is given by

$$\tilde{b}_{\circ,\mathrm{h}}(\rho_\mathrm{h}) = \left\langle \rho_\mathrm{h} b_{\circ,\mathrm{m}}(\rho_\mathrm{m}(\rho_\mathrm{h})) \right\rangle\ \xi_{\circ,\mathrm{m}} b_{\circ,\mathrm{m}}(\rho_\mathrm{m}(\rho_\mathrm{h}))\,, \tag{25}$$

where the prefactor encodes the difference of the correlation function $\sqrt{\xi_{\circ,\mathrm{h}}/\xi_{\circ,\mathrm{m}}} = \langle \rho_\mathrm{h} b_{\circ,\mathrm{m}}(\rho_\mathrm{m}(\rho_\mathrm{h})) \rangle$ and is tabulated using a fifth-order Taylor expansion of $\rho_\mathrm{h}(\rho_\mathrm{m})$ near one.

Using this sample, a non-linear model fit is implemented for the two functions $\mathcal{P}_\mathrm{h}(\rho_\mathrm{h})$ and $b_{\circ,\mathrm{h}}(\rho_\mathrm{h})$ with weights determined by the errors from the measured PDF and bias function (using bootstrapping over eight subsamples of the simulation). The result of the fit for the parameters and the associated uncertainties is given in Table 2 (see also Fig. 8 for the corresponding figures of merit) and agrees very well with the directly measured values reported in Table 1. In particular, the sphere bias (i.e the two-point statistics of density in spheres) is shown as anticipated to break the degeneracy. Since the dark matter correlation function $\xi_{\circ,\mathrm{m}}$ enters as an overall amplitude, the degeneracy is broken by the information contained in the shape of the sphere bias function, rather than its amplitude, as can be seen perturbatively in Appendix C. As the noise is more important in the two-point sphere bias than in the one-point density

[7] Note also that the tracer PDF's boundaries depend on the bias parameters, which, combined with the fact that the LDS model is only accurate on a finite range of densities adds an extra layer of complexity to the likelihood exploration.

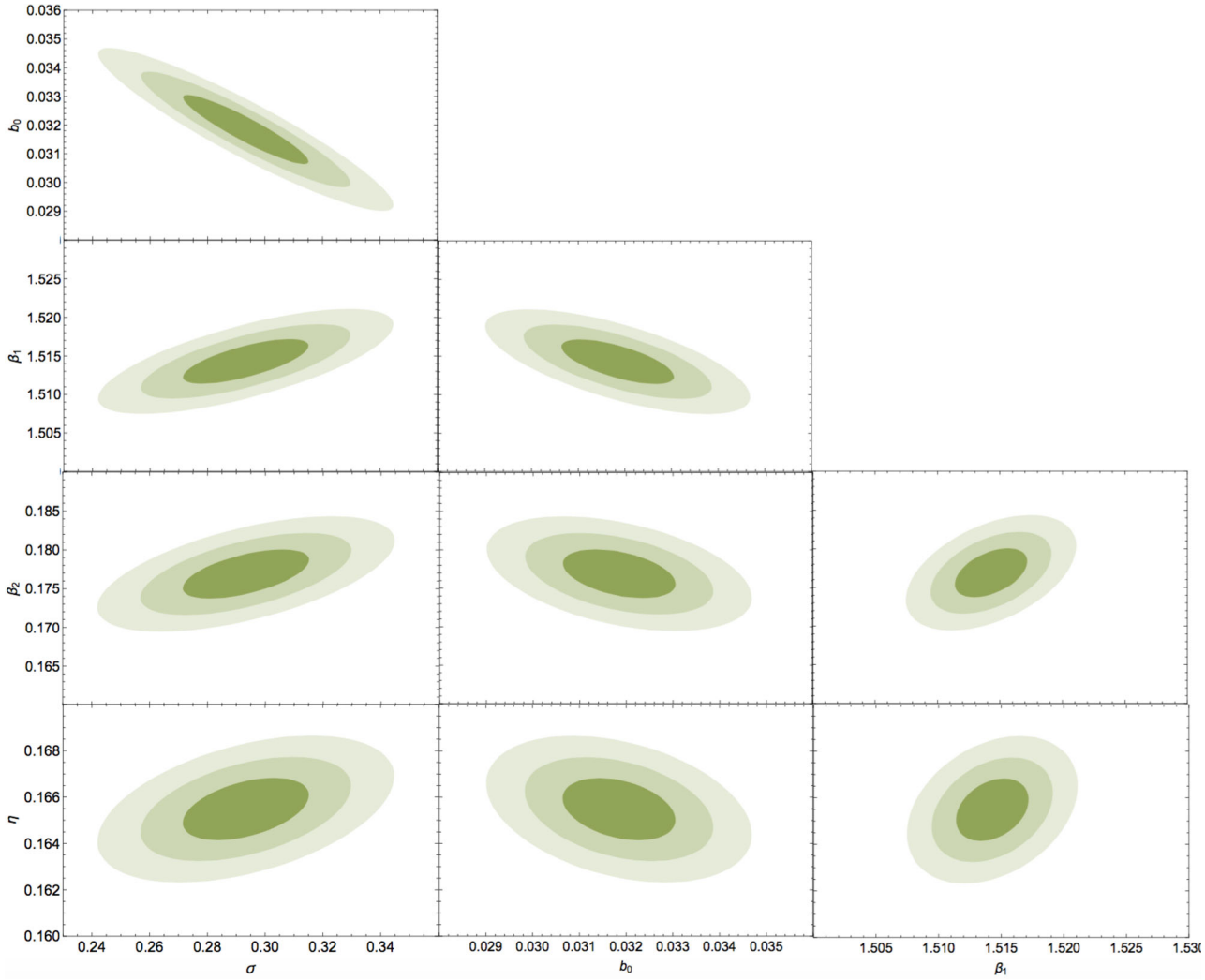

Figure 8. One, two and three sigmas contours obtained by fitting the density PDF and the bias function at $z = 1$ and for spheres of radius $R = 15\,\mathrm{Mpc}\,h^{-1}$ where $\eta = \xi/\sigma^2$, $\beta_i = b_i/\sigma$ and $\sigma = \sigma_{\mu,\mathrm{m}}$.

PDF, the error budget on the parameters of the model is dominated by the accuracy on the measurement of $\tilde{b}_{\circ,\mathrm{h}}$.

The total number of spheres ($\approx 10^6$) is of the order of the number of spheres that a survey like *Euclid* will probe at a redshift around $z \approx 1$ (Codis et al. 2016a). Hence, one can expect this novel idea to be applicable to real data in a very near future, which will allow us to measure consistently the growth of fluctuations across cosmic time (through the dark matter variance $\sigma_{\mu,\mathrm{m}}$) and to characterize galaxy biasing (through a set of bias parameters at different redshifts). The accuracy of the constraints on those parameters depends on the accessible survey volume and therefore the number of spheres N, in a way which can be studied by subsampling the simulation. Redoing the above-described analysis on eight subcubes of the simulation, yields the average best-fitting values (notably 0.29, 0.030, 1.518, 0.181, 0.161 for $\sigma_{\mu,\mathrm{m}}$, b_0, β_1, β_2, η) are consistent with the parameters estimated from the full box (0.29, 0.032, 1.514, 0.177, 0.165), as seen on Fig. 9. The mean standard deviation are, respectively, 0.033, 0.0030, 0.0052, 0.0056, 0.0023 (to be compared with the 1σ error bars from the full volume: 0.016, 0.00089, 0.0021, 0.0023, 0.00099), which is consistent with a $1/\sqrt{N}$ scaling. Overall, the typical one-sigma errors evolve as $\Delta\sigma_{\mu,\mathrm{m}} = 0.016\sqrt{10^6/N}$ and $\Delta\xi_{\circ,\mathrm{m}} = 0.0017\sqrt{10^6/N}$.

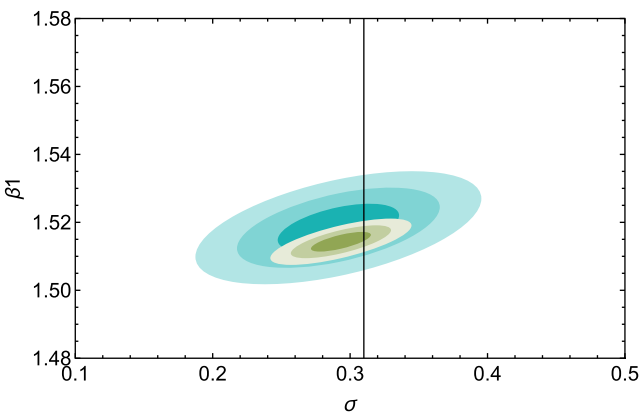

Figure 9. Mean 1σ, 2σ, 3σ contours obtained from the eight subcubes by averaging the best fits and covariance matrix (cyan). For comparison, the figure of merit of the whole volume is superimposed in dark green and a line at the target value $\sigma \equiv \sigma_{\mu,\mathrm{m}} = 0.31$ is displayed. As expected the constraints on the model parameters shrink when the accessible volume increases.

The above presented experiment is of course fairly idealized at various levels. It may turn out to be too optimistic, but should none the less provide a framework in which to implement a dark energy experiment based on count-in-cells.

6 CONCLUSIONS

Starting from a very accurate model for the dark matter density in cells, we extended it to biased tracers such as dark haloes or galaxies and compared them to the state-of-the-art *N*-body simulation HR4 in real and redshift space. Our main findings can be summarized as follows:

(i) on scales of the order of 10 Mpc h^{-1}, mass-weighted subhalo densities show considerably less scatter than their number-weighted version; they can be accurately fit with a quadratic bias model in the log-densities and closely resemble the bias relation of mass-weighted galaxy densities.

(ii) Using a quadratic mean bias model for log-densities and neglecting the scatter is sufficient to obtain a one-point halo PDF and two-point sphere bias that are as accurate as the underlying dark matter results when compared against simulations, see Figs 5 and 6. Combining the quadratic bias model with fitted coefficients with the dark matter PDF from LDS with the measured dark matter variance, the accuracy of the halo PDF is well within 5 per cent over a wide range of densities, in both real and redshift space.

(iii) The one-point PDF yields access to a one-dimensional manifold in the four-dimensional parameter space of dark matter variance and quadratic bias.

(iv) Combining the one-point halo PDF and the two-point halo sphere bias, one can jointly constrain the non-linear dark matter variance and correlation as well as the bias parameters, and hence disentangle tracer bias from non-linear gravitational evolution. This is of interest both from the point of view of dark energy and non-linear power spectra estimation. The density-dependent clustering signal encoded in the two-point sphere bias is related to the concept of 'sliced' or 'marked' correlation functions (see e.g. Sheth 2005; White & Padmanabhan 2009; Neyrinck et al. 2016) which hence might contain valuable information about bias and could be used to break the degeneracy between linear bias and the clustering amplitude in the two-point correlation.

(v) Comparison to counts extracted from 'full-physics' hydrodynamical simulations suggest that our findings will scale from dark haloes to galaxies.

The excellent accuracy of the analytical prediction for the dark matter PDF and two-point bias plays a critical role in disentangling the dark matter variance from biasing when applied to tracers. Hence, this formalism should be applied to constrain cosmology using counts-in-cells statistics in ongoing or upcoming surveys like DES, *Euclid*, *WFIRST*, LSST, KiDs, following the fiducial dark energy experiment presented in Codis et al. (2016a).

ACKNOWLEDGEMENTS

This work is partially supported by the grants ANR-12-BS05-0002 and ANR-13-BS05-0005 of the French *Agence Nationale de la Recherche*. CU is supported by the Delta-ITP consortium, a program of the Netherlands organization for scientific research (NWO) funded by the Dutch Ministry of Education, Culture and Science (OCW). We thank Tobias Baldauf, Karim Benabed, Donghui Jeong, Marcello Musso, Fabian Schmidt, Ravi Sheth and the participants of the workshops 'Statistics of Extrema in Large Scale Structure' and the 'Biased Tracers of Large-Scale Structure' for discussions. We thank Iary Davidzon for having run the SED-fitting on the photometry of the simulated galaxies in the Horizon-AGN simulation in order to compute mock observed stellar masses. CU thanks IAP and CITA, while MF, DP and SC also thank KIAS for hospitality while some of this work was done. Many thanks to Stéphane Rouberol for smoothly running the Horizon cluster which is hosted by the Institut d'Astrophysique de Paris, and to our colleagues who produced and post-processed the Horizon-AGN/ HR4 simulations. We also thank the KIAS Center for Advanced Computation for providing computing resources.

REFERENCES

Abbas U., Sheth R. K., 2007, MNRAS, 378, 641
Abell G. O., 1958, ApJS, 3, 211
Adelberger K. L., Steidel C. C., Giavalisco M., Dickinson M., Pettini M., Kellogg M., 1998, ApJ, 505, 18
Angulo R. E., Baugh C. M., Lacey C. G., 2008, MNRAS, 387, 921
Arnouts S. et al., 2002, MNRAS, 329, 355
Aubert D., Pichon C., Colombi S., 2004, MNRAS, 352, 376
Bahcall N. A., Soneira R. M., 1983, ApJ, 270, 20
Baldauf T., Seljak U., Senatore L., Zaldarriaga M., 2011, J. Cosmol. Astropart. Phys., 10, 031
Balian R., Schaeffer R., 1989, A&A, 220, 1
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Baugh C. M., 2006, Rep. Prog. Phys., 69, 3101
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Bel J., Marinoni C., 2012, MNRAS, 424, 971
Bel J. et al., 2016, A&A, 588, A51
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Bernardeau F., 1992, ApJ, 392, 1
Bernardeau F., 1994a, A&A, 291, 697
Bernardeau F., 1994b, ApJ, 427, 51
Bernardeau F., 1996, A&A, 312, 11
Bernardeau F., Kofman L., 1995, ApJ, 443, 479
Bernardeau F., Reimberg P., 2016, Phys. Rev. D, 94, 063520
Bernardeau F., Schaeffer R., 1992, A&A, 255, 1
Bernardeau F., Pichon C., Codis S., 2014, Phys. Rev. D, 90, 103519
Bernardeau F., Codis S., Pichon C., 2015, MNRAS, 449, L105
Blanton M., Cen R., Ostriker J. P., Strauss M. A., 1999, ApJ, 522, 590
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Casas-Miranda R., Mo H. J., Sheth R. K., Boerner G., 2002, MNRAS, 333, 730
Casas-Miranda R., Mo H. J., Boerner G., 2003, MNRAS, 339, 872
Cen R., Ostriker J., 1992, ApJ, 393, 22
Chan K. C., Scoccimarro R., 2012, Phys. Rev. D, 86, 103519
Clerkin L. et al., 2017, MNRAS, 466, 1444
Codis S., Pichon C., Bernardeau F., Uhlemann C., Prunet S., 2016a, MNRAS, 460, 1549
Codis S., Bernardeau F., Pichon C., 2016b, MNRAS, 460, 1598
Coles P., 1986, MNRAS, 222, 9P
Coles P., Jones B., 1991, MNRAS, 248, 1
Colombi S., 1994, ApJ, 435, 536
Cooray A., Milosavljević M., 2005, ApJ, 627, L89
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Coupon J. et al., 2015, MNRAS, 449, 1352
Dekel A., Lahav O., 1999, ApJ, 520, 24
Desjacques V., Jeong D., Schmidt F., 2016, ApJS, 224, 24
Dressler A., 1980, ApJ, 236, 351
Dubinski J., Kim J., Park C., Humble R., 2004, New A, 9, 111
Dubois Y., Devriendt J., Slyz A., Teyssier R., 2012, MNRAS, 420, 2662
Dubois Y. et al., 2014a, MNRAS, 444, 1453
Dubois Y. et al., 2014b, MNRAS, 444, 1453
Dwek E., 1998, ApJ, 501, 643
Fosalba P., Gaztanaga E., 1998, MNRAS, 301, 503

Frusciante N., Sheth R. K., 2012, J. Cosmol. Astropart. Phys., 11, 016
Fry J. N., 1985, ApJ, 289, 10
Fry J. N., Gaztanaga E., 1993, ApJ, 413, 447
Fry J. N., Colombi S., Fosalba P., Balaraman A., Szapudi I., Teyssier R., 2011a, MNRAS, 415, 153
Fry J. N., Colombi S., Fosalba P., Balaraman A., Szapudi I., Teyssier R., 2011b, MNRAS, 415, 153
Gaztañaga E., Fosalba P., Elizalde E., 2000, ApJ, 539, 522
Haardt F., Madau P., 1996, ApJ, 461, 20
Hamaus N., Seljak U., Desjacques V., Smith R. E., Baldauf T., 2010, Phys. Rev. D, 82, 043515
Hurtado-Gil L., Martínez V. J., Arnalte-Mur P., Pons-Bordería M. J., Pareja-Flores C., Paredes S., 2017, A&A, 601, 13
Ilbert O. et al., 2006, A&A, 457, 841
Jee I., Park C., Kim J., Choi Y.-Y., Kim S. S., 2012, ApJ, 753, 11
Jonsson P., 2006, MNRAS, 372, 2
Juszkiewicz R., Bouchet F. R., Colombi S., 1993, ApJ, 412, L9
Juszkiewicz R., Weinberg D. H., Amsterdamski P., Chodorowski M., Bouchet F., 1995, ApJ, 442, 39
Kaiser N., 1984, ApJ, 284, L9
Kauffmann G., Nusser A., Steinmetz M., 1997, MNRAS, 286, 795
Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999, MNRAS, 303, 188
Kaviraj S. et al., 2017, MNRAS, 467, 4739
Kim J., Park C., 2006, ApJ, 639, 600
Kim J., Park C., L'Huillier B., Hong S. E., 2015, J. Korean Astron. Soc., 48, 213
Komatsu E. et al., 2011, ApJS, 192, 18
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Kulier A., Ostriker J. P., 2015, MNRAS, 452, 4013
L'Huillier B., Park C., Kim J., 2014, New A, 30, 79
Laigle C. et al., 2016, preprint (arXiv:1604.02350)
Manera M., Gaztañaga E., 2011, MNRAS, 415, 383
Matsubara T., 2011, Phys. Rev. D, 83, 083518
McCullagh N., Neyrinck M., Norberg P., Cole S., 2016, MNRAS, 457, 3652
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution, Cambridge Univ. Press, Cambridge
Mobasher B. et al., 2015, ApJ, 808, 101
Moster B. P., Naab T., White S. D. M., 2017, MNRAS, preprint (arXiv:1705.05373)
Munshi D., Sahni V., Starobinsky A. A., 1994, ApJ, 436, 517
Munshi D., Melott A. L., Coles P., 2000, MNRAS, 311, 149
Neyrinck M. C., Aragón-Calvo M. A., Jeong D., Wang X., 2014, MNRAS, 441, 646
Neyrinck M. C., Szapudi I., McCullagh N., Szalay A., Falck B., Wang J., 2016, MNRAS, preprint (arXiv:1610.06215)
Ohta Y., Kayo I., Taruya A., 2003, ApJ, 589, 1
Paranjape A., Sefusatti E., Chan K. C., Desjacques V., Monaco P., Sheth R. K., 2013, MNRAS, 436, 449
Pforr J., Maraston C., Tonini C., 2012, MNRAS, 422, 3285
Pichon C., Thiébaut E., Prunet S., Benabed K., Colombi S., Sousbie T., Teyssier R., 2010, MNRAS, 401, 705
Porto R. A., 2016, Phys. Rep., 633, 1
Repp A., Szapudi I., 2017a, preprint (arXiv:1708.00954)
Repp A., Szapudi I., 2017b, MNRAS, 464, L21
Schmidt F., Jeong D., Desjacques V., 2013, Phys. Rev. D, 88, 023515
Scoccimarro R., 2000, ApJ, 542, 1
Scoccimarro R., Frieman J., 1996, ApJS, 105, 37
Seljak U. et al., 2005, Phys. Rev. D, 71, 043511
Seljak U., Hamaus N., Desjacques V., 2009, Phys. Rev. Lett., 103, 091303
Senatore L., 2015, J. Cosmol. Astropart. Phys., 11, 007
Sheth R. K., 2005, MNRAS, 364, 796
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
Sheth R. K., Mo H. J., Saslaw W. C., 1994, ApJ, 427, 562
Shin J., Kim J., Pichon C., Jeong D., Park C., 2017, ApJ, 843, 73
Sigad Y., Branchini E., Dekel A., 2000, ApJ, 540, 62
Somerville R. S., Lemson G., Kolatt T. S., Dekel A., 2000, MNRAS, 316, 479
Szapudi I., Pan J., 2004, ApJ, 602, 26
Szapudi I., Meiksin A., Nichol R. C., 1996, ApJ, 473, 15
Teyssier R., 2002, A&A, 385, 337
Uhlemann C., Codis S., Pichon C., Bernardeau F., Reimberg P., 2016, MNRAS, 460, 1529
Uhlemann C., Codis S., Hahn O., Pichon C., Bernardeau F., 2017a, MNRAS, 469, 2481
Uhlemann C., Codis S., Kim J., Pichon C., Bernardeau F., Pogosyan D., Park C., L'Huillier B., 2017b, MNRAS, 466, 2067
Valageas P., 2002a, A&A, 382, 412
Valageas P., 2002b, A&A, 382, 450
Vale A., Ostriker J. P., 2004, MNRAS, 353, 189
Vale A., Ostriker J. P., 2006, MNRAS, 371, 1173
White M., Padmanabhan N., 2009, MNRAS, 395, 2381
Wolk M., McCracken H. J., Colombi S., Fry J. N., Kilbinger M., Hudelot P., Mellier Y., Ilbert O., 2013, MNRAS, 435, 2
Yang A., Saslaw W. C., 2011, ApJ, 729, 123
Yang X., Mo H. J., van den Bosch F. C., Zhang Y., Han J., 2012, ApJ, 752, 41
Zu Y., Mandelbaum R., 2015, MNRAS, 454, 1161

APPENDIX A: LOGARITHMIC SCATTER PLOTS

A logarithmic view on the scatter plot similar to Fig. 1 is shown in Fig. A1. In there, one can also find a comparison between the best-fitting linear and quadratic bias model for log-densities. As evident from the plot, the linear model cannot fit the extreme density regions. According to Neyrinck et al. (2014), an exponential relationship between the halo and dark matter density is expected in the very low density tail. We also find that a linear relation for log-densities does not fit the low-density end and that the quadratic term is essential there, but we do not see the need for another higher order term. This observation could be related to differences in halo densities under

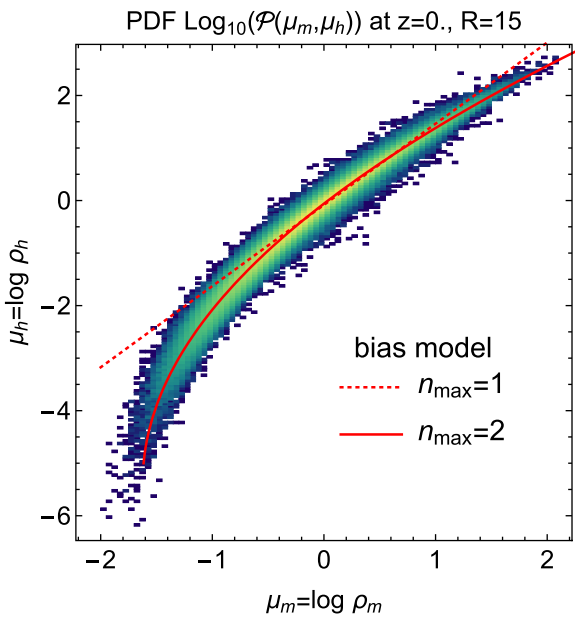

Figure A1. Density scatter plot of the halo log-density $\mu_{\rm h} = \log \rho_{\rm h}$ with mass-weighting versus the dark matter log-density $\mu_{\rm m} = \log \rho_{\rm m}$ for radius $R = 15\,{\rm Mpc}\,h^{-1}$ at redshift $z = 0$. The figure also shows the the best-fitting linear and quadratic bias model for the log-density obtained from a fit to the CDF bias function (dotted and solid line, respectively)

consideration, in particular the mass-range and weighting scheme for the haloes or the size and shape of the cells. Note, however, that unlike Neyrinck et al. (2014), we did not apply strategies to suppress discreteness and stochasticity, so the difference could also be due to an insufficient halo sampling level for the very low density end.

APPENDIX B: LOGNORMAL RECONSTRUCTION

Let us compare the LDS approach to the well-known lognormal models. The lognormal PDF, proposed first from a dynamical model for dark matter in Coles & Jones (1991) but nowadays being used as a phenomenological parametrization for PDFs of dark matter and its tracers, has the following form

$$\mathcal{P}_{\rm LN}(\rho\,|\,\sigma_\mu,\bar{\mu}) = \frac{1}{\sqrt{2\pi\sigma_\mu}}\frac{1}{\rho}\exp\left[-\frac{(\log\rho-\bar{\mu})^2}{2\sigma_\mu^2}\right], \tag{B1}$$

where the variance σ_μ of the log-density $\mu = \log\rho$ can be treated as free parameter and the mean of the log-density is connected to the variance via $\bar{\mu} = \langle\log\rho\rangle \approx -\sigma^2/2$ by requiring a unit mean density $\langle\rho\rangle = 1$. The skewed lognormal PDF as introduced in Colombi (1994), involves an Edgeworth expansion around the lognormal PDF and reads

$$\mathcal{P}_{\rm SLN}(\rho\,|\,\sigma_\mu,\bar{\mu},\epsilon_3,\epsilon_4) = \mathcal{P}_{\rm LN}(\rho\,|\,\sigma_\mu,\bar{\mu}) \times\left[1+\frac{\epsilon_3}{6}H_3(\hat{\mu})+\frac{\epsilon_4}{24}H_4(\hat{\mu})+\frac{\epsilon_3^2}{72}H_6(\hat{\mu})\right], \tag{B2}$$

with the normalized log-density $\hat{\mu} = (\mu-\bar{\mu})/\sigma_\mu$, its rescaled cumulants $\epsilon_n = \langle\hat{\mu}\rangle_c$ and the probabilist's Hermite polynomials H_n.[8] A comparison between the accuracy of the three different lognormal based models, when the underlying parameters (the mean $\bar{\mu}$, variance σ_μ, skewness ϵ_3 and kurtosis ϵ_4 of the log-density) are measured from the simulated halo densities as shown in Fig. B1. The generalized normal distribution $\mathcal{N}_{v2}$ adopted by Shin et al. (2017) to fit dark matter PDFs has very similar properties to the skewed lognormal PDF in the range of radii we consider and hence will not be discussed here.

The lognormal dark matter PDF can be combined with a polynomial bias model for the log-densities. For a linear bias model of the log-densities, the resulting halo PDF is again lognormal with variance and mean given by $\sigma_{\mu,\rm h} = \sigma_{\mu,\rm m}/b_1$ and $-\bar{\mu}_{\rm h} = b_0/b_1 + b_1\sigma^2_{\mu,\rm h}/2$, once the dark matter mean density is fixed to one so that $-\bar{\mu}_{\rm m} = \sigma^2_{\mu,\rm m}/2$. In addition, the halo mean density being one, one gets an additional constraint which relates the constant bias shift to the linear bias factor and the variance according to $b_0 = b_1(1-b_1)\sigma^2_{\mu,\rm h}/2$ and agrees with the leading order perturbative result. In this model, there is a full degeneracy between the linear bias b_1 and the log-variance of the underlying dark matter $\sigma_{\mu,\rm m}$.

Let us now consider a quadratic log-bias model. Even if the dark matter PDF was close to lognormal (which is typically the case at ∼10 per cent accuracy; see Uhlemann et al. 2017b), this non-linear mapping induces extra terms in the exponential. If one expands these terms in an Edgeworth-like fashion, one can see that non-linear bias naturally feeds into higher order cumulants, in particular the skewness and kurtosis, which is why it is necessary to go to the skewed lognormal forms to fit the measured halo PDF. The residuals obtained when augmenting the lognormal dark matter PDF with the quadratic bias model with measured parameters are shown as comparison in the lower panel of Fig. B1. In this case, the predicted PDF is slightly less accurate than the large-deviation prediction, with two additional parameters that cannot easily be related to bias because they mix in contributions from the dark matter PDF, which is significantly better fitted by a skewed lognormal. This is in contrast with the LDS formalism that clearly disentangles the effect of gravitational evolution (parametrized through the non-linear dark matter variance) from non-linear biasing (parametrized through the bias parameters).

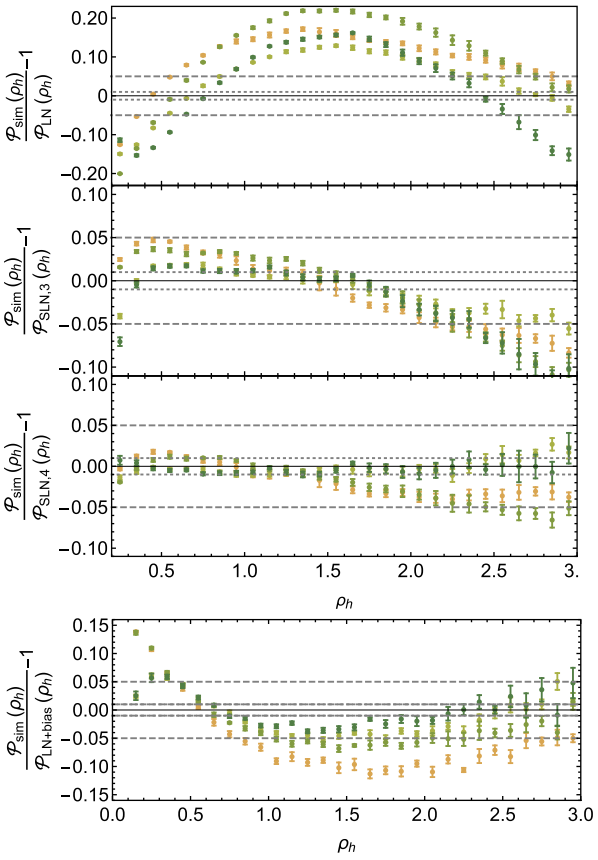

Figure B1. Halo mass-density PDFs $\mathcal{P}_{\rm h}$ for direct measurements based on halo catalogues at redshifts $z = 0$, 1 for radii $R = 10$, $15\,{\rm Mpc}\,h^{-1}$ in comparison to the recovered PDF of lognormal models with measured mean and variance (upper panel), the skewed lognormal model with measured cumulants up to skewness (upper middle panel) and up to kurtosis (lower middle panel) and the PDF assuming a lognormal model with measured matter variance and quadratic bias model for the log-densities (lower panel).

APPENDIX C: BREAKING DEGENERACIES

The main text has shown that with a practical implementation of the LDS formalism, one can accurately measure bias parameters and dark matter variance, and break the degeneracies by including information from two-point statistics, an idea also followed by Bel & Marinoni (2012) in another context. Let us illustrate these findings using perturbation theory.

[8] Note that Edgeworth expansions are known to be problematic in the tails of the distribution because the expression in the brackets can eventually become negative depending on the size of the corrections in the cumulant expansion.

C1 One-point PDF

From equation (10), one can easily compute the relation between halo and matter contrast within the quadratic log bias model

$$\delta_{\rm h} = -1 + \exp\left[\frac{\sqrt{b_1^2 - 4b_2(b_0 - \log(1+\delta_{\rm m}))} - b_1}{2b_2}\right]. \quad \text{(C1)}$$

Expanding this relation for small contrasts yields perturbative bias consistency relations. First, imposing a zero mean for the halo contrast allows us to get b_0 at all orders in the dark matter variance σ

$$b_0 = \textstyle\sum_i b_0^{(i)}\sigma^{2i}\,, \quad \text{(C2)}$$

with

$$b_0^{(0)} = 0\,, \quad b_0^{(1)} = \frac{b_1 - 2b_2 - b_1^2}{2b_1^2}. \quad \text{(C3)}$$

The measured halo variance then imposes a relation between the dark matter variance and the bias parameters which reads

$$\sigma_{\rm h}^2 = \left(\frac{\sigma}{b_1}\right)^2\left[1 + \sigma^2\Delta_{\rm NL}\right]\,, \quad \text{(C4)}$$

with

$$\Delta_{\rm NL}^{(0)} = \frac{(S_3 - 3)(b_1 - 2b_2 - b_1^2)}{b_1^2} + \frac{20b_2^2 - 8b_1b_2 - b_1^4}{2b_1^4}. \quad \text{(C5)}$$

The constraints are therefore dominated by this degeneracy between b_1 and σ at first order in PT. After the mean and the variance, the PDF will typically pick up the information from the skewness. Let us therefore compute perturbatively the skewness of the halo density field. At first order, it reads

$$S_{3,\rm h} = 3 + b_1(S_3 - 3) - 6\frac{b_2}{b_1} + \mathcal{O}(\sigma). \quad \text{(C6)}$$

This latter equation gives a relation between σ and b_2 at first order

$$b_2 = \frac{\sigma(3 - S_{3,\rm h})}{6\sigma_{\rm h}}\left(1 + \frac{\sigma(S_3 - 3)}{\sigma_{\rm h}(S_{3,\rm h} - 3)}\right) + \mathcal{O}(\sigma^2). \quad \text{(C7)}$$

Equations (C4) and (C7) predict a linear degeneracy between on the one hand σ and b_1 and on the other hand σ and b_2 which is indeed observed when performing the model fitting (see Fig. 7). This model fitting described in Section 5 eventually gathers all the information coming from the mean, variance, skewness and higher order cumulants in a fully consistent way (because LPD provides the PDF and therefore the full statistics). In principle, the knowledge of the full hierarchy of cumulants eventually breaks those degeneracies if the LDS model is exact. In practice, (i) sample noise prevents accurate measurements of the higher order cumulants (kurtosis etc.) which scale like higher power of the variance (σ^6 and above) and (ii) loop corrections in the skewness that are not accounted for in the LDS model appears at the same perturbative order as those higher order cumulants and therefore do not allow us to fully break the degeneracy between the parameters. To break this degeneracy, one must involve two-point statistics as described in the next section.

C2 Two-point PDF

Let us assume that the two-point PDF of the matter density is well described by its large-scale approximation given by equation (6). The sphere bias $b_{\circ,\rm m}(\rho_{\rm m})$ can be exactly computed using the large-deviation principle (Codis et al. 2016b; Uhlemann et al. 2017b). A fair approximation for small densities is given by

$$b_{\circ,\rm m}(\rho_{\rm m}) = \frac{\tau_{\rm SC}(\rho_{\rm m})}{\sigma_{\rm L}^2\left(R\rho_{\rm m}^{1/3}\right)}. \quad \text{(C8)}$$

Remarkably, plugging in the bias relation in $b_\circ(\rho)$, shows that the sphere bias of the halo density field at small density behaves as

$$b_{\circ,\rm h}(\rho_{\rm h}) = \frac{\delta_{\rm h}b_\circ^{(1)} + b_\circ^{(0)}}{\sigma_{\rm L}^2\left(R\sqrt[3]{{\rm e}^{b_0}+1}\right)}\,, \quad \text{(C9)}$$

where

$$b_\circ^{(1)} = \frac{2{\rm e}^{b_0}b_1\nu\gamma\left(\sqrt[3]{{\rm e}^{b_0}+1}\right)}{3\left({\rm e}^{b_0}+1\right)^{2/3}}\left[\left({\rm e}^{b_0}+1\right)^{-\frac{1}{\nu}} - 1\right] + \frac{{\rm e}^{b_0}b_1}{\left({\rm e}^{b_0}+1\right)^{\frac{1+\nu}{\nu}}}\,,$$

$$b_\circ^{(0)} = \nu\left(1 - \left({\rm e}^{b_0}+1\right)^{-\frac{1}{\nu}}\right)\,.$$

and $\gamma = \sigma'/\sigma$. Obviously, the overall amplitude in $b_{\circ,\rm h}$ cannot be measured because it is degenerate with the unknown dark matter correlation function $\xi_{\circ,\rm m}$. In contrast, the ratio between the slope called $b_\circ^{(1)}$ and intercept $b_\circ^{(0)}$ can be obtained

$$\frac{b_\circ^{(1)}}{b_\circ^{(0)}} = -\frac{2{\rm e}^{b_0}b_1\gamma\left(\sqrt[3]{{\rm e}^{b_0}+1}\right)}{3\left({\rm e}^{b_0}+1\right)^{2/3}} + \frac{{\rm e}^{b_0}b_1\left({\rm e}^{b_0}+1\right)^{-1}}{\nu\left(\left({\rm e}^{b_0}+1\right)^{\frac{1}{\nu}} - 1\right)}. \quad \text{(C10)}$$

This ratio is in particular proportional to b_1 and does not depend on the variance. Constraining this ratio, as is done in the joint fit presented in the main text will therefore break the degeneracy in equation (C4).

APPENDIX D: THE HORIZON-AGN SIMULATION

Let us briefly describe the cosmological hydrodynamical simulation used in the main text, Horizon-AGN (Dubois et al. 2014b). The simulation http://www.horizon-simulation.org/ is run with a ΛCDM cosmology with total matter density $\Omega_{\rm m} = 0.272$, dark energy density $\Omega_\Lambda = 0.728$, amplitude of the matter power spectrum $\sigma_8 = 0.81$, baryon density $\Omega_{\rm b} = 0.045$, Hubble constant $H_0 = 70.4\,{\rm km\,s^{-1}\,Mpc^{-1}}$ and $n_{\rm s} = 0.967$ compatible with the WMAP-7 data (Komatsu 2011). The size of the simulation box is $L_{\rm box} = 100\,h^{-1}\,{\rm Mpc}$ on a side, and the volume contains 1024^3 DM particles, corresponding to a DM mass resolution of $M_{\rm DM,res} = 8\times10^7\,{\rm M}_\odot$. The simulation is run with the RAMSES code (Teyssier 2002), and the initially coarse 1024^3 grid is adaptively refined down to $\Delta x = 1$ proper kpc, with refinement triggered in a quasi-Lagrangian manner: if the number of DM particles becomes greater than 8, or the total baryonic mass reaches eight times the initial DM mass resolution in a cell. It leads to a typical number of 6.5×10^9 gas resolution elements (leaf cells) in the simulation at $z = 1$. Heating of the gas from a uniform UV background takes place after redshift $z_{\rm reion} = 10$ following Haardt & Madau (1996).

Star formation occurs in regions of gas number density above $n_0 = 0.1\,{\rm H\,cm^{-3}}$ following a Schmidt law: $\dot\rho_* = \epsilon_*\rho_{\rm g}/t_{\rm ff}$, where $\dot\rho_*$ is the star formation rate mass density, $\rho_{\rm g}$ the gas mass density, $\epsilon_* = 0.02$ the constant star formation efficiency, and $t_{\rm ff}$ the local free-fall time of the gas. Feedback from stellar winds, supernovae type Ia and type II are included into the simulation with mass, energy and metal release. The simulation also follow the formation of black holes (BHs), which can grow by gas accretion at a Bondi-capped-at-Eddington rate and coalesce when they form a tight enough binary.

BHs release energy in a quasar/radio (heating/jet) mode when the accretion rate is, respectively, above and below one per cent of Eddington, with efficiencies tuned to match the BH-galaxy scaling relations at $z = 0$ (see Dubois et al. 2012, for details). A light-cone has been generated from the simulation, as described in Pichon et al. (2010). The area of the light-cone is 5 deg^2 below $z = 1$ and 1 deg^2 above. A mock photometric galaxy catalogue has been extracted from the light-cone in order to mimic observational data sets (see Laigle et al. in preparation for more details). Galaxies have been identified from the stellar particles distribution using the ADAPTAHOP halo finder (Aubert, Pichon & Colombi 2004). The local density is computed from a total of 20 neighbours, and a density threshold ρ_t of 178 times the average matter density is required to select structures. Once identified mock galaxies in the light-cone, a BC03 simple stellar population (SSP) has been attached to any stellar particle in each galaxy, according to its mass and stellar metallicity. The spectrum of the galaxy is then obtained by adding the SEDs of all the SSPs. The (possibly redshifted) spectra are then convolved with photometric filter passbands, in order to get absolute and apparent magnitudes in the following 13 bands: *NUV*, *u*, *B*, *V*, *r*, i^+, z^{++}, *Y*, *J*, *H*, K_s, 3.6μm, 4.5μm. Dust attenuation is also taken into account along the line of sight of each stellar particle in the galaxy, assuming the dust mass scales with the gas metal mass, with a dust-to-metal ratio of 0.4 (Dwek 1998; Jonsson 2006). In order to get observed stellar masses, the SED-fitting code LEPHARE (Arnouts et al. 2002; Ilbert et al. 2006) has been run using as input photometry the virtual magnitudes included in the mock catalogue and with a configuration similar to Laigle et al. (2016).

In closing, the galaxy population was shown to reproduce in overall the luminosity function of observed galaxies in Kaviraj et al. (2017) (see Fig. D1 for a qualitative representation of the count-in-cells within its light-cone).

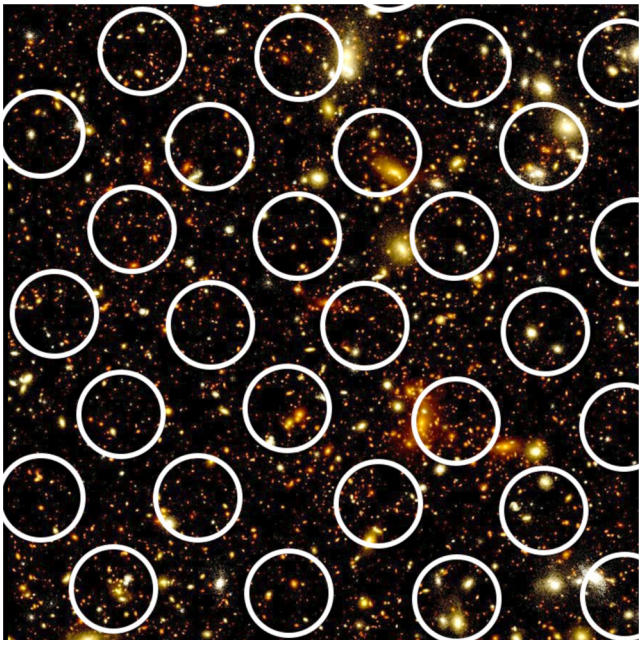

Figure D1. Qualitative distribution of spheres in the simulated galaxies in Horizon-AGN. The background represents synthetic galaxies produced by the simulation while converting cold gas into stars. Realistic colours are post processed using spectral synthesis (Kaviraj et al. 2017).

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