A new method to measure galaxy bias

Jennifer E. Pollack¹*, Robert E. Smith²,³†, & Cristiano Porciani¹‡

¹Argelander Institut für Astronomie der Universität Bonn, Auf dem Hügel 71, D-53121 Bonn, Germany
²Max-Planck Institute for Astrophysics, Karl-Schwarzschild-Str.1, Postfach 1523, 85740 Garching, Germany
³Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH

ABSTRACT
We present a new approach for modelling galaxy/halo bias that utilizes the full non-linear information contained in the moments of the matter density field, which we derive using a set of numerical simulations. Although our method is general, we perform a case study based on the local Eulerian bias scheme truncated to second-order. Using 200 N-body simulations covering a total comoving volume of $675 h^{-3}$ Gpc$^3$, we measure several 2- and 3-point statistics of the halo distribution to unprecedented accuracy. We use the bias model to fit the halo-halo power spectrum, the halo-matter cross spectrum and the corresponding three bispectra for wave numbers in the range $0.04 \lesssim k \lesssim 0.12 h$ Mpc$^{-1}$. We find the constraints on the bias parameters obtained using the full non-linear information differ significantly from those derived using standard perturbation theory at leading order. Hence, neglecting the full non-linear information leads to biased results for this particular scale range. We also test the validity of the second-order Eulerian local biasing scheme by comparing the parameter constraints derived from different statistics. Analysis of the halo-matter cross-correlation coefficients defined for the 2- and 3-point statistics reveals further inconsistencies contained in the second-order Eulerian bias scheme, suggesting it is too simple a model to describe halo bias with high accuracy.

Key words: cosmology: theory, large-scale structure

1 INTRODUCTION
The clustering statistics of the galaxy distribution contain a wealth of information about the cosmological model. However, in the absence of a robust theory for galaxy formation, extracting this information can only be achieved in part. In practice, to do this requires us to assume a specific phenomenological relationship between the density field of galaxies and that of the underlying matter, more commonly referred to as galaxy bias. Whilst still incomplete, our leading theories of galaxy formation, do provide a great deal of insight about the distribution of galaxies. For instance they predict that galaxies should only reside in dark-matter haloes and be strongly associated with the distribution of sub-structures (for a detailed review of galaxy formation see Mo, van den Bosch & White 2010). This greatly simplifies our ability to construct a phenomenological model for the galaxy distribution on large scales: it should be closely related to a weighted average of the dark-matter-halo overdensities (e.g. Smith, Scoccimarro & Sheth 2007).

There are a number of detailed analytical approaches for characterizing the bias of dark-matter haloes with respect to the mass distribution (for a recent review see Porciani 2013). However, it has yet to be determined which model provides the most accurate description of galaxy bias. In the simplest method, the local Eulerian bias model (hereafter LEB), one assumes that the overdensities of the biased tracers can be written as some function of the matter-density field at the same location. If both densities are smoothed over the patch scale $R$, then the biased field may be written as a Taylor-series expansion (Fry & Gaztanaga 1993). If one considers sufficiently large patches, then high-order corrections are guaranteed to be small and the series may be truncated after a finite number of terms.

Halo-clustering predictions of the LEB expressed in terms of standard perturbation theory (hereafter SPT, for a review see Bernardeau et al. 2002) have been examined in numerous works (Scoccimarro et al. 2001; Smith, Scoccimarro & Sheth 2007; Guo & Jing 2009; Roth & Porciani 2011; Manera & Gaztanaga 2011; Pollack, Smith & Porciani 2012; Chuen Chan & Scoccimarro 2012). One of the results to emerge from these studies is that, when the model is applied to halo counts within finite volumes of linear size...
The coefficients of the bias expansion show a running with the "cell" size. However, halo-clustering statistics such as the n-point correlation functions (or the corresponding n-spectra) do not contain any smoothing scale and should not depend on $R$. There has been much debate in the literature on how to reconcile these seemingly contrasting results (see Porciani 2013, for a concise summary).

This has led some to discuss an "effective" or "renormalized" bias approach where the scale-dependence of the bias coefficients is compensated by the contribution of small-scale perturbations in the matter density $\delta$ (Heavens, Matarrese & Verde 1998; McDonald 2006; Schmidt, Jeong & Desjacques 2012). Whilst such a scheme may be plausible (Jeong & Komatsu 2008; Smith, Hernández-Monteagudo & Seljak 2009), the development of a unique renormalization method is still ongoing, especially for dynamically evolved configurations in Euclidean space. On the other hand, it was recently proposed by Chuen Chan & Scoccimarro (2012) that the bias parameters obtained counting halos within cells of size $R$ are only relevant for describing perturbations of wavenumber $k \approx 0.8/R$ in the halo distribution. While there is no challenge to their argument when analyzing power-spectrum data, it does present a complication when using higher-order statistics such as the bispectrum. In order to interpret the galaxy bispectrum one would be required to compute bias coefficients separately for each configuration of wavevectors. This approach appears somewhat cumbersome to implement.

Currently, most observational analyses of galaxy clustering assume that galaxy bias can be described by the truncated LEB and that the statistical properties of the non-linear matter density field can be modelled using SPT. To leading order in the perturbations, this requires only one bias parameter for 2-point statistics of the tracers and two parameters for 3-point statistics. Present-day galaxy surveys, however, do not cover enough comoving volume to accurately sample the spatial scales at which tree-level results provide an accurate description of galaxy clustering. The presence of rare large-scale structures, for instance, significantly alters the measurements of three-point statistics (e.g. Nichol et al. 2006). On smaller scales, where data are more robust, dynamical non-linearities pose a serious challenge to the models. Adopting the simplified LEB+SPT model may therefore generate systematic errors and thus influence the characterisation of the bias or the estimation of the cosmological parameters.

The LEB truncated to second order is the standard workhorse for studying three-point statistics of galaxy clustering. Its predictions to leading perturbative order have been used to interpret measurements from the two-degree field galaxy redshift survey (Verde et al. 2002, Jing & Börner 2004, Wang et al. 2004, Gaztañaga et al. 2003), the Sloan Digital Sky Survey (Kayo et al. 2004, Hikage et al. 2005, Pan & Szapudi 2003), Kulkarni et al. 2007, Nishimichi et al. 2007, Marin 2011, McBride et al. 2011, McBride et al. 2011, Guo et al. 2013), and the WiggleZ Dark Energy Survey (Marín et al. 2013). In our previous study (Pollack, Smith, & Porciani 2012), we demonstrated that, in order to robustly model three-point statistics with the LEB, one must necessarily have an accurate model for the clustering statistics of the non-linear matter density on the relevant scales. This is imperative to recover the correct values of the bias parameters in controlled numerical experiments. Therefore, it is not surprising that past investigations based on the LEB+SPT model reached inconsistent conclusions. For example, studying the galaxy bispectrum on scales $0.1 < k < 0.5 h$ Mpc$^{-1}$, Verde et al. (2002) concluded that 2dF galaxies are unbiased tracers of the mass distribution. On the other hand, using the complete 2dF sample, Gaztañaga et al. (2003) found strong evidence for non-linear biasing from the analysis of the three-point correlation function with triangle configurations that probe separations between 9 and 36 $h^{-1}$Mpc (see also Jing & Börner 2004; Wang et al. 2004).

In this paper, we build upon our past experience and present a general method to model the clustering of biased tracers of the mass distribution on mildly non-linear scales $k < 0.1 h$ Mpc$^{-1}$. This is key to extend studies of galaxy clustering to smaller spatial separations where observational data are less certain. Our method relies on using N-body simulations to measure the relevant statistics for the clustering of the underlying mass distribution. Related approaches have been presented by Sigad, Branchini & Dekel (2000) and Szapudi & Park (2004) for galaxy counts in cells (see also Pan & Szapudi 2003 for an application to correlation functions). We apply our general framework to the modelling of n-point clustering statistics of non-linear, Euclidian, locally biased tracers. In our framework, bias parameters run with the patch scale $R$. We address the running of the bias by treating the filter scale as a nuisance parameter to be marginalized over. The major advantage of our scheme is that we exactly recover the matter poly-spectra used in the bias model at every order. The only truncation necessary in the model is the choice as to what level to truncate the bias expansion, and this may be selected by the data in a Bayesian model comparison. We test our modelling framework up to quadratic order in the local bias expansion (as commonly done in recent observational studies), for the power- and bi-spectra of haloes and their cross-spectra with matter measured from a large ensemble (200 realizations) of measurements from a series of large ΛCDM N-body simulations. This ensemble of simulations resolves the halos that should host luminous red galaxies over a total comoving volume of $675 h^{-3}$ Gpc$^3$, and so provides us with a very stringent statistical test ground for our model.

The sections are organized as follows. In §2 we set our mathematical notation and introduce the LEB. The numerical simulations used in this work are briefly described in §3 and used in §4 to measure several statistical quantities for the matter and halo distributions. In §5 we use Bayesian statistics to estimate the free parameters of the LEB and describe our main results. Finally, in §6 and 7 we further discuss our findings and present our conclusions.

2 A NEW FRAMEWORK FOR MODELLING THE CLUSTERING OF BIASED TRACERS

2.1 General formalism

Consider some discrete tracers of the large-scale structure (dark-matter haloes or galaxies) with mean density $\bar{n}_h$ and physical density $n_\delta(x) = [1 + \delta_n(x)] \bar{n}_h$. We want to relate this random field to the underlying distribution of matter
with local density $\rho(x) = [1 + \delta(x)] \bar{\rho}$. If we assume that the density contrast of the tracers averaged over some patch of linear size $R$, $\delta_h(x|R)$, is locally related to the density of matter in the same patch, then we may write

$$\delta_h(x|R) = F[\delta(x|R)]$$

(1)

where $F$ denotes a generic function $\mathbb{R} \to \mathbb{R}$ and the symbols

$$\delta_n(x|R) \equiv \int d^3y W(|x - y|, R) \delta_n(y)$$

(2)

(where $a$ stands for haloes or matter) denote smoothed overdensity fields, $W$ being a rotation-invariant filter function with size $R$.

Since we are dealing with smooth mathematical functions we may Taylor expand Eq. (1) to obtain

$$\text{Fry & Gaztanaga (1994):}$$

$$\delta_h(x|R) = \sum_{n=1}^{\infty} \frac{b_n}{n!} \Delta_n^{(n)}(x|R)$$

(3)

where the terms $b_n$ are the Eulerian bias coefficients of order $n$, which depend on both the smoothing scale and the exact definition of the tracers (e.g. halo mass, etc.). Note that the subtraction of the terms $\Delta_n^{(n)}(x|R)$ at each order ensures that $\delta_h(x|R) = 0$, where $\langle \ldots \rangle$ denote an ensemble average. On Fourier transforming the above relation one finds, for $|k| \neq 0$,

$$\delta_h(k|R) = \int_{-\infty}^{\infty} \frac{d^3q}{(2\pi)^3} \delta^{D}(k - q_{1...n}) \prod_{i=1}^{n} \delta(q_i|R) \frac{d^3q_i}{(2\pi)^3}$$

(5)

where $\Delta_n^{(n)}(k|R)$ can be written as

$$\Delta_n^{(n)}(k|R) \equiv (2\pi)^3 \int \delta^{D}(k - q_{1...n}) \prod_{i=1}^{n} \delta(q_i|R) \frac{d^3q_i}{(2\pi)^3}.$$

In the last expression $\delta^{D}(k)$ denotes the Dirac-delta distribution and we have made use of the compact notation $q_1 = q_2 + \cdots + q_n$ and $\delta(q_i|R) = \delta(q_iW(q|R))$.

We now define the power spectrum of the biased tracers and their cross-spectrum with the matter in the terms of the correlators:

$$\langle \delta_h(k_1|R) \delta_h(k_2|R) \rangle \equiv \langle \delta_q(k_1|R) \delta_q(k_2|R) \rangle.$$  

(6)

Similarly, the corresponding bispectra can be defined as

$$\langle \delta_h(k_1|R) \delta_h(k_2|R) \delta_h(k_3|R) \rangle \equiv \langle \delta_q(k_1|R) \delta_q(k_2|R) \delta_q(k_3|R) \rangle.$$  

(7)

where we have suppressed the dependence of the bispectrum on the third wavevector, since the Dirac-delta distribution imposes $k_3 = -k_1-k_2$. On inserting Eq. (4) into Eq. (6), we find:

$$\langle \delta_h(k_1|R) \delta_h(k_2|R) \rangle = \sum_{l,m,n=1}^{\infty} \frac{\Gamma_l^m \Gamma_m^n}{l! m! n!} \langle \delta_h^{(l)}(k_1|R) \delta_h^{(m)}(k_2|R) \rangle.$$  

(8)

with $\Gamma_l^m = b_l$ and $\Gamma_l^m = \delta_{l}^{m}$ (for haloes and matter, respectively) where $\delta_{l}^{m}$ denotes the Kronecker-delta function. Similarly for Eq. (7) we have:

$$\langle \delta_h(k_1) \delta_h(k_2) \delta_h(k_3) \rangle = \sum_{l,m,n=1}^{\infty} \frac{\Gamma_l^m \Gamma_m^n \Gamma_n^m}{l! m! n!} \times \langle \delta_h^{(l)}(k_1|R) \delta_h^{(m)}(k_2|R) \delta_h^{(n)}(k_3|R) \rangle.$$  

(9)

It is convenient to introduce the functions $P_{l,m}$ and $B_{l,m,n}$ such that

$$\langle \delta_h^{(l)}(k_1|R) \delta_h^{(m)}(k_2|R) \rangle = (2\pi)^3 \delta^{D}(k_{12}) \mathcal{P}_{l,m}(k_1),$$

and

$$\langle \delta_h^{(l)}(k_1|R) \delta_h^{(m)}(k_2|R) \delta_h^{(n)}(k_3|R) \rangle = (2\pi)^3 \delta^{D}(k_{123}) \mathcal{B}_{l,m,n}(k_1,k_2,k_3).$$

(10)

In simple words, $\mathcal{P}_{l,m}(k_1)$ denotes the cross power spectrum between the smoothed random fields $\langle \delta(x|R) \rangle$ and $\langle \delta(x|R) \rangle$, whereas $\mathcal{B}_{l,m,n}(k_1,k_2,k_3)$ is the corresponding bispectrum. Thus for the halo and matter power and bispectra we have:

$$\mathcal{P}_{\alpha\beta}(k_1) = \sum_{l,m=1}^{\infty} \frac{\Gamma_l^m \Gamma_m^\alpha \Gamma_m^\beta}{l! m!} \mathcal{P}_{l,m}(k_1),$$

and

$$\mathcal{B}_{\alpha\beta\gamma}(k_1,k_2,k_3) = \sum_{l,m,n=1}^{\infty} \frac{\Gamma_l^m \Gamma_m^n \Gamma_m^\alpha \Gamma_n^\beta \Gamma_n^\gamma}{l! m! n!} \mathcal{B}_{l,m,n}(k_1,k_2,k_3).$$

(11)

The above sets of equations provide us with models for the power spectra and the bispectra of halo counts in cells of size $R$. However, what we really want to model is the halo 2- and 3-point functions, $P_{\alpha\beta}$ and $B_{\alpha\beta\gamma}$. We assume that these quantities can be approximately recovered by “de-smoothing” $\mathcal{P}_{\alpha\beta}$ and $\mathcal{B}_{\alpha\beta\gamma}$ (Smith, Scoccimarro & Sheth 2007; Smith, Sheth & Scoccimarro 2008; Sefusatti 2008).

$$P_{\alpha\beta}(k_1) = \frac{P_{\alpha\beta}(k_1)}{W^2(k_1)};$$

and

$$B_{\alpha\beta\gamma}(k_1,k_2,k_3) = \frac{B_{\alpha\beta\gamma}(k_1,k_2,k_3)}{W(k_1)W(k_2)W(k_3)W(k_{12})W(k_{13})W(k_{23})}.$$

(12)

Note that when considering a model of halo bias beyond linear order this operation does not fully remove the dependence of the theory on $R$. In Section 5 we will use the models presented in Eq. (12) and Eq. (13) to fit simulation data. Nevertheless, our choice to “de-smooth” the theoretical model is equivalent to analyzing counts in cell data with a smoothed model. This is due to the fact that in Fourier-space the smoothing kernels can be treated as multiplicative factors, which means that if we factorize the expressions by dividing out the product of the window functions the relation between the model and the data still holds. Hence, fitting counts in cells data with a smoothed model is indifferent to analyzing unsmoothed data with a “de-smoothed” or factorized model.

The smoothing scale must therefore be considered as a free parameter of the model, and so it must be either determined by fitting a set of data or marginalized over.

In $\mathcal{A}_1$ and $\mathcal{A}_2$ we show how the terms $\mathcal{P}_{l,m}$ and $\mathcal{B}_{l,m,n}$ are related to the $p$-point matter spectra, where $p = l + m$ or $p = l + m + n$, respectively. In $\mathcal{A}_3$ we prove that the functions $\mathcal{P}_{l,m}$ are totally symmetric in $l$ and $m$. For $l \neq m \neq n$, the functions $\mathcal{B}_{l,m,n}(k_1,k_2,k_3)$ are not in general symmetric in $l$, $m$, and $n$, unless the wavevectors $k_i$ are also exchanged, i.e. whilst $\mathcal{B}_{l,m,n}(k_1,k_2,k_3) = \mathcal{B}_{m,n,l}(k_3,k_2,k_1)$, $\mathcal{B}_{l,m,n}(k_1,k_2,k_3) \neq \mathcal{B}_{m,l,n}(k_1,k_2,k_3)$. Note that in this study we choose to work with $n$-point spectra, $\mathcal{P}_{(n_1...,n_n)}$, that are symmetric to an exchange of their vectorial arguments, and we accomplish this through the symmetrization.

© 0000 RAS, MNRAS 000, 000-000
and cross-power spectra with the total mass up to second-order in the bias model, may be written:

\[ B_{(2,1,1)}^{(s)} \equiv \frac{1}{3} \int \frac{d^3 q}{(2\pi)^3} P_4(q, k_1 - q, k_2, k_3) + 2 \text{ cyc} , \]

\[ B_{(2,2,1)}^{(s)} \equiv \frac{1}{3} \int \frac{d^3 q_1}{(2\pi)^3} \frac{d^3 q_2}{(2\pi)^3} P_5(q_1, k_1 - q_1, q_2, k_2 - q_2, k_3) + 2 \text{ cyc} ; \]

\[ B_{(2,2,2)}^{(s)} \equiv \int \frac{d^3 q_1}{(2\pi)^3} \frac{d^3 q_2}{(2\pi)^3} \times P_6(q_1, k_1 - q_1, q_2, k_2 - q_2, q_3, k_3 - q_3) . \]

In \( \text{eq} \) we show how one may estimate \( P_{(l,m)} \) and \( B_{(l,m,n)}^{(s)} \) directly from an N-body simulation.

### 3 N-BODY SIMULATIONS

In order to test the LEB and also to determine the covariance matrices of the various spectra we have simulated 200 realizations of a flat ΛCDM cosmological model. The specific cosmological parameters that we have adopted are: \( \sigma_8 = 0.8, \Omega_m = 0.25, \Omega_b = 0.04, h = 0.7, n_s = 1.0 \) where \( \sigma_8 \) is the variance of linear mass fluctuations in top-hat spheres of radius \( R = 8 h^{-1}\text{Mpc} \); \( \Omega_m \) and \( \Omega_b \) are the matter and baryon density parameters; \( h \) is the dimensionless Hubble parameter in units of \(100 \text{ km s}^{-1} \text{ Mpc}^{-1} \); and \( n \) is the power-law index of the primordial density power spectrum. Our adopted values were inspired by the results from the WMAP experiment (Komatsu et al. 2001).

All of the N-body simulations were run using the publicly available Tree-PM code GADGET-2 (Springel 2005). This code was used to follow with high accuracy the non-linear evolution under gravity of \( N = 750^3 \) equal mass particles in a periodic comoving cube of length \( L = 1500 h^{-1}\text{Mpc} \), giving a total sample volume of \( V = 675 h^{-3} \text{ Gpc}^3 \). Newtonian two-body forces were softened below scales \( \lambda_{\text{soft}} = 60 \text{ kpc} h^{-1} \). The transfer function for the simulations was generated using the publicly available cmbfast code (Seljak & Zaldarriaga 1996), with high sampling of the spatial frequencies on large scales. Initial conditions were laid down at redshift \( z = 49 \) using the serial version of the publicly available 2LPT code (Crocce, Pueblas & Scoccimarro 2006).

We use only the simulation outputs at redshift \( z = 0 \) for analysis and identify dark matter haloes using the code BFGnF. This is a Friends-of-Friends algorithm (Davis et al. 1984), where we adopted a linking length corresponding to \( b = 0.2 \) times the mean inter-particle spacing. The minimum number of particles an object must contain to be considered a bound halo was set to 20. This implies a minimum halo mass of \( M_{\text{min}} = 1.11 \times 10^{13} h^{-1} M_{\odot} \) and a mean number density of \( \bar{n}_h \approx 3.7 \times 10^{-4} h^{-3} \text{ Mpc}^{-3} \). Further details regarding this set of N-body simulations can be found in Smith (2007) and Smith et al. (2013).

### 4 ESTIMATING THE SPECTRA

In this section we describe how we estimate all the halo and matter power spectra that enter the second-order LEB from the N-body simulations at redshift \( z = 0 \).
4.1 The halo auto- and cross-power and bispectra

To begin, the halo and matter density fields are interpolated onto a cubical Cartesian mesh using the cloud-in-cell (CIC) algorithm. Throughout we use mesh sizes corresponding to \( N_{\text{cell}} = 1024^3 \). We then Fourier transform these grids using the Fast Fourier Transform technique and correct each mode for the CIC assignment. The three power spectra \( P_{\text{mm}}, P_{\text{hm}}, P_{\text{hh}} \), and the four bispectra, \( B_{\text{mmm}}, B_{\text{hmn}}, B_{\text{hmm}}, B_{\text{hhh}} \), are then estimated using the expressions:

\[
\hat{P}_{\alpha\beta}(\mathbf{k}_1) = \frac{L^3}{N(\mathbf{k}_1)} \sum_i^N(\mathbf{k}_1) \delta_\alpha(\mathbf{k}_1) \delta_\beta(-\mathbf{k}_1),
\]

\[
\hat{B}^{\alpha\beta,\gamma}_{\alpha\beta\gamma}(\mathbf{k}_1, \mathbf{k}_2, \theta_{12}) = \frac{1}{3} \frac{L^6}{N_{\text{tri}}} \sum_{i,j,k} \delta_\alpha(\mathbf{k}_i) \delta_\beta(\mathbf{k}_j) \times \\
\times \delta_\gamma(-\mathbf{k}_i - \mathbf{k}_j) + 2 \text{cyc},
\]

where \( N(\mathbf{k}_1) \) is the number of Fourier modes in a narrow shell centred on \( k_1 \), \( \epsilon(\mathbf{k}_i, \mathbf{k}_j) \) represents the pair of vectors which lie in thin shells centred on \( k_1 \) and \( k_2 \), whose angular separation lies in the angular bin centred on \( \theta_{12} \), and \( N_{\text{tri}} \equiv N_{\text{tri}}(k_1, k_2, \theta_{12}) \) is the total number of triangles with this configuration in Fourier space. The superscript “d” denotes that these are spectra of a discrete distribution of points (i.e. haloes) and must be corrected for shot noise. The forms of the Poissonian shot-noise corrections we adopt were presented in Pollack, Smith & Porciani (2012):

\[
\hat{P}_{\alpha\beta}^{\text{shot}}(\mathbf{k}_1) = \frac{\delta^K_{\alpha\beta}}{\bar{n}_\alpha}
\]

\[
\hat{B}^{\alpha\beta,\gamma}_{\alpha\beta\gamma}^{\text{shot}}(\mathbf{k}_1, \mathbf{k}_2) = \frac{1}{3} \frac{\delta^K_{\alpha\beta}}{\bar{n}_\alpha} [P_{\beta\gamma}(\mathbf{k}_1) + 2 \text{cyc}]
\]

\[
+ \frac{1}{3} \frac{\delta^K_{\beta\gamma}}{\bar{n}_\beta} [P_{\alpha\gamma}(\mathbf{k}_1) + 2 \text{cyc}]
\]

\[
+ \frac{1}{3} \frac{\delta^K_{\alpha\gamma}}{\bar{n}_\gamma} [P_{\alpha\beta}(\mathbf{k}_1) + 2 \text{cyc}] + \frac{\delta^K_{\alpha\beta} \delta^K_{\beta\gamma} + \delta^K_{\alpha\beta} \delta^K_{\alpha\gamma} + \delta^K_{\beta\gamma} \delta^K_{\alpha\gamma}}{\bar{n}_\alpha \bar{n}_\beta \bar{n}_\gamma}
\]

where \( \bar{n}_\alpha \) denotes the mean number density of either the matter particles or the halo population.

Figure 1 presents the various power- and bi-spectra averaged over the 200 realizations with the corresponding standard errors on the mean. All spectra were corrected for shot noise using Eqs (29) and (30). The bispectra were measured for triangle configurations with fixed lengths \( k_1 = 0.04 \, h \, \text{Mpc}^{-1} \) and \( k_2 = 2k_1 \), but with varying angle \( \theta_{12} \). We adopt the convention \( \theta_{12} = 0 \) for \( k_1 \) and \( k_2 \) parallel. In order to use the same range of wavenumbers, the power spectra were measured over the scale range \( 0.04 < k < 0.12 \, h \, \text{Mpc}^{-1} \).

4.2 Estimating \( P_{(l,m)} \) and \( B_{(l,m,n)}^{(s)} \)

The polyspectra \( P_{(l,m)} \) and \( B_{(l,m,n)}^{(s)} \) that enter the expressions for the halo power- and bi-spectra in the LEB are affected by the non-linear evolution of the matter fluctuations. While these terms are usually approximated with perturbative techniques, we measure them directly from our N-body
Bound of $R_k$ not want the largest variance of the density perturbations, $\sigma$, as a function of the filter scale (see Table 1). Our results show a number of smoothing scales within the range $2 \lesssim R \lesssim 18 \, h^{-1}\text{Mpc}$ in comparison with our basic CIC grid (see the main text for more details). The errorbars denote the standard error on the mean.

Table 1. Level of non-linearity in the smoothed mass-density field at redshift $z = 0$. Column 1: filter scale, $R$; Column 2: variance of density fluctuations, $\sigma^2(R)$; Column 3: volume fraction with $|\delta(R)| > 1$, $f$.

| $R [\, h^{-1}\text{Mpc}]$ | $\sigma^2(R)$ | $100 \times f$ |
|--------------------------|--------------|--------------|
| 2                        | 2.44         | 10.0         |
| 4                        | 0.71         | 8.4          |
| 6                        | 0.38         | 6.0          |
| 7                        | 0.28         | 4.9          |
| 8                        | 0.22         | 3.8          |
| 10                       | 0.15         | 2.2          |
| 12                       | 0.10         | 1.1          |
| 14                       | 0.08         | 0.5          |
| 16                       | 0.06         | 0.2          |
| 18                       | 0.05         | 0.1          |

$\delta \geq 1$ is $f \lesssim 0.1$ for all of the filter scales considered. Furthermore, in our previous work (Pollack, Smith & Porciani 2012), we evaluated the scatter plots of $\delta_k$ versus $\delta$ measured from our $N$-body simulations for different smoothing radii. We found that expressing $\delta_k$ as a polynomial function at second-order in $\delta$ can describe reasonably well the mean trend of the scatter.

4.3 Results for $P_{(l,m)}$ and $B_{(l,m,n)}$

Figures 2 and 3 show the results for the ensemble-averaged de-smoothed power and bispectra, $P_{(l,m)}$ and $B_{(l,m,n)}$, respectively. Focusing on the power spectrum, the panels show (from left to right) the matter power spectrum $P = P_{(1,1)} = P_{\text{mm}}$ followed by the terms $P_{(2,1)}$, and $P_{(2,2)}$.

For the bispectra, the panels show: the matter bispectra, $P_{(l,m)}$, followed by the terms $B_{(l,m,n)}$ for $n = 1$, $n = 2$, and $n = 3$. We have restricted the triangle configurations to those which enter the auto- and cross- halo bispectra shown in Figure 4. Each panel shows six sets of points with errorbars which denote the error bars for those points. The red crosses denote the resulting polynomials when no Gaussian

Simulations. We do this as follows. First, we correct each Fourier mode of the mass-density field for convolution with the CIC grid. Then we multiply the result by a Gaussian smoothing function $W(kR) = \exp[-(kR)^2/2]$ and inverse transform back to real space. Next, we generate the density field $\delta(x|R)$ for the relevant values of $l$ and re-transform them into Fourier space. We then deconvolve these fields for the original smoothing, which means simply multiplying each Fourier mode by $W^{-1}(kR)$. Finally, the required $P_{(l,m)}$ and $B_{(l,m,n)}$ terms, defined in terms of $\Delta^{(i)}(k|R)$ (see Eq. 7), can be estimated as follows

$$\hat{P}_{(l,m)}(k_1) = \frac{L^3}{N(k_1)} \sum_i N_{(k_i)} \Delta^{(i)}(k_1|R) \Delta^{(m)}(-k_1|R), \tag{31}$$

and

$$\hat{B}_{(l,m,n)}(k_1, k_2, \theta_{12}) = \frac{L^6}{3 N(k_1)} \sum_{k_1,k_2} \Delta^{(i)}(k_1|R) \times \Delta^{(m)}(k_2|R) \Delta^{(n)}(-k_1 - k_2|R) + 2 \text{cyc}. \tag{32}$$

We note that the functions $P_{(l,m)}$ and $B_{(l,m,n)}$ slowly vary with $R$ and so can be smoothly interpolated. Based on this knowledge, we measure the spectral functions over the range: $R = [2, 18] \, h^{-1}\text{Mpc}$, in increments of $\Delta R = 2 \, h^{-1}\text{Mpc}$, but including an additional measurement at $R = 7 \, h^{-1}\text{Mpc}$. The lower limit was adopted because we do not wish to smooth below the Lagrangian size of haloes, which for our sample is of the order of $\sim 3 - 4 \, h^{-1}\text{Mpc}$. The upper bound of $R = 18 \, h^{-1}\text{Mpc}$ we justifiy by noting that we do not want the largest $k$-mode entering our computations of the halo power- and bispectra to be too heavily smoothed.

Before inspecting the functions $P_{(l,m)}$ and $B_{(l,m,n)}$, we first report the level of non-linearity present in the smoothed matter-density field, $\delta(x|R)$. We quantify this by measuring the variance of the density perturbations, $\sigma^2(R)$ and the fraction of cells where the density contrast exceeds unity, $f$, as a function of the filter scale (see Table 1). Our results show that $\sigma^2(R) < 1$ for $R \gtrsim 4 \, h^{-1}\text{Mpc}$. We therefore expect the quadratic bias model to be a poor description for smaller values of $R$. However, we note that the fraction of the cells with

Figure 2. Measurements of the de-smoothed terms $P_{(1,1)}$, $P_{(2,1)}$, and $P_{(2,2)}$ averaged over 200 $N$-body simulations. We show results for a number of smoothing scales within the range $2 \lesssim R \lesssim 18 \, h^{-1}\text{Mpc}$ in comparison with our basic CIC grid (see the main text for more details). The errorbars denote the standard error on the mean.
A new method to measure galaxy bias

Figure 3. Same as for Figure 2 but for the bispectrum terms $B_{(1,1,1)}$, $B_{(2,1,1)}^{(s)}$, $B_{(2,2,1)}^{(s)}$, and $B_{(2,2,2)}^{(s)}$.

Comparing the different panels reveals how the amplitudes of the de-smoothed quantities vary. Obviously, for the matter power and bispectra, $P_{(1,1)}$ and $B_{(1,1,1)}$, all of the spectra overlap with the CIC result as the smoothing and the de-smoothing procedures perfectly cancel each other out. However, for the remaining $P_{(l,m)}$ and $B_{(l,m,n)}$ functions, the de-smoothed quantities vary with the scale $R$. In particular, as $R$ decreases, the overall amplitude of the spectra increases due to the contributions of small-scale modes. For the largest smoothing scales, the configuration dependence of the spectra is also modified. In order to gain some insight into the origin of this behaviour, let us consider, for instance, the term

$$P_{(2,1)}(k) = \int \frac{d^3q}{(2\pi)^3} B(q, k - q, -k) W(q, k - q)$$  \hspace{1cm} \text{(33)}$$

where $W$ is a generic weighting function defined as

$$W(k_1, k_2) = \frac{W(k_1 R) W(k_2 R)}{W(|k_1 + k_2| R)}$$ \hspace{1cm} \text{(34)}$$

For Gaussian smoothing, the weighting function in Eq. (33) can be re-expressed as $W(q, k, q) = \exp \left[-R^2 (q^2 - kq\mu)\right]$, with $\mu = k \cdot q / (kq)$ the cosine of the angle between $k$ and $q$. The contribution to the integral from all modes with $qR \gg 1$ is exponentially suppressed (i.e. $W \ll 1$).

The contribution to the integral from all modes with $qR \gg 1$ is exponentially suppressed (i.e. $W \ll 1$). However, at fixed $k$, $W$ assumes values larger than unity for $\mu > 0$ and $q < k\mu$ (independently of $R$) and presents an absolute maximum for $q = k/2$ and $\mu = 1$ where it takes the value $W_{\text{max}} = \exp[(kR)^2/4]$. Note that, when $kR \ll 1$, $W_{\text{max}} \simeq 1 + (kR)^2/4 \simeq 1$ so that all configurations where $W > 1$ receive nearly the same weight. In this case, the parameter $R$ regulates how quickly the function $W$ drops...
when $q$ moves away from the region where $W > 1$. In other words, $W$ behaves nicely as a smoothing function. This is not true, however, when $kR \gg 1$ and the value of $W_{\text{max}}$ grows very large. In this case, $P_{(2,1)}$ receives dominant contributions from a narrow shell of modes located at $q \approx k/2$ and $\mu \lesssim 1$. This effect is clearly seen in Figure 2 for $R = 18$ $h^{-1}\text{Mpc}$ where the over-smoothing (i.e. the fact that $kR$ is significantly larger than unity for $k \sim 0.1$) leads to a change in shape for $P_{(2,1)}$ which is particularly evident for the configurations with the largest wavenumbers.

It is interesting to investigate why, for $kR \ll 1$, the configuration dependence of $P_{(2,1)}$ changes very little with $R$ and only the overall normalisation appears to depend on the smoothing scale. If we assume that the amplitude of the bispectrum $B(k_1, k_2, -k_1 - k_2)$ keeps nearly constant at all scales assuming a value $\approx B_0$, Equation (33) then gives

$$ P_{(2,1)}(k) \approx \frac{\pi^{3/2}}{R^3} \exp \left( \frac{(kR)^2}{4} \right) \frac{B_0}{R^3} . \tag{35} $$

The first term on the right-hand-side gives the $q$-space volume over which the bispectrum is averaged to get $P_{(2,1)}$. At fixed $k$, this expression diverges as $R^{-3}$ when $R \to 0$ and exponentially as $R \to \infty$ while it shows broad minimum around $kR \sim 2.5$.

Clearly, had we not smoothed the density field, the resulting $P_{(l,m)}$ and $B_{(l,m,n)}$ would be divergent in any $\Lambda \text{CDM}$ cosmology.

### 4.4 Modelling $P_{(2,1)}$ and $B_{(2,1,1)}^{(s)}$ with SPT

In order to better understand what drives the amplitude and functional form of the $P_{(l,m)}$ and $B_{(l,m,n)}^{(s)}$ terms we have attempted to model their signal with SPT. For simplicity, we have focused on the lowest-order non-trivial terms $P_{(2,1)}$ and $B_{(2,1,1)}^{(s)}$.

To leading order in the perturbations, the matter bispectrum can be written as

$$ B(k_1, k_2, k_3) = 2P_2(k_1, k_2)P_0(k_3) + 2 \text{cyc} + P_3(k_1, k_2, k_3) , $$

where $P_0$, $P_2$, and $P_3$ are the power spectra for density fluctuations, two-point correlation function, and three-point correlation function, respectively, and $\text{cyc}$ refers to cyclic permutations of the arguments. The SPT contribution to $B_{(2,1,1)}^{(s)}$ is

$$ B_{(2,1,1)}^{(s)}(k_1, k_2, k_3, \theta_{12}) = \frac{2}{3} \left[ P_0(k_2)P_0(k_3)W(k_2, k_3) + 2 \text{cyc} \right] + \frac{1}{3} \int \frac{d^3q_1}{(2\pi)^3} T(q_1, k_1 - q_1, k_2, k_3)W(q_1, k_1 - q_1)
+ 2 \text{cyc} , \tag{36} $$

where $W$ denotes the matter trispectrum (i.e. the connected part of the 4-point correlator). The SPT contribution to lowest non-vanishing order is simply

$$ B_{(2,1,1)}^{(s)}(k_1, k_2, k_3) \simeq \frac{2}{3} \left[ P_0(k_2)P_0(k_3)W(k_2, k_3) + 2 \text{cyc} \right] . \tag{37} $$

In the right panel of Figure 3 we show that this approximation (star-shaped points) strongly underestimates the outcome from the $N$-body simulations (solid symbols with errorbars) and does not display the same scaling behaviour with $k$ and $R$ as the data. A common approach performed during observational data analysis is to substitute in place of the linear power spectrum, $P_0$, shown in Eq. 37 the fully non-linear power spectrum, $P_{(1,1)}$. We found that perform-
ing this substitution has little effect on the resulting amplitudes, remaining roughly equivalent as the lowest nonvanishing contributions. We then go one step further and compute the next-to-leading-order corrections to \( B_{(2,1,1)} \) which are of sixth-order in terms of the linear density field. This gives

\[
B_{(2,1,1)}^{(0)}(k_1, k_2, k_3) \simeq \frac{2}{3} W(k_2, k_3) \left[ P_{(0)}(k_2)P_{(0)}(k_3) + P_{(1)}(k_2)P_{(1)}(k_3) + P_{(1)}(k_2)P_{(0)}(k_3) \right] + 2 \text{cyc} \\
+ \frac{1}{3} \int \frac{d^3 q_1}{(2\pi)^3} V(q_1, k_1 - q_1) T_{(0)}(q_1, k_1 - q_1, k_2, k_3) + 2 \text{cyc} ,
\]

where \( P_{(1)} \) denotes the first loop correction to the power spectrum (i.e. \( P \simeq P_{(0)} + P_{(1)} + \ldots \)) and the term \( T_0 \) represents the tree-level contribution to the connected trispectrum. In Appendix B we provide the expressions needed for evaluating all these quantities, which are de-smoothed according to Eq. (15). Our final results are shown in Figure 4 (solid lines). The SPT approximation shows the correct scaling with \( R \), but for \( R > 2 \) h\(^{-1}\)Mpc it tends to overpredict the amplitude for collinear (i.e. \( \theta \simeq 0 \) and \( \theta \simeq \pi \)) configurations. For \( 4 < R < 8 \) h\(^{-1}\)Mpc it also underpredicts the amplitude for triangles in which \( k_1 \) and \( k_2 \) are nearly perpendicular. However, as \( R \) increases the discrepancy lessens and at \( R = 10 \) h\(^{-1}\)Mpc SPT performs better. This suggests that using SPT to fit galaxy bispectra in the scale range \( 0.04 \lesssim k \lesssim 0.12 \) h\(^{-1}\)Mpc may possibly lead to seriously biased estimates for the parameters of the LEB.

Nevertheless, whilst the analytic calculations of \( P_{(2,1)} \) and \( B_{(2,1,1)}^{(0)} \) are feasible, computing higher-order terms becomes increasingly challenging. However, estimating these quantities from simulations is no more demanding than measuring the low-order terms and so our approach offers a distinct advantage over the classical SPT calculations.

5 ESTIMATION OF HALO BIAS

5.1 Bayesian parameter estimation

The second-order LEB contains three parameters: \( \theta \equiv (b_1, b_2, R) \). In this section, we use Bayesian statistics to determine their values that best represent the halo power and bispectra extracted from our simulations. For simplicity, we assume that the cosmological parameters are perfectly known and that the measurement errors are Gaussian distributed, i.e.

\[
\mathcal{L}(x|\theta) = (2\pi)^{-N/2} |C|^{-1/2} e^{-\frac{1}{2}(x - \mu(\theta))^T C^{-1}(x - \mu(\theta))} = (39) \\
= (2\pi)^{-N/2} |C|^{-1/2} e^{-\frac{1}{2}(x - \theta)^T},
\]

where \( x^T \) is the \( N \)-dimensional vector containing the power spectra or bispectra for different configurations, \( \mu(\theta) \) is the model prediction and \( C \) is the covariance matrix. In theory \( C \) is a model dependent quantity, however owing to the technical challenge of estimating this matrix and its inverse, we have decided to determine \( C \) directly from the data.

Equation (39) gives the likelihood of the data given the model, but what we need in order to perform parameter estimation is the posterior probability of the model parameters given the data. This can be obtained using Bayes’ theorem:

\[
P(\theta|x) = \frac{\Pi(\theta) \mathcal{L}(x|\theta)}{p(x)}, \quad (40)
\]

where \( \Pi(\theta) \) is the prior probability for the model parameter while the evidence,

\[
p(x) \equiv \int \Pi(\theta) \mathcal{L}(x|\theta) d^3 \theta , \quad \quad (41)
\]

simply acts as a normalizing factor and does not influence the search for the best fit. In what follows we will always assume flat priors on the parameters, but bounded over a finite domain which is much more extended than the likelihood function. Moreover, \( b_1 \) and \( R \) will always be assumed to be positive.

5.2 Covariance matrix estimation

The sample covariance matrix

\[
\hat{\Sigma} \equiv \frac{N}{N - 1} \langle \Delta x^T \Delta x \rangle_N ; \quad \Delta x \equiv x - \langle x \rangle_N , \quad (42)
\]

where \( \langle \ldots \rangle_N \) denotes the arithmetic mean over \( N \) independent measurements, provides an unbiased estimator of the covariance matrix for the measurement errors.

However, this estimator is extremely unstable and inefficient. It generally provides matrices where the smallest eigenvalue is too small and the largest one is too big. Very large samples are thus needed to obtain accurate estimates of the covariance.

On using our ensemble of 200 simulations for both the power and the bispectra, we could measure the diagonal elements of the covariance with an accuracy of \( \sim 10 \) per cent. On the other hand, the off-diagonal elements had a much smaller absolute value and were scattering around zero with errors of the order of \( \sim 100 \) per cent. All this suggests that the covariance should be close to diagonal as expected for a Gaussian random field with infinitesimally narrow bins in \( k \)-space.

Due to these large uncertainties in the off-diagonal elements, we opted for implementing a shrinkage method to better estimate the covariance matrices of our power and bispectra. Shrinkage estimation is a variance reduction technique that shrinks an empirical estimation of the covariance like \( \hat{\Sigma} \) towards a theoretical model for how the covariance should be, represented by a structured matrix \( T \) (the target covariance). The shrunk estimator is given by the convex linear combination

\[
\hat{C} = \lambda T + (1 - \lambda) \hat{\Sigma} \quad (43)
\]

where \( 0 < \lambda < 1 \) is the shrinkage intensity. This ensures the resulting covariance matrix to be positive definite even if \( \hat{\Sigma} \) is singular (because it is determined from \( N \sim \dim(x) \) observations).

It has been demonstrated that shrinkage techniques provide a regularized estimate of the covariance which is both more accurate and statistically efficient than either of the individual estimators \( \hat{\Sigma} \) and \( T \), and they do so in a systematic way (Schäfer & Strimmer 2002). Without the need for specifying an underlying probability distribution, Ledoit & Wolf (2004) provided a theorem that determines the optimal value for \( \lambda \) through minimization of a quadratic
loss function such as the mean-square error of the covariance matrix. This can be expressed in terms of the squared Frobenius norm

\[ L(\lambda) = \| \hat{\Sigma} - \Sigma \|^2_F = \| \lambda T + (1 - \lambda) \hat{S} - \Sigma \|^2_F = \sum_{i,j=1}^{p} (\lambda s_{ij} + (1 - \lambda) s_{ij} - \sigma_{ij})^2. \]  

(44)

which gives a measure of the distance between the true population covariance, \( \Sigma \), and the inferred one, namely, \( \hat{\Sigma} \). The key is to select a suitable target, and we assume it to be a diagonal matrix with unequal variances coinciding with the sample variances:

\[ t_{ij} = \begin{cases} s_{ii}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \]  

(45)

Minimizing Eq. (44) gives the expression for the optimal shrinkage intensity:

\[ \lambda = \frac{\sum_{j>1} \text{Var}(s_{ij})}{\sum_{j>1} \text{Var}(s_{ij}) + \sigma_{jj}^2} = \frac{\sum_{j>1} \text{Var}(s_{ij})}{\sum_{j>1} E(s_{ij}^2)}, \]  

(46)

where \( E(...) \) denotes the expectation value of a random variable. Following Schäfer & Strimmer (2002), we estimate the sampling variance of the elements of the sample covariance using

\[ \text{Var}(s_{ij}) = \frac{N}{(N-1)^3} \sum_{j=1}^{N} \left( \Delta x_j^T \Delta x_j - \langle \Delta x^T \Delta x \rangle_N \right) s_{ij}^2. \]  

(47)

However, while these authors approximate \( E(s_{ij}^2) \) in Eq. (46) with the square of the point estimate \( s_{ij} \) thus overestimating \( \lambda_s \), we adopt the square of the sample covariances \( s_{ij}^2 \) as a proxy for \( \sigma_{ij}^2 \) (e.g. Kwan 2011). In all cases, we found that the optimal shrinkage intensity was roughly \( \lambda_s \sim 0.45 \) for the power spectra covariance and \( \lambda_s \sim 0.23 \) for the bispectra covariance, respectively. Note that the adopted algorithm only performs shrinkage of the off-diagonal elements of the covariance matrix.

5.3 Constraining the bias parameters: \( b_1 \), \( b_2 \) and \( R \)

We now determine the best-fit model parameters for the various power and bispectra that we have estimated from the simulations within the scale range \( 0.04 < k < 0.12 \) h Mpc\(^{-1}\). We consider two second-order LEB models that differ in the polynomials describing the non-linear matter distribution (see below for the details). In both cases, we map the likelihood function within a finite volume of the parameter space that we slice into a regular Cartesian mesh.

5.3.1 SPT tree-level model

The first model uses SPT at the lowest non-vanishing order to approximate the \( P(l,m) \) and \( B(l,m,n) \). This is what is most commonly done in the literature. For the power spectrum, the \( P(l,m) \) terms expressed at tree-level of SPT are:

\[ P_{(1,1)}^{\text{tree}} = P_{(0)}(k), \]  

(48)

\[ P_{(i,j)}^{\text{tree}} = 0 \text{ for } i + j > 2, \]  

(49)

where \( P_{(0)}(k) \) denotes the linear matter power spectrum.

Thus, the bias relation is linear and carries no dependence on the filter scale, \( R \), and on \( b_2 \).

For the bispectrum, the evaluation of \( B(l,m,n) \) using only tree-level contributions gives:

\[ B_{(l,m,n)}^{(s),\text{tree}} = 2 P_{(0)}(k_1) P_{(0)}(k_2) F_2(k_1, k_2) + 2 \text{ cyc}, \]  

(50)

\[ B_{(l,m,n)}^{(s),\text{tree}} = \frac{2}{3} P_{(0)}(k_1) P_{(0)}(k_2) W(k_1, k_2) + 2 \text{ cyc}, \]  

(51)

\[ B_{(i,j,k)}^{(s),\text{tree}} = 0 \text{ for } i + j + k > 4, \]  

(52)

where \( F_2(k_1, k_2) \) is the second-order mode-coupling kernel function from SPT (e.g. Bernardeau et al. 2002).

5.3.2 Fully non-linear model

The second model considers the fully non-linear matter polynomials extracted from the simulations. Note that, while evaluating the LEB when varying \( b_1 \) and \( b_2 \) at fixed \( R \) is a trivial exercise, varying \( R \) would, in theory, require recomputing all the relevant \( P_{l,m} \) and \( B_{l,m,n} \). However, as we mentioned earlier in our discussion of Fig. 3, these functions change smoothly with \( R \). We therefore use a cubic-spline interpolation of \( \log [P_{l,m}] \) and \( \log [B_{l,m,n}] \) to model the \( R \)-dependence of the theory. This enabled us to map the likelihood function with arbitrary resolution.

A final comment is in order regarding the details of how the fit is performed. There is some arbitrariness in defining what exactly are the “observables” and what is the “model” in the simulations. For instance, we could have fit the outcome of each \( N \)-body simulation separately using the polynomials extracted from the very same realization. While being a valid test of the LEB, this method would have not been a valid test of the LEB, this method would have not had much in common to actual galaxy redshift surveys (or even to the SPT model discussed above), where the underlying mass distribution is unknown and needs to be modeled independently. In fact, the presence of the same noise structure in the matter and halo power and bispectra would result in overfitting. There are a couple of alternative approaches one could follow to prevent this. The first is to generate smooth versions of the \( P_{l,m} \) and \( B_{l,m,n} \) terms by averaging over the entire ensemble of simulations. One can then use these “theoretical models” to simultaneously fit the halo statistics extracted from all of our 200 independent realizations. The other alternative is to subdivide the total ensemble of simulations into two subsets, where one subset would be used to construct the smooth \( P_{l,m} \) and \( B_{l,m,n} \) terms by averaging over the total number of simulations in the subsample and the other subset would serve as the halo statistics to be analyzed. The partition of the ensemble of realizations into two distinct subsets ensures that the “model” and “data” are indeed independent. Furthermore, one can exchange the roles of “model” and “data” for the two subsets and then sum the \( \chi^2 \)s obtained from the two sets of analysis. We carried out both approaches however we only report the results from averaging over the 200 simulations as the bias model constraints compared to the partitioning approach are in extremely good agreement.
5.4 Goodness of fit

In this Section we use the classic $\chi^2$ goodness-of-fit test to quantify how well the second-order LEB fit our simulated data. We minimise the $\chi^2$ function over the parameter space using the simplex method. The best-fit models determined this way basically coincide with those that minimise the $\chi^2$ function in the dense grid used for our Bayesian analysis. Since for all power and bispectra we have always used 20 bins in $k$ or $\theta$ and the covariance matrices are full rank, the number of degrees of freedom totalled $\nu = 200 \cdot 20 - 3 = 3997$ for each fit.

5.4.1 Power spectra

The tree-level SPT models for the halo power spectra provide very poor fits to our data. The minimum $\chi^2$ value is much larger than the number of degrees of freedom, reaching $\chi^2_{\text{min}} \simeq 7465$ for $P_{\text{hm}}$ and $\chi^2_{\text{min}} \simeq 139,821$ for $P_{\text{hh}}$. These results may serve as indicators that halo biasing is non-linear and/or a result of the breakdown of linear SPT. To check the latter, we re-fit both spectra using Equation (48) but after replacing $P_{\text{tree}}^{(1,1)}$ by the fully non-linear matter power spectrum $P_{\text{tree}}^{(1,1)}$. In this case, we acquire $\chi^2_{\text{min}} \simeq 3903$ for $P_{\text{hm}}$ and $\chi^2_{\text{min}} \simeq 4442$ for $P_{\text{hh}}$. This significant improvement to the tree-level results demonstrates the need to model non-linearities in the matter distribution very accurately. Using the fully non-linear model with the additional free-parameters $b_2$ and $R$ only slightly improves the goodness of fit for $P_{\text{hm}}$, giving $\chi^2_{\text{min}} \simeq 3901$. On the other hand, the improvement is marked for $P_{\text{hh}}$ for which we obtain $\chi^2_{\text{min}} \simeq 3915$.

It is interesting to see how the $\chi^2_{\text{min}}$ value changes when $R$ is kept fixed. In this case, we find that all the fits to $P_{\text{hm}}$ are equally good. However, for $P_{\text{hh}}$, the values of $\chi^2_{\text{min}}$ undergo a sharp decrease (from 3944 $\lesssim \chi^2_{\text{min}} \lesssim 3921$) for $2 < R \lesssim 3.66 \, h^{-1}\text{Mpc}$, then decrease slowly to the absolute minimum value at $R \sim 13.2 \, h^{-1}\text{Mpc}$ and finally begin to slowly rise again to our cutoff scale of $R = 18 \, h^{-1}\text{Mpc}$. Hence, it appears that there is a range of preferred smoothing scales that best fit the simulation data for $P_{\text{hh}}$.

5.4.2 Bispectra

Turning now to the bispectra, we find that the fully non-linear model provides slightly better fits to the numerical data ($\chi^2_{\text{min}} \simeq 3906,3908$ and 3913 for $B_{\text{hm},h}$, $B_{\text{hm},mh}$, and $B_{\text{hh},h}$, respectively) than the tree-level model ($\chi^2_{\text{min}} \simeq 3923,3922$ and 3925) which, however, already supplies $\chi^2_{\text{min}}/\nu \lesssim 1$.

In all cases, if we keep $R$ fixed and only consider 2-parameter models, we find that the $\chi^2_{\text{min}}$ value does not change much for $2 < R < 13 \, h^{-1}\text{Mpc}$ while it rapidly grows adopting larger smoothing scales. In terms of goodness of fit, the non-linear model for $B_{\text{hm},h}$ outperforms the tree-level SPT model for all values of $R$. On the other hand, when $B_{\text{hm},h}$ and $B_{\text{hh},h}$ are considered, the non-linear model gives a better fit only for $R \lesssim 15 \, h^{-1}\text{Mpc}$.

5.4.3 Posterior mean

In order to give a visual impression of the best-fit models, in Figure 6 we show the posterior mean (black line) and the posterior rms error (shaded gray region) for the halo power and bispectra resulting from our fits with the fully non-linear model in comparison with the simulation data. In all cases, the models agree with the simulations remarkably well. Note that the rms error on the best-fit models for $P_{\text{hh}}$ and $P_{\text{hm}}$ is hardly visible on the scale of the plots.

5.5 Bias from the power spectrum

5.5.1 Effective bias

Due to its highly compressed ordinate, Figure 5 gives the true $b_{\text{eff}}$ $k$ dependence ratio $P_{\text{hm}}/P_{\text{mm}}$ (orange symbols) and $b_{\text{hh}} = (P_{\text{hh}}/P_{\text{mm}})^{1/2}$ (blue symbols) extracted from our simulations as a function of the wavenumber. The black solid line and shaded regions indicate the mean and the rms value of the effective bias obtained by averaging the predictions of the second-order LEB over the posterior probability of the model parameters.

Figure 5. The effective halo-bias parameters $b_{\text{hm}} = P_{\text{hm}}/P_{\text{mm}}$ (orange symbols) and $b_{\text{hh}} = (P_{\text{hh}}/P_{\text{mm}})^{1/2}$ (blue symbols) extracted from our simulations as a function of the wavenumber. The black solid line and shaded regions indicate the mean and the rms value of the effective bias obtained by averaging the predictions of the second-order LEB over the posterior probability of the model parameters.

In order to give a visual impression of the best-fit models, in Figure 6 we show the posterior mean (black line) and the posterior rms error (shaded gray region) for the halo power and bispectra resulting from our fits with the fully non-linear model in comparison with the simulation data. In all cases, the models agree with the simulations remarkably well. Note that the rms error on the best-fit models for $P_{\text{hh}}$ and $P_{\text{hm}}$ is hardly visible on the scale of the plots.

$1$ Very similar results are obtained if one averages the ratios instead of taking the ratio of the averages.
roughly constant while it shows a significant scale dependence for $k > 0.06 \ h \ \text{Mpc}^{-1}$, whereas $b_{hm}$ shows the opposite trend although its scale dependence is weaker for the large scales. At $k \approx 0.04 \ h \ \text{Mpc}^{-1}$, $b_{hm}$ and $b_{hh}$ assume very similar values. However, $b_{hm} > b_{hh}$ for all wavenumbers. Our high-quality data also provide some hints for the presence of weak oscillatory features in the effective bias parameters on the scales of baryonic acoustic oscillations.

Figure 6 also tests how the fully non-linear second-order LEB model is able to reproduce the scale-dependence of $b_{hm}$ and $b_{hh}$ in fine details. The black curves represent the posterior mean of the effective bias coefficients and the shaded grey regions denote the corresponding rms value of their posterior distribution. Although the models are not able to reproduce all the features which are present in the numerical data, they are in reasonable agreement with the simulations, especially for $k > 0.08 \ h \ \text{Mpc}^{-1}$. Nevertheless, we see that for both $b_{hm}$ and $b_{hh}$ the power spectrum models actually are less accurate at small $k$ (i.e. on the large scales) in the proximity of the point where the trend from constant-to-scale dependence (and vice versa) occurs. On these scales, the models systematically overpredict the effective biases and the largest discrepancy is of the order of $\sim 0.3$ per cent.

### 5.5.2 Marginal credible regions

Now we compare the level of the consistency between the model-parameter constraints deriving from the fits to the halo power spectra, $P_{hh}$ and $P_{hm}$. Figure 8 shows (from left to right) the marginal posterior distributions for the parameter pairs $b_1–b_2$, $b_1–R$ and $b_2–R$ of our fully non-linear model. The black and green contours denote the 68.3\% and 95.4\% credible regions for the parameters of the LEB obtained from analyzing $P_{hh}$ and $P_{hm}$, respectively. The first apparent observation is that the contours of $P_{hm}$ and $P_{hh}$ span different regions of the parameter space: while the $P_{hm}$ data prefer $b_1 \lesssim 1.5$ and $b_2 \geq 0$, $P_{hh}$ favours $b_1 \gtrsim 1.5$ combined with $-0.15 \lesssim b_2 \lesssim -0.2$. In other words, the second-order LEB model provides a successful fit to $P_{hh}$ or $P_{hm}$ but requires two incompatible parameter sets. Improper modeling of the shot noise in $P_{hh}$ might be the primary cause of the inconsistency (e.g. Hamaus et al. 2010). Note, however, that the best-fit values for $b_1$ and $b_2$ that we derive from $P_{hh}$ are in good agreement with the predictions of theories that follow the collapse of dark-matter halos (e.g. Equation (14) and (15) in Scoccimarro et al. 2003). It is also worth mentioning that, for Gaussian fluctuations in the matter density, the cross-spectrum of locally-biased tracers is always exactly proportional to $P_{hm}$ even though this is not apparent from the mathematical formulation of the LEB (Frusciante & Sheth 2012). The fact that our measurement of $P_{hm}$ needs $b_2 \sim 0$ might simply suggest that a similar relation holds true also in the presence of non-Gaussian perturbations (at least approximately, since $b_{hm}$ keeps nearly constant with $k$ as shown in Fig. 5).

### 5.6 Bias from the bispectrum

#### 5.6.1 Effective bias

To investigate the bias relation as a function of scale using the halo bispectra, we define a set of coefficients by taking the following ratios: $b_{hm} = \langle P_{hm} \rangle_{200}/\langle P_{mm} \rangle_{200}$, $b_{hh} = \langle \langle P_{hh} \rangle_{200}/\langle P_{mm} \rangle_{200} \rangle^{1/2}$, $b_{hhh} = \langle \langle \langle P_{hhh} \rangle_{200}/\langle P_{mmm} \rangle_{200} \rangle \rangle^{1/3}$. The results are shown in Figure 7 all the effective bias coefficients present a characteristic configuration dependence and are in agreement within their 1$\sigma$ uncertainties (although $b_{hhh}$ tends to assume slightly higher values for all triangle configurations). The posterior means of the effective bias coefficients from the fully non-linear models also closely match the data as expected from the $\chi^2$ test presented in 5.4.2 All this suggests the second-order LEB provides a suitable description of the bias relation at the three-point level.

#### 5.6.2 Marginal credible regions

We now evaluate the consistency of the model-parameter constraints for the halo bispectra. Figure 8 shows the marginal posterior distribution for the parameter pairs $b_1–b_2$, $b_1–R$ and $b_2–R$, respectively. Each panel refers to a particular bispectrum, as indicated from the label in the bottom left corner. The black contours denote the 68.3\% and 95.4\% credible regions for the parameters of the fully non-linear
model. The red contours, instead, indicate the corresponding regions for the SPT tree-level model described in §5.3.1.

The first thing that may be noticed is that the estimates for $b_1$ and $b_2$ from the tree-level and fully non-linear models are in disagreement: the tree-level constraints show a systematic shift, preferring lower $b_1$ and slightly more negative $b_2$ values. This implies that inferences made about the non-linearity of galaxy bias using the galaxy bispectrum and tree-level perturbation theory will be significantly biased. Note that this statement also applies to rather large scales ($k_1 \lesssim 0.12\, h$ Mpc$^{-1}$). If one uses triangle configurations on smaller scales (e.g. Verde et al. 2002) then the discrepancy becomes larger. Therefore, our program to use N-body simulations for determining the matter terms in the bias relation is key to correctly estimate the bias (and thus the cosmological parameters) from forthcoming observational data.

The second important point to notice is that fits to $B_{hhm}$, $B_{hmm}$ and $B_{hmm}$ with the fully non-linear model give consistent constraints for $b_1$, $b_2$ and $R$. The precision with which we are able to determine the bias parameters increases as we go from $B_{hmm}$ to $B_{hhm}$. This finding is consistent with our earlier results (Pollack, Smith & Porciani 2012).

Note that the best-fit values for $b_1$ appear relatively insensitive to variations in $R$. One has to consider rather large smoothing scales in order to see any changes in the best-fit models. However, the best-fit solutions for $b_2$ are strongly degenerate with $R$: as $R$ decreases, $b_2$ becomes less negative and tends towards zero. This owes to the fact that, on changing $R$ in the interval $R \in [2, 13]\, h^{-1}$ Mpc, one can always find different combinations of $b_1$ and $b_2$ that fit the data with the same accuracy as previously described in §4. We can more directly understand the origin of the $b_2 - R$ degeneracy as follows. Let us consider the bias model for $B_{hmm}$ since this only contains the terms $B_{(1,1,1)}^{(s)}$ and $B_{(2,1,1)}^{(s)}$. As shown in the top panels of Figure 4, the de-smoothing of $B_{(1,1,1)}^{(s)}$ results in the matter bispectrum. However, the de-smoothing of $B_{(2,1,1)}^{(s)}$ results in a function that carries a dependence on $R$. If we take the ratio of $B_{(2,1,1)}^{(s)}$ defined for $R_a$ with the same function but defined for $R_b$, then we will find something close to a constant for $R = [4, 10]\, h^{-1}$ Mpc. Thus we can identify the degenerate combination:

$$b_2^s = \frac{B_{(1,1,1)}^{(s)}}{B_{(2,1,1)}^{(s)}}(k_1, k_2, R_a) b_2^b \approx A(R_a, R_b) b_2^b \quad (53)$$

where $A(R_a, R_b)$ is a function that is independent of the triangle configuration. Hence, the value of $b_2$ is correlated with the size of the smoothing scale, $R$.

### Table 2. Posterior mean and rms error of the bias parameters $b_1$, $b_2$ and $R$ obtained fitting various halo statistics with the full non-linear bias model.

| Statistic | $b_1 \pm \sigma_{b_1}$ | $b_2 \pm \sigma_{b_2}$ | $R \pm \sigma_R$ ($h^{-1}$Mpc) |
|-----------|--------------------------|--------------------------|----------------------------------|
| $P_{hh}$  | 1.53 ± 0.02              | -0.18 ± 0.02             | 12.0 ± 3.1                       |
| $P_{hhm}$ | 1.48 ± 0.02              | 0.02 ± 0.01              | 10.6 ± 4.1                       |
| $B_{hhm}$ | 1.49 ± 0.03              | -0.18 ± 0.07             | 7.2 ± 2.6                        |
| $B_{hmm}$ | 1.51 ± 0.03              | -0.26 ± 0.10             | 7.8 ± 2.8                        |
| $B_{hmm}$ | 1.52 ± 0.05              | -0.31 ± 0.14             | 9.1 ± 3.1                        |
show the results from the tree-level perturbation theory for the bispectra. The corresponding mean and rms values of the marginal probabilities for the full non-linear model are reported in Table 2.

Considering the values for $b_1$ from the bispectra, we see that, as noted earlier, the parameter constraints for the non-linear model are consistent with one another and are significantly different from the best-fit $b_1$ obtained from the tree-level expressions. On comparing the bispectra results with the power-spectra results we find reasonable consistency for the non-linear modelling, whereas for the tree-level bispectrum model, the results disagree at high significance (see also Pollack, Smith & Porciani 2012). However, the marginal distributions for $b_1$ from $P_{\text{hh}}$ and $P_{\text{hm}}$ overlap very little. In fact, they exhibit opposite skewness although they are both narrow and located around $b_1 \simeq 1.5$. The marginal distribution for $b_1$ computed from $P_{\text{hm}}$ agrees remarkably well with the effective bias $b_{\text{hm}} = 1.503 \pm 0.002$. This is because the data require $b_2 \simeq 0$ in this case. On
A new method to measure galaxy bias

Figure 9. Marginal probability distributions for the single bias parameters $b_1$ (left), $b_2$ (center) and $R$ (right) obtained fitting various halo statistics (from top to bottom: $P_{hh}$, $P_{hm}$, $B_{hhh}$, $B_{hhm}$, $B_{hmm}$). Results obtained with the full non-linear model (black) are compared with those derived using tree-level SPT (red).

the other hand, $b_1 > b_{hh} = 1.49 \pm 0.002$ in the marginal distribution extracted from $P_{hh}$ which requires $b_2 < 0$.

Examining the results of the fully non-linear model for $b_2$, from the bispectra we find that the marginal posterior distributions are fairly broad and are peaked towards negative values ($b_2 \approx -0.2$ for $B_{hhh}$, $b_2 \approx -0.3$ for $B_{hhm}$ and $B_{hmm}$). Overall, the various bispectra give consistent constraints. Note that the sharp cutoff in the marginal distributions at $b_2 \approx 0$ is due to the fact that our prior for $R$ does not consider values $R < 2 \, h^{-1}\text{Mpc}$. Considering the results obtained using the tree-level SPT model, we see that the distributions for $b_1$ and $b_2$ shift towards different values (approximately the posterior mean of the bias parameters moves by $\Delta b_1 \approx \Delta b_2 \approx -0.15$). On comparing with the results obtained from the halo power spectra, we see that the marginal distribution for $b_2$ extracted from $P_{hh}$ and $P_{hm}$ are narrowly peaked around $b_2 \approx -0.18$ and $b_2 \approx 0.02$, respectively.

We now turn to the question of whether there is a preferred smoothing scale for the haloes we have considered. On inspecting the bispectra, we see that the marginal distributions for $R$ are reasonably consistent and display a broad peak between 5 and $12 \, h^{-1}\text{Mpc}$. The power spectra, instead, tend to prefer slightly larger values of $R$: $10 < R < 15 \, h^{-1}\text{Mpc}$ for $P_{hh}$ and $R > 5 \, h^{-1}\text{Mpc}$ for $P_{hm}$, consistent with the behaviour of the goodness of fit reported in [5.4].

In all cases, these optimal smoothing scales correspond to a few Lagrangian radii of the halos. They are also comparable to (but a bit smaller than) the scales that we sample with the measurements of the power spectra and bispectra. Note that a sphere of radius $\sim 10 \, h^{-1}\text{Mpc}$ contains $\sim 1.5$ halos on average so that counts in cells of this extension are subject
to sizable random fluctuations that create stochasticity in the bias relation.

5.8 Cross-correlation coefficients

There are three possible explanations as to why $P_{hm}$, $P_{hh}$ and the bispectra show disagreement for the full non-linear model. One, we may require higher-order terms in the bias expansion, e.g. $b_3$ etc; two, the LEB may be wrong; three, there may be uncorrelated stochasticity in the relation between halo overdensity and mass overdensity. We shall now explore this latter possibility.

A number of studies have demonstrated, using N-body simulations, that the relation between $\delta_h(x|R)$ against $\delta(x|R)$ contains scatter, and that this scatter depends on the scale which one uses to compute the density field (e.g. Dekel & Lahav 1993; Sfakianakis & Warren 2004; Smith, Scoccimarro & Sheth 2007; Manera & Gaztañaga 2011; Roth & Porciani 2011; Pollack, Smith & Porciani 2012; Chuen Chan & Scoccimarro 2012).

Another way to explore the stochasticity is through the cross-correlation coefficient between Fourier modes. For two-point statistics this can be defined (Dekel & Lahav 1993):

$$ r \equiv \frac{\langle \delta_{hm} \delta_{hh} \rangle}{\sqrt{\langle \delta_{hm}^2 \rangle \langle \delta_{hh}^2 \rangle}} = \frac{b_{hm}}{b_{hh}}. $$

(54)

If $\delta^b$ is a deterministic linear function of $\delta$, then $r = \pm 1$. However, if there is uncorrelated random noise present, i.e. $\delta^b(x) = b_0 \delta(x) + \epsilon(x)$, then the halo power spectrum would be $P_{hh}(k) = b_0^2 P_{mm}(k) + b_1(k)$, where $P_\ell$ denotes the power spectrum of the noise distribution. This leads to:

$$ r = \left( 1 + \frac{P_{1}}{P_{mm}} \right)^{-1/2} < 1. $$

(55)

We note that non-linearity in the bias relationship will also introduce deviations of $r$ away from unity: consider the quadratic relation $\delta_h(x) = b_1 \delta(x) + b_2 \delta^2(x)/2$, then one finds that the cross-correlation can be written:

$$ r = \left[ 1 + \frac{c_2 P_{2(2,1)}}{2 P_{(1,1)}} \right]^{1/2} \left[ 1 + \frac{c_2 P_{2(2,1)}}{2 P_{(1,1)}} + \frac{c_2^2 P_{2(2,2)}}{4 P_{(1,1)}} \right]^{-1/2} \approx 1 - \frac{c_2^2 P_{2(2,2)}}{8 P_{(1,1)}}. $$

(56)

(57)

where $c_2 \equiv b_2/b_1$ and where the second equality follows for the case where $P_{2(1,1)} \ll P_{(1,1)}$ and $P_{2(2,1)} \ll P_{(1,1)}$.

In this case, we see that the cross-correlation function can be either greater or less than unity depending on the sign and magnitude of $c_2$.

Figure 10 shows the cross-correlation coefficient estimated from our ensemble of N-body simulations along with the standard errors on the mean. The open symbols show the result before we correct $P_{hh}$ for shot noise, the solid symbols show the result after the usual inverse number-density correction. We see that before correcting for the shot noise the function is less than 1 and decreases with scale. After the correction, $r$ is brought within a few percent from unity and is always larger than one. Note that the difference from unity is very significant given the numbers of realizations and the comoving volume covered by our simulations.

In order to derive $r$ from the fully non-linear model, we jointly fit the numerical data for $P_{hm}$ and $P_{hh}$. We acknowledge that utilizing 200 simulations prevents us from accurately estimating a $40 \times 40$ covariance matrix, in particular the cross covariances between the different spectra. Therefore, we performed the joint fit in the following manner. In order to ensure that the different spectra can be treated as independent of each other, we generated two ensembles consisting of 100 simulations each to estimate a particular spectrum. We then computed a 20x20 block covariance matrix, selecting every other bin from our auto- and cross-power spectrum estimates. The off-diagonal blocks of the covariance matrix were set equal to zero. The resulting best-fit model ($b_1 \approx 1.5, b_2 \approx -0.09, R \approx 18$) does not match to the data ($\chi^2_{\text{min}} \approx 2242/1997$ with a contribution of 1170 coming from $P_{hm}$) meaning that it is impossible to simultaneously fit $P_{hh}$ and $P_{hm}$ with the second-order LEB. Consequently, we find that the posterior mean of the cross-correlation coefficient, obtained by multiplying the likelihoods of the single fits to $P_{hh}$ and $P_{hm}$, is always smaller than unity and does not provide a good description to the data (see the black line and the shaded region in Figure 10).

To investigate this further, we recompute $r$ using the posterior means of $P_{hh}$ and $P_{hm}$ shown in the left panel of Figure 1. Inserting them in Equation (54), we find excellent agreement with the data (see the blue line in Figure 1).

As shown previously, the best-fit models for $P_{hh}$ and $P_{hm}$ prefer different values for $b_1$ and $b_2$ when analyzed independently. The joint analysis of $P_{hm}$ and $P_{hh}$, in this manner, shows more clearly the inconsistency obtained when using the second-order LEB as a model for halo biasing.

One can also define cross-correlation coefficients for higher-order statistics. The second equality in Eq. (54) gives us a clear path to make this generalization. From the 3-point effective bias coefficients we may form two independent ratios:

$$ r_1 \equiv \frac{b_{hhm}}{b_{hhh}} = \frac{B_{hhm}}{B_{mm}^{2/3} B_{hh}^{1/3}} $$

(58)

$$ r_2 \equiv \frac{b_{hmh}}{b_{hhh}} = \frac{B_{hmh}}{B_{mm}^{1/6} B_{hh}^{1/3}}, $$

(59)

where the dependence on the triangle configuration is understood. Note, that a third (dependent) ratio may be also computed: $r_3 = b_{hhm}/b_{hmh} = r_1/r_2$. For a deterministic linear bias model with bias coefficient $b$, $r_1 = 1$ and $r_2 = \text{sgn}(b)$. Once again, additional stochasticity or non-linear biasing will alter the cross-correlation coefficients.

In the right panel of Figure 10 we present the cross-correlation coefficients $r_1$ and $r_2$ extracted from the simulations as a function of the triangle configuration. Both functions are always a few per cent below unity even after shot-noise subtraction. In the same figure, we also plot the posterior mean and variance for the $r$ coefficients obtained from joint fits to two bispectra (black line and shaded region) performed in the same manner as for the power spectrum. These results are very close to unity and do not adequately describe the simulated data. In fact, the joint fits prefer less negative values for $b_2$ than the single fits (for example, the best simultaneous fit to $B_{hh}$ and $B_{hm}$ gives $b_1 \approx 1.50, b_2 \approx -0.15$ and $R \approx 5.5$ with $\chi^2_{\text{min}}/\nu \approx 1954/1997$). On the other hand, if we compute $r_1$ and $r_2$ from the individual posterior means
A new method to measure galaxy bias

Figure 10. Left: Linear cross-correlation coefficient between the fluctuations in the halo and matter density, \( r = b_{hm}/b_{hh} \), for different wavenumbers. Closed and open symbols show the results obtained from the simulations when \( P_{hh} \) is and is not corrected for shot noise, respectively. The black solid line and the shaded region around it indicate the mean and the rms value of the correlation coefficient obtained by averaging the predictions of the second-order LEB over the posterior probability of the model parameters derived from a joint fit to \( P_{hh} \) and \( P_{hm} \). The solid curve, instead, shows the values of \( r \) that are computed using the means for \( P_{hh} \) and \( P_{hm} \) over the posterior distributions for the individual fits to \( P_{hh} \) and \( P_{hm} \), respectively. Right: As in the left panel but for the 3-point coefficients \( r_1 \) and \( r_2 \) defined in Equations (58) and (59). In this case, the shaded region is obtained averaging the model over the posterior distribution for the parameters derived from a joint fit to the relevant bispectra, while the solid curve uses the different means from the fits to the individual bispectra.

Figure 11. The halo bispectrum, \( B_{hhh} \), measured from the simulations (solid symbols) is compared with the fully non-linear LEB adopting the parameters that best fit \( B_{hmm} \). The line and shaded region show the mean and rms value of the LEB model for \( B_{hhh} \) averaged over the posterior distribution for \( b_1, b_2 \) and \( R \) coming from a fit to \( B_{hmm} \). This shows that the parameter sets that nicely fit \( B_{hmm} \) (see Figure 4) are not able to reproduce all the features seen in \( B_{hhh} \).

6 DISCUSSION

Our high-precision measurements of the halo-halo and halo-matter spectra and bispectra enabled us to carry out a series of consistency tests of the second-order LEB. As seen in Figure 9 and in Table 2, the marginal posterior distributions for \( b_1 \) and \( b_2 \) determined from \( P_{hh}, B_{hhh}, B_{hmm} \) and \( B_{hmm} \) are all consistent with one another and, yet, are inconsistent with the constraints derived from the halo-matter cross spectrum. The primary reason is that \( P_{hm} \) requires a positive \( b_2 \) that is close to 0, whereas the fits to the other spectra prefer a negative value for \( b_2 \). In terms of statistical significance, the stronger discrepancy is with \( P_{hh} \) as the posterior distributions for \( b_2 \) extracted from all bispectra are rather for \( B_{hhh}, B_{hmm} \) and \( B_{hmm} \), we get results that are in good agreement with the data. This is somewhat puzzling as the fits to the various bispectra appear to give consistent bias parameters. However, in order to test how congruous the different fits really are, we derive models for one bispectrum type (say \( B_{hmm} \)) averaging over the joint posterior distribution for the bias parameters derived by fitting one of the other bispectra (\( B_{hmm} \) or \( B_{hhh} \)). An example is shown in Figure 11: the fit based on \( B_{hmm} \) matches well the data for \( B_{hhh} \) for collinear triangles but systematically underestimates the halo bispectrum in all the other configurations. It is exactly in this more precise comparison that we see the failure of the non-linear local bias model when analyzing the bispectra data.
broad. The incompatibility between the bias parameters obtained from \( P_{hh} \) and \( P_{hm} \) might indicate a breakdown in the modelling due to either incorrect shot-noise subtraction or incorrect parameterization of halo biasing. To better understand this issue, it is interesting to focus for a moment onto the shot-noise free spectra \( P_{hm} \) and \( B_{hmm} \). Comparing their mathematical expressions given in Equations (17) and (21), we see that they have the same parametric form in terms of \( b_1 \) and \( b_2 \); it is only the non-linear matter terms which are different. Since we directly measure these terms from the simulations and shot-noise does not play any role here, the fact that the model-parameter constraints from \( P_{hm} \) and \( B_{hmm} \) are incompatible suggests that the LEB truncated to second-order is incorrect or, at the very least, incomplete. The simplest improvement would be to consider higher-order terms in the bias expansion given in Equation (3). However, there are good reasons to believe that more sophisticated corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. Recent numerical work has provided strong evidence that dark-matter halos form out of linear corrections are needed. 

We use a set of 200 N-body simulations to study the clustering properties of dark-matter halos and their relation to the underlying matter distribution with unprecedented accuracy. Our halo catalogs cover a total comoving volume of \( 675 h^{-3} \text{Gpc}^3 \), much larger than the effective volume of the SDSS LRG sample (0.26 \( h^{-3} \text{Gpc}^3 \)), the BOSS BAO sample (2.4 \( h^{-3} \text{Gpc}^3 \)) and the planned spectroscopic survey of the Euclid satellite (19.7 \( h^{-3} \text{Gpc}^3 \)). We consider halos with mass \( M > 1.11 \times 10^{13} h^{-1} M_\odot \) corresponding to a number density of \( 3.7 \times 10^{-4} h^{-3} \text{Mpc}^3 \) so that the effective volume (i.e. the actual volume weighted by the factor \( n_{hh} \)) roughly coincides with the total volume for the wavenumbers analyzed here (0.04 \( \leq k \leq 0.12 h \text{Mpc}^{-1} \)) that match the observable scales of current and future surveys. All this allows us to measure the halo power spectrum to sub-percent accuracy (better than 0.3 per cent at \( k \approx 0.04 h \text{Mpc}^{-1} \)) and the halo bispectrum to a few per cent accuracy.

We make a twofold use of our simulations: to measure the moments of the non-linear matter distribution on several scales (and compare them against SPT predictions) and to test how well the LEB truncated to quadratic order fits several statistics of the halo distribution. In particular, we consider the halo power spectrum, \( P_{hh} \), the halo-mass cross-spectrum, \( P_{hm} \), as well as all the possible bispectra \( B_{hhh}, B_{hmm}, B_{hmh} \). Our main results can be summarized as follows:

(i) In a ΛCDM model at \( z = 0 \), tree-level SPT does not accurately model non-linearities in the momenta of the matter distribution on spatial scales of the order of \( 10 - 30 h^{-1} \text{Mpc} \).

(ii) The simple second-order LEB fits very well all halo spectra and bispectra when either N-body simulations or tree-level SPT are used in the modelling of the clustering amplitudes for the matter distribution. However, the bias parameters derived from the models based on SPT are heavily biased with respect to the case when non-linearities are accurately modelled. This might explain why studies that interpreted different statistics of the galaxy distribution based on SPT reached inconsistent conclusions regarding the non-linear bias of optically selected galaxies (e.g. Verde et al. 2002; Gaztañaga et al. 2004).

(iii) The LEB models applied to counts in cells requires an optimal smoothing scale of several \( h^{-1} \text{Mpc} \) to match the halo statistics from the simulations. For our halos, this corresponds to a few Lagrangian radii but is also of the same order of the spatial scales under analysis.

(iv) Comparing the parameter constraints for the fully non-linear LEB obtained from the different spectra, we find some inconsistencies. In particular, the non-linear bias parameter extracted from the cross-spectrum \( P_{hm} \) is incompatible with the results from all the other statistics. The main difference is that \( P_{hm} \) strongly favours a positive value for \( b_2 \) that is very close to zero, whereas the posterior distributions derived from all other spectra prefer a negative \( b_2 \) in the range \( -0.3 \lesssim b_2 \lesssim -0.2 \). General agreement, instead, is found for the linear bias parameter, \( b_1 \).

(v) Non-trivial shot-noise corrections in \( P_{hm} \) might be invoked to reconcile the bias-parameters extracted from \( P_{hm} \) and \( P_{hh} \). However, this cannot explain the differences between the constraints from the shot-noise free statistics \( P_{hm} \) and \( B_{hmm} \). This suggests that further complexity should be
A new method to measure galaxy bias

Heavens A. F., Matarrese S., Verde L., 1998, MNRAS, 301, 797
Hikage C., Matsubara T., Suto Y., Park C., Szalay A. S., Brinkmann J., 2005, Publications of Astronomical Society of Japan, 57, 709
Jeong D., Komatsu E., 2008, ArXiv e-prints, 805
Jing Y. P., Börner G., 2004, ApJ, 607, 140
Kayo I. et al., 2004, Publications of Astronomical Society of Japan, 56, 415
Komatsu E. et al., 2009, ApJS, 180, 330
Kulkarni G. V., Nichol R. C., Sheth R. K., Seo H.-J., Eisenstein D. J., Gray A., 2007, MNRAS, 378, 1196
Kwan C. C. Y., 2011, eJournal Spreadsheets in Education, 4, 3, 6
Ledoit O., Wolf M., 2003, Journal of Empirical Finance, 10, 603
Ludlow A. D., Porciani C., 2011, MNRAS, 413, 1961
Manera M., Gaztañaga E., 2011, MNRAS, 415, 383
Marín F., 2011, ApJ, 737, 97
Marín F. A. et al., 2013, MNRAS, 432, 2654
Matsubara T., 2011, PRD, 83, 083518
McBride C. K., Connolly A. J., Gardner J. P., Scranton R., Newman J. A., Scoccimarro R., Zehavi I., Schneider D. P., 2011, ApJ, 726, 13
McBride C. K., Connolly A. J., Gardner J. P., Scranton R., Scoccimarro R., Berlind A. A., Marn F., Schneider D. P., 2011, The Astrophysical Journal, 739, 85
McDonald P., 2006, PRD, 74, 103512
Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution
Nichol R. C. et al., 2006, MNRAS, 368, 1507
Nishimichi T. et al., 2007, Publications of Astronomical Society of Japan, 59, 1049
Pan J., Szapudi I., 2005, MNRAS, 362, 1363
Pollack J. E., Smith R. E., Porciani C., 2012, MNRAS, 2312
Porciani C., 2013, Publications of Astronomical Society of Australia, 30, 30
Roth N., Porciani C., 2011, MNRAS, 415, 829
Schäffer J., Strimmer K., 2005, Stat. Appl. Genet. Mol. Biol., 4, 1544
Schmidt F., Jeong D., Gaztañaga E., 2012, ArXiv e-prints
Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, 20
Sefusatti E., 2009, PRD, 80, 123002
Seljak U., Warren M. S., 2004, MNRAS, 355, 129
Seljak U., Zaldarriaga M., 1996, ApJ, 469, 437
Sigad Y., Branchini E., Dekel A., 2000, ApJ, 540, 62
Smith R. E., 2009, MNRAS, 1337
Smith R. E., Hernández-Monteagudo C., Seljak U., 2009, PRD, 80, 063528
Smith R. E., Reed D. S., Potter D., Marian L., Croccia M., Moore B., 2012, ArXiv e-prints
Smith R. E., Scoccimarro R., Sheth R. K., 2007, PRD, 75, 063512
Smith R. E., Sheth R. K., Scoccimarro R., 2008, PRD, 78, 023523
Springel V., 2005, MNRAS, 364, 1105
Szapudi I., Pan J., 2004, ApJ, 602, 26
Verde L. et al., 2002, MNRAS, 335, 432
Wang Y., Yang X., Mo H. J., van den Bosch F. C., Chu Y., 2004, MNRAS, 353, 287

© 0000 RAS, MNRAS 000, 000–000

ACKNOWLEDGEMENTS

JEP thanks Xun Shi and Andrés Balaguera-Antolínez for useful discussions. We thank V. Springel for making public GADGET-2 and for providing his B=FOF halo finder, and R. Scoccimarro for making public his 2LPT code. JEP and CP were partially supported by funding provided through the SFB-Transregio 33 The Dark Universe by the Deutsche Forschungsgemeinschaft. RES acknowledges support from ERC Advanced Grant 245797 GALFORMOD.

REFERENCES

Angulo R. E., White S. D. M., 2010, MNRAS, 405, 143
Baldauf T., Seljak U., Desjacques V., McDonald P., 2012, PRD, 86, 083540
Barnabeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep. , 367, 1
Catelan P., Lucchin F., Matarrese S., Porciani C., 1998, MNRAS, 297, 692
Catelan P., Porciani C., Kamionkowski M., 2000, MNRAS, 318, L39
Chan K. C., Scoccimarro R., Sheth R. K., 2012, PRD, 85, 083509
Chuen Chan K., Scoccimarro R., 2012, ArXiv e-prints
Crocce M., Pueblas S., Scoccimarro R., 2006, MNRAS, 373, 369
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Dekel A., Lahav O., 1999, ApJ, 520, 24
Francesco N., Sheth R. K., 2012, Journal of Cosmology and Astro-Particle Physics, 11, 16
Fry J. N., Gaztañaga E., 1993, ApJ, 413, 447
Gaztañaga E., Norberg P., Baugh C. M., Croton D. J., 2005, MNRAS, 364, 620
Guo H., Jing Y. P., 2009, ApJ, 702, 425
Guo H., Li C., Jing Y. P., Boerner G., 2013, ArXiv e-prints
Hamaus N., Seljak U., Desjacques V., Smith R. E., Baldauf T., 2010, PRD, 82, 043515

A new method to measure galaxy bias

added to second-order LEB in order to match all halo statistics.
(vi) Analysis of the cross-correlation coefficients defined for the two-point and three-point statistics reveal further subtle inconsistencies contained in the LEB truncated to second order, suggesting it is too simple a model to describe halo bias with high accuracy.

A final remark is in order. Our numerical study is based on simulations with a fixed background cosmology and focuses on retrieving the bias parameters when the cosmological parameters are perfectly known. However, this is not the case for actual galaxy surveys where bias and cosmology are generally estimated simultaneously. To transform our method into a resourceful tool for data analysis, we will need to explore how the shapes and amplitudes of the moments of the non-linear matter density field depend on the unknown cosmological parameters without having to run an exorbitant amount of N-body simulations (e.g. Angulo & White 2010) – a topic we shall explore in future work.
APPENDIX A: SPECTRAL RELATIONSHIPS

A1 Relationship between $\mathcal{P}_{(l,m)}$ and the $n$-point multispectra

We now derive the relation between the functions $\mathcal{P}_{(l,m)}$ and the multi-point matter spectra.

To begin, the functions $\mathcal{P}_{(l,m)}$ are defined:

$$\left\langle \Delta^{(l)}(k_1|R) \Delta^{(m)}(k_2|R) \right\rangle \equiv (2\pi)^3 \delta^D(k_1 + k_2) \mathcal{P}_{(l,m)}(k_1)$$

$$= \int \prod_{i=1}^{l} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} (2\pi)^3 \delta^D(k_1 - q_{1...l})(2\pi)^3 \delta^D(k_2 - p_{1...m}) \times$$

$$\times \left\langle \hat{\delta}(q_{1|R}) \ldots \hat{\delta}(q_l|R) \hat{\delta}(p_1|R) \ldots \hat{\delta}(p_m|R) \right\rangle$$

(A1)

The ensemble average of the $l + m$ density fields may be evaluated to give:

$$\left\langle \hat{\delta}(q_{1|R}) \ldots \hat{\delta}(q_l|R) \hat{\delta}(p_1|R) \ldots \hat{\delta}(p_m|R) \right\rangle \equiv (2\pi)^3 \delta^D(q_{1...l} + p_{1...m}) \mathcal{P}_{(l+m)}(q_1, \ldots, q_l, p_1, \ldots, p_m).$$

(A2)

On inserting the above definition into Eq. (A2) we find

$$\left\langle \Delta^{(l)}(k_1|R) \Delta^{(m)}(k_2|R) \right\rangle \equiv \int \prod_{i=1}^{l-1} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m-1} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} (2\pi)^3 \delta^D(k_1 - q_{1...l})(2\pi)^3 \delta^D(k_2 - p_{1...m}) \times$$

$$\times (2\pi)^3 \delta^D(q_{1...l} + p_{1...m}) \mathcal{P}_{(l+m)}(q_1, \ldots, q_l, p_1, \ldots, p_m)$$

(A3)

Integrating over the first two Dirac delta functions in yields:

$$\left\langle \Delta^{(l)}(k_1|R) \Delta^{(m)}(k_2|R) \right\rangle \equiv (2\pi)^3 \delta^D(k_1 + k_2) \int \prod_{i=1}^{l-1} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m-1} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \times$$

$$\times \mathcal{P}_{(l+m)}(q_1, \ldots, q_{l-1}, k_1 - q_{1...(l-1)}, p_1, \ldots, p_{m-1}, k_2 - p_{1...(m-1)})$$

(A4)

Hence,

$$\mathcal{P}_{(l,m)}(k_1) = \int \prod_{i=1}^{l-1} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m-1} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \mathcal{P}_{(l+m)}(q_1, \ldots, q_{l-1}, k_1 - q_{1...(l-1)}, p_1, \ldots, p_{m-1}, k_2 - p_{1...(m-1)})$$

(A5)

Lastly, we may change integration variables in the following way: $q_2 \rightarrow q_1$, $q_3 \rightarrow q_2 - q_1$, ..., upon which the above expression may be written as:

$$\mathcal{P}_{(l,m)}(k_1) = \int \prod_{i=1}^{l-1} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m-1} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \mathcal{P}_{(l+m)}(q_1, q_2 - q_1, \ldots, k_1 - q_{1...(l-1)}, p_1, p_2 - p_1, \ldots, k_2 - p_{1...(m-1)})$$

(A6)

Terms up to and including the quadraspectrum may be written:

$$\mathcal{P}_{(1,1)}(k_1) = \mathcal{P}_{(2)}(k_1, k_2) = \mathcal{P}(k_1)$$

(A7)

$$\mathcal{P}_{(2,1)}(k_1) = \int \frac{d^3q_1}{(2\pi)^3} \mathcal{P}_{(3)}(q_1, k_1 - q_1, k_2) = \int \frac{d^3q_1}{(2\pi)^3} \mathcal{P}(q_1, k_1 - q_1, k_2)$$

(A8)

$$\mathcal{P}_{(3,1)}(k_1) = \int \left\{ \frac{d^3q_1}{(2\pi)^3} \right\} \frac{d^3q_2}{(2\pi)^3} \mathcal{P}_{(4)}(q_1, q_2 - q_1, k_1 - q_2, k_2)$$

(A9)

$$\mathcal{P}_{(2,2)}(k_1) = \int \frac{d^3q_1}{(2\pi)^3} \frac{d^3p_1}{(2\pi)^3} \mathcal{P}_{(4)}(q_1, k_1 - q_1, p_1, k_2 - p_1)$$

(A10)

$$\mathcal{P}_{(4,1)}(k_1) = \int \frac{d^3q_1}{(2\pi)^3} \frac{d^3q_2}{(2\pi)^3} \frac{d^3q_3}{(2\pi)^3} \mathcal{P}_{(5)}(q_1, q_2 - q_1, q_3 - q_2, k_1 - q_3, k_2)$$

(A11)

$$\mathcal{P}_{(3,2)}(k_1) = \int \frac{d^3q_1}{(2\pi)^3} \frac{d^3q_2}{(2\pi)^3} \frac{d^3p_1}{(2\pi)^3} \mathcal{P}_{(5)}(q_1, q_2 - q_1, k_1 - q_2, p_1, k_2 - p_1)$$

(A12)

A2 Relationship between $\mathcal{B}_{(l,m,n)}$ and the $n$-point multispectra

In a similar fashion, we may now derive the relation between the functions $\mathcal{B}_{(l,m,n)}$ and the multi-point matter spectra.

To begin, the functions $\mathcal{B}_{(l,m,n)}$ are defined:

$$\left\langle \Delta^{(l)}(k_1|R) \Delta^{(m)}(k_2|R) \Delta^{(n)}(k_3|R) \right\rangle \equiv \int \prod_{i=1}^{l} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \prod_{k=1}^{n} \left\{ \frac{d^3s_k}{(2\pi)^3} \right\} (2\pi)^3 \delta^D(k_1 - q_{1...l})(2\pi)^3 \times$$

$$\times \delta^D(k_2 - p_{1...m})(2\pi)^3 \delta^D(k_3 - s_{1...n}) \left\langle \hat{\delta}(q_{1|R}) \ldots \hat{\delta}(q_l|R) \hat{\delta}(p_1|R) \ldots \hat{\delta}(p_m|R) \hat{\delta}(s_1|R) \ldots \hat{\delta}(s_n|R) \right\rangle$$

(A13)
The ensemble average of the $l + m + n$ density fields may be evaluated to give:

$$
\langle \delta(q_1|R) \ldots \delta(q_l|R) \delta(p_1|R) \ldots \delta(p_m|R) \delta(s_1|R) \ldots \delta(s_n|R) \rangle \equiv (2\pi)^3 \delta^D(q_{1 \ldots l} + p_{1 \ldots m} + s_{1 \ldots n}) \times 
\times \mathcal{P}_{(l+m+n)}(q_1, \ldots, q_l, p_1, \ldots, p_m, s_1, \ldots, s_n) .
$$ (A15)

On inserting the above definition into Eq. (A14) we find

$$
\langle \Delta(l) (k_1|R) \Delta^*(m) (k_2|R) \Delta^*(n) (k_3|R) \rangle \equiv \int \prod_{i=1}^l \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^m \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \prod_{k=1}^n \left\{ \frac{d^3s_k}{(2\pi)^3} \right\} (2\pi)^{12} \delta^D(k_1 - q_{1 \ldots l}) \delta^D(k_2 - p_{1 \ldots m}) \times 
\times \delta^D(k_3 - s_{1 \ldots n}) \mathcal{P}_{(l+m+n)}(q_1, \ldots, q_l, p_1, \ldots, p_m, s_1, \ldots, s_n) .
$$ (A16)

On integrating over the first three Dirac delta functions gives:

$$
\langle \Delta(l) (k_1|R) \Delta^*(m) (k_2|R) \Delta^*(n) (k_3|R) \rangle \equiv (2\pi)^3 \delta^D(k_1 + k_2 + k_3) \int \prod_{i=1}^l \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^m \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \prod_{k=1}^n \left\{ \frac{d^3s_k}{(2\pi)^3} \right\} 
\times \mathcal{P}_{(l+m+n)}(q_1, \ldots, k_1 - q_{1 \ldots l}, p_1, \ldots, k_2 - p_{1 \ldots m}, s_1, \ldots, k_3 - s_{1 \ldots n}) .
$$ (A17)

As for the power spectrum, we may change integration variables in the following way, $q_2 \rightarrow q_2 - q_1$, $q_3 \rightarrow q_3 - q_2$, $\ldots$. After which we find,

$$
\mathcal{B}_{(l,m,n)}(k_1, k_2, k_3) = \int \prod_{i=1}^{l-1} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m-1} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \prod_{k=1}^{n-1} \left\{ \frac{d^3s_k}{(2\pi)^3} \right\} 
\times \mathcal{P}_{(l+m+n)}(q_1, q_2 - q_1, \ldots, k_1 - q_{1 \ldots l}, p_1, \ldots, k_2 - p_{1 \ldots m}, s_1, \ldots, k_3 - s_{1 \ldots n}) .
$$ (A18)

### A3 Proof of the symmetry of $\mathcal{P}_{(l,m)}$

We now prove that the spectra $\mathcal{P}_{(l,m)}$ are symmetric in their indices $m$ and $l$:

$$
\mathcal{P}_{(l,m)} = \mathcal{P}_{(m,l)} .
$$ (A19)

Consider Eq. (A17), on relabelling the variables $p_i = q_{i+l-1}$, and writing $P_{(l+m)} = P_{(m+l)}$, we find,

$$
\mathcal{P}_{(l,m)}(k_1) = \int \left\{ \frac{d^3q_1}{(2\pi)^3} \right\} \ldots \left\{ \frac{d^3q_{l+m-2}}{(2\pi)^3} \right\} \mathcal{P}_{(m+l)}(q_1, \ldots, k_1 - q_{1 \ldots l}, q_{l+1}, \ldots, k_2 - q_{l+m-2}) .
$$ (A20)

On changing the integration variables to $p_i = -q_i$, we find

$$
\mathcal{P}_{(l,m)}(k_1) = \int \left\{ \frac{d^3p_1}{(2\pi)^3} \right\} \ldots \left\{ \frac{d^3p_{l+m-2}}{(2\pi)^3} \right\} \mathcal{P}_{(m+l)}(-p_1, \ldots, k_1 - p_{1 \ldots l}, -p_{l+1}, \ldots, k_2 + p_{l+m-2}) .
$$ (A21)

Parity invariance of the $n$-point correlation functions means that $\mathcal{P}_n(p_1, \ldots, p_n) = \mathcal{P}_n(-p_1, \ldots, -p_n)$ (for a proof see Smith, Sheth & Scoccimarro 2008). Under parity invariance, we find

$$
\mathcal{P}_{(l,m)}(k_1) = \int \left\{ \frac{d^3p_1}{(2\pi)^3} \right\} \ldots \left\{ \frac{d^3p_{l+m-2}}{(2\pi)^3} \right\} \mathcal{P}_{(m+l)}(-p_1, \ldots, -k_1 - p_{1 \ldots l}, -p_{l+1}, \ldots, -k_2 - p_{l+m-2}) .
$$ (A22)

We may switch $k_1 = -k_2$ and $k_2 = -k_1$,

$$
\mathcal{P}_{(l,m)}(k_1) = \int \left\{ \frac{d^3p_1}{(2\pi)^3} \right\} \ldots \left\{ \frac{d^3p_{l+m-2}}{(2\pi)^3} \right\} \mathcal{P}_{(m+l)}(p_1, \ldots, k_2 - p_{1 \ldots l}, p_{l+1}, \ldots, k_1 - p_{l+m-2}) .
$$ (A23)

Next we may rearrange the arguments of the $n$-point spectra, since it is totally symmetric under exchange symmetry: $\mathcal{P}_n(p_1, \ldots, p_n) = \mathcal{P}_n(p_{l+1}, \ldots, p_n, p_1, \ldots, p_l)$, whereupon

$$
\mathcal{P}_{(l,m)}(k_1) = \int \left\{ \frac{d^3p_1}{(2\pi)^3} \right\} \ldots \left\{ \frac{d^3p_{l+m-2}}{(2\pi)^3} \right\} \mathcal{P}_{(m+l)}(p_1, \ldots, k_2 - p_{1 \ldots l}, p_{l+1}, \ldots, k_1 - p_{l+m-2}) .
$$ (A24)

Finally on changing variables $p_{i+l-1} = q_i$, we obtain

$$
\mathcal{P}_{(l,m)}(k_1) = \int \prod_{i=1}^{l-1} \left\{ \frac{d^3q_i}{(2\pi)^3} \right\} \prod_{j=1}^{m-1} \left\{ \frac{d^3p_j}{(2\pi)^3} \right\} \mathcal{P}_{(m+l)}(q_1, \ldots, k_2 - q_{m-1}, p_1, \ldots, k_1 - p_{l-1}) .
$$ (A25)

and this completes the proof of Eq. (A19).
APPENDIX B: MODELLING $B_{(2,1,1)}^{(s)}$ WITH PERTURBATION THEORY

In order to better understand what drives the amplitude and functional form of the $B_{(1,m,n)}^{(s)}$ we have attempted to model the signal with standard perturbation theory techniques. Rather than modelling all of the spectra, we have focused on the lowest order non-trivial term $B_{(2,1,1)}$.

To begin, consider again Eq. (24), this came from:

$$\left\langle \Delta^{(2)}(k_1|R)\Delta^{(3)}(k_2|R)\Delta^{(1)}(k_3|R) \right\rangle = \int \frac{d^3q_1}{(2\pi)^3} \frac{d^3q_2}{(2\pi)^3} \frac{d^3q_3}{(2\pi)^3} \delta^D(k_1 - q_1 - q_2) \left\langle \delta(q_1|R)\delta(q_2|R)\delta(q_3|R) \right\rangle$$  \hspace{1cm} (B1)

The above ensemble averaged product can be broken into connected and disconnected terms. Hence:

$$\left\langle \delta(q_1|R) . . . \delta(q_4|R) \right\rangle = \left\langle \delta(q_1|R) . . . \delta(q_4|R) \right\rangle + \left\langle \delta(q_1|R)\delta(q_2|R)\right\rangle \left\langle \delta(q_3|R)\delta(q_4|R) \right\rangle + \left\langle \delta(q_1|R)\delta(q_3|R)\right\rangle \left\langle \delta(q_2|R)\delta(q_4|R) \right\rangle + \left\langle \delta(q_1|R)\delta(q_4|R)\right\rangle \left\langle \delta(q_2|R)\delta(q_3|R) \right\rangle$$  \hspace{1cm} (B2)

If we consider the disconnected terms, these may be written in terms of power spectra (c.f. Eq. (3)). We note that the first disconnected term is vanishing unless $k_1$ is the null vector. The remaining terms are:

$$\left\langle \delta(q_1|R) . . . \delta(q_4|R) \right\rangle = (2\pi)^3\delta^D(q_1 . . . q_4)T(q_1, q_2, q_3, q_4) + (2\pi)^3\mathcal{P}(q_1)\mathcal{P}(q_2) \left[ \delta^D(q_1 + q_3)\delta^D(q_2 + q_4) + \delta^D(q_1 + q_4)\delta^D(q_2 + q_3) \right]$$  \hspace{1cm} (B3)

On inserting the above expression into Eq. (B1) and computing the integral over $q_2$, we find:

$$B_{(2,1,1)}^{(s)} = \frac{2}{3} \left[ \mathcal{P}(k_2)\mathcal{P}(k_3) + 2\text{cyc} \right] + \frac{1}{3} \int \frac{d^3q_1}{(2\pi)^3} T(q_1, k_1 - q_1, k_2, k_3) + 2\text{cyc} \right] ,$$  \hspace{1cm} (B4)

where we have factored out the Dirac delta. The power spectrum and trispectrum may be evaluated using standard Eulerian perturbation theory (for a review see [Bernardeau et al.] [2003]). The important results that we will need are that

At one loop level we have

$$B_{(2,1,1)}^{(s)} = \frac{2}{3} \left[ \mathcal{P}(k_2)\mathcal{P}(k_3) + 2\text{cyc} \right] + \frac{1}{3} \int \frac{d^3q_1}{(2\pi)^3} T(q_1, k_1 - q_1, k_2, k_3) + 2\text{cyc} \right] .$$  \hspace{1cm} (B5)

The function $\mathcal{P}(0)$ denotes the smoothed linear matter power spectrum and $\mathcal{P}_{(1)}$ denotes the ‘1-loop’ correction. The 1-Loop term may be written as the

$$\mathcal{P}_{1\ell}(k) = \mathcal{P}_{22}(k) + \mathcal{P}_{13}(k)$$  \hspace{1cm} (B6)

where the loop integrals are ...
where \( P_i \equiv P_{\text{lin}}(q_i) \), \( P_{ij} \equiv P_{\text{lin}}(|q_i + q_j|) \) and \( q_{ij} \equiv q_i + q_j \). The calculation of the second-order coupling functions is straightforward. The result is

\[
F_2(s)(q_1, q_2) = \frac{5}{14} \left[ \alpha(q_1, q_2) + \alpha(q_2, q_1) \right] + \frac{2}{7} \beta(q_1, q_2) ; \quad (B12)
\]

\[
G_2(s)(q_1, q_2) = \frac{3}{14} \left[ \alpha(q_1, q_2) + \alpha(q_2, q_1) \right] + \frac{4}{7} \beta(q_1, q_2) ; \quad (B13)
\]

\[
F_3(s)(q_1, q_2, q_3) = \frac{7}{54} \left[ \alpha(q_1, q_{23}) F_2(s)(q_2, q_3) + \alpha(q_2, q_{13}) F_2(s)(q_1, q_3) + \alpha(q_1, q_{12}) F_2(s)(q_1, q_2) \right]
+ \frac{4}{54} \left[ \beta(q_{12}, q_{23}) G_2(s)(q_2, q_3) + \beta(q_{21}, q_{13}) G_2(s)(q_1, q_3) + \beta(q_{12}, q_{13}) G_2(s)(q_1, q_2) \right]
+ \frac{7}{54} \left[ \alpha(q_{12}, q_{13}) G_2(s)(q_1, q_2) + \alpha(q_{13}, q_{21}) G_2(s)(q_1, q_3) + \alpha(q_{23}, q_{11}) G_2(s)(q_2, q_3) \right] . \quad (B14)
\]

where we introduced the two fundamental mode coupling functions

\[
\alpha(q_1, q_2) = \frac{(q_1 + q_2) \cdot q_1}{q_1^2} ; \quad \beta(q_1, q_2) = \frac{|q_1 + q_2|^2(q_1 \cdot q_2)}{2q_1^2q_2^2} . \quad (B15)
\]

We then proceeded with integrating as in equation (24) taking care to cut the integral off at scales that exceed twice the filter radius.