Terms of use:

The halo 3-point correlation function: a methodological analysis.

A. Veropalumbo,[a,1] **A. Binetti, E. Branchini,**[c,d] **M. Moresco,**[e,f] **P. Monaco,**[g,h,i,l] **A. Oddo, A. G. Sánchez,**[k] **E. Sefusatti**[h,i,j]

[a]Dipartimento di Fisica, Università di Roma Tre, Via della Vasca Navale 84, I-00146 Roma, Italy

[b]Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, Via Celoria 16, I-20133 Milano, Italy

[c]INFN - Sezione di Roma Tre, Via della Vasca Navale 84, I-00146 Roma, Italy

[d]Dipartimento di Fisica, Università degli Studi di Genova, and INFN Sezione di Genova, Via Dodecaneso 33, I-16146, Genova, Italy

[e]Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Università di Bologna, Via Piero Gobetti 93/2, I-40129 Bologna, Italy

[f]INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, I-40129 Bologna, Italy

[g]Dipartimento di Fisica - Sezione di Astronomia, Università di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy

[h]IFPU, Institute for Fundamental Physics of the Universe, Via Beirut 2, I-34151 Trieste, Italy

[i]INAF-Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, I-34143 Trieste, Italy

[j]INFN, Sezione di Trieste, Via Valerio 2, I-34127 Trieste, Italy

[k]Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, D-85748 Garching, Germany

[l]SISSA - International School for Advanced Studies, Via Bonomea 265, I-34136 Trieste, Italy

[1]Corresponding author.

E-mail: alfonso.veropalumbo@uniroma3.it

Abstract.   Upcoming galaxy surveys will provide us with an unprecedented view of the Large-Scale Structure of the Universe and the realistic chance to extract valuable astrophysical and cosmological information from higher-order clustering statistics. This perspective poses new challenges, requiring both accurate and efficient estimators and a renewed assessment of possible systematic errors in the theoretical models and likelihood assumptions. This work investigates these issues in relation to the analysis of the 3-point correlation function (3PCF) in configuration space. We measure the 3PCF of 300 halo catalogs from the MINERVA simulations covering a total volume of $1000h^{-3}\mathrm{Gpc}^3$. Each 3PCF measurement includes *all* possible triangular configurations with sides between 20 and $130h^{-1}\,\mathrm{Mpc}$. In the first place, we test different estimates of the covariance matrix, a crucial aspect of the analysis. We compare the covariance computed numerically from the limited but accurate benchmark simulations set to the one obtained from 10000 approximate halo catalogs generated with the PINOCCHIO code. We demonstrate that the two numerically-estimated covariance matrices largely match, confirming the validity of approximate methods based on Lagrangian Perturbation Theory for generating mocks suitable for covariance estimation. We also compare the numerical covariance with a theoretical prediction in the Gaussian approximation. We find a good match between the two for separations above $40h^{-1}\,\mathrm{Mpc}$. We test the 3PCF tree-level model in Perturbation Theory. The model is adopted in a likelihood analysis aimed at the determination of bias parameters. We find that, for our sample of halos at redshift $z = 1$, the tree-level model performs well for separations $r \geq 40h^{-1}\,\mathrm{Mpc}$. Results obtained with this scale cut are robust against different choices of covariance matrix. We compare to the analogous analysis of the halo bispectrum already presented in a previous publication, finding a remarkable agreement between the two statistics. We notice that such comparison relies, to the best of our knowledge for the first time, on a robust and consistent covariance estimate and on the inclusion of essentially all measurable configurations in Fourier as in configuration space. We then test different assumptions to build the model defining a robust combination of hypotheses that lead to unbiased parameter estimates. Our results confirm the importance of 3PCF, supplying a solid recipe for its inclusion in likelihood analyses. Moreover, it opens the path for further improvements, especially in modelling, to extract information from non-linear regimes.

Contents

1 Introduction

It has long been realized that the analysis of the large-scale structure (LSS) of the Universe traced by the spatial distribution of galaxies provides a wealth of information complementary to other cosmological probes as weak gravitational lensing or the Cosmic Microwave Background. As a result, several galaxy redshift surveys have been carried out and planned to probe an increasingly larger fraction of the observable Universe, mapping the spatial distributions of tens of millions of galaxies (e.g. DESI [1, 2], the Euclid spectroscopic survey [3] and the Roman Telescope high-latitude survey [4]). So far, measurements of the of two-point statistics have been the main tool for the analysis of the LSS. Such statistics can be either the power spectrum in Fourier space, or the 2-point correlation function in configuration space. Yet, larger survey volumes will allow more precise measurements of higher-order clustering statistics, able to quantify the non-Gaussian nature of cosmological perturbations at low redshift. Such measurements can therefore help to disentangle different sources of non-Gaussianity, either of intrinsic interest if it's of primordial origin or possibly responsible for relevant degeneracies with cosmological parameters in the case of non-linear galaxy bias.

Indeed, the analysis of higher-order statistics in recent datasets is becoming routine, particularly in the case of the galaxy bispectrum, the 3-point correlation function in Fourier space

[5–10]. The bispectrum, in fact, presents some advantages with respect to its configuration-space counterpart. In the first place, the computational cost of measuring the bispectrum of the largest galaxy catalogs is moderate, allowing one to accurately compute its covariance matrix if an adequate number of simulated catalog is available [5, 11–16]. In the second place, theoretical models for the bispectrum, in some cases even up to one loop in Perturbation Theory, have been extensively validated with large, simulated data-sets [15, 17–23].

Similar developments have taken place for three-point correlation function (3PCF) in configuration space at a relatively slower pace due to the higher computational cost of its estimation in large galaxy catalogs and simulations ([24–28] and references therein). In this respect, the situation has changed dramatically with the advent of a new class of efficient 3PCF estimators [29–32] that reduced the computational scaling cost from N^3 to N^2, where N is the number of objects in the catalog. This allowed the analysis of large datasets and for the first detection of baryonic acoustic oscillation features in the 3PCF of galaxies [33] and clusters [34], and to perform a joint two- and three-point correlation analysis [35] paralleling those performed in Fourier space [5]. Such advances triggered a development on the theoretical side that pushed the 3PCF model beyond the analysis of the sole monopole, taking advantage of its redshift-space anisotropy [33, 36, 37].

The main advantage of the 3PCF over the bispectrum is its ability to deal with survey geometry. The footprint of spectroscopic surveys is characterized by irregular boundaries and holes in correspondence to regions that could not be observed. A complex survey geometry can be naturally dealt with in configuration space by using appropriate estimators as, in the 3PCF case, the Szapudi-Szalay one [38]. This is not the case in Fourier space, where the survey footprint induces mode coupling that needs to be accounted for by either deconvolving the measured correlation signal or by convolving the theory prediction with the window [31, 39, 40].

All these considerations justify the renewed interest in the 3PCF statistics and the urgency to carry out a dedicated study aimed at investigating how various methodological choices in the 3PCF analysis can affect the inference of model parameters. In this work we will assess the effect of these choices on the determination of galaxy bias parameters, one of the traditional motivations to study the 3PCF.

The most relevant one is the estimation of the covariance matrix. Numerical estimates of the 3PCF covariance matrix that rely on individual measurements from a set of mock catalogs are time-consuming, both because of mocks production and 3PCF measurements, despite the fast estimator mentioned above. A number of workarounds have been proposed to reduce the computational cost while keeping the accuracy at an acceptable level. Computational efficiency can be increased either by speeding up the creation of mocks by means of approximate halo catalogs (see [41] for a recent review) or by approximating the covariance matrix itself. In this work we will explore both possibilities. First of all, we will compare the 3PCF covariance matrix estimated from catalogs of dark matter halos obtained from the MINERVA set of full N-body simulations to the one estimated from a set of halo catalogs obtained with the approximated method implemented in the PINOCCHIO code, based on Lagrangian Perturbation Theory [42, 43]. Then we shall compare the performances of different 3PCF covariance estimators, considering the brute-force numerical estimate from mocks, the analytical Gaussian model of [29] and the shrinkage method [44], i.e. a hybrid combination of numerical and analytical estimates. To evaluate the goodness of each method we will perform the same likelihood analysis with different covariance matrices and compare the recovered best-fit estimates of the halo bias parameters. We will explore as well additional aspects of

the standard 3PCF analysis. One is the selection of the triangular configurations included in the analysis. A related one is the choice of the separation bin size, defining such triangles. In both cases we will evaluate the impact on the determination of constraints of the bias parameters. Finally, we will consider different relations among bias parameters, either theoretically-motivated or based on fits to numerical simulations, as a way to reduce the parameter space.

Our results will ultimately provide an extensive assessment of the validity of the tree-level PT model of the 3PCF under a variety of assumptions. The posterior constraints on bias parameters will be compared to the similar results obtained in [15] from the halo bispectrum measured from the same data set. This constitutes, to the best of our knowledge, the most direct and fair comparison between the two statistics presented so far in the literature, based, *on both sides*, on all measurable configurations, as a function of the smallest scale included, and taking advantage of a robust estimate of the covariance properties from the same set of approximate mocks.

The paper is structured as follows: in Sec. 2 we briefly describe the two halo catalogs used in our analysis and their parent simulations. In Section 3 we describe the estimators used to measure the 3PCF and its covariance matrix, then we compare the covariance matrices numerically estimated from the two different sets of mock catalogs obtained from different types of simulations. In Sec. 4 we present the model for the halo 3PCF adopted in this work and used, along with the measured 3PCF and its covariance matrix, to perform the likelihood analysis described in Sec. 5 and to estimate the halo bias parameters. The results of this analysis, aimed at assessing the impact of the various methodological options described above are presented in Sec. 6 and discussed in Sec. 7, in which we draw our main conclusions.

2   Data sets

All results presented in this work are based on two different sets of halo catalogs. The first is constituted by 298 catalogs of dark matter halos obtained from the MINERVA set of full N-body simulations. The second is a much larger set of 10,000 approximate halo catalogs produced with the PINOCCHIO code. Both sets, sharing the same simulation box and cosmological parameters, are described below.

2.1   MINERVA halo catalogs

The MINERVA simulations [45, 46] have been performed with the GADGET-III code [47] evolving 1000^3 dark matter particles of mass $m_p \simeq 2.67 \cdot 10^{10} \, h^{-1} \mathrm{M}_\odot$ in a cubic box of side $L = 1500 \, h^{-1} \, \mathrm{Mpc}$.

Dark matter halos were identified with a Friends-of-Friends algorithm of linking length equal to 0.2 times the mean inter-particle distance. Following [15], we only consider the $z = 1$ output and halos with masses larger than $1.12 \cdot 10^{13} \, h^{-1} \mathrm{M}_\odot$, corresponding to a mean number density $(\bar{n} = 2.13 \cdot 10^{-4} \, h^3 \mathrm{Mpc}^{-3})$, roughly matching that of H-α line emission galaxies that will be observed in wide surveys like Euclid [3].

2.2   PINOCCHIO halo catalogs

The second set of halo catalogs has been obtained with the PINOCCHIO code [42, 43]. Unlike the MINERVA set, here the dynamical evolution of the cosmic density field is approximated using third order Lagrangian Perturbation Theory. Dark matter particles are grouped in halos using a procedure based on the ellipsoidal collapse model. Since PINOCCHIO realisations are

computationally cheaper than N-body simulations, a larger number of mock halo catalogs can be generated. Here we consider the same $10,000$ catalogs used by [15] and adopt the same threshold mass, originally chosen to match the amplitude of the halo power spectrum measured in the MINERVA simulations at large scales. It has been shown that this matching minimizes the difference between the power spectrum and bispectrum variance estimated from the PINOCCHIO mocks and the one from the MINERVA simulations, both in real [15, 17] as in redshift-space [23]. We highlight that the first 300 PINOCCHIO mocks share the same random seed for the initial condition of the MINERVA simulations. This allows a direct comparison that is not affected by sample variance.

3 Measuring the 3PCF

3.1 3PCF estimator

To measure the 3PCF of the DM halos in the PINOCCHIO and MINERVA catalogs we use the Szapudi & Szalay estimator [38] as implemented in [29]. That estimator relies on the local Spherical Harmonics Decomposition (SHD) of the 3PCF

$$\zeta(r_{12}, r_{13}, \mu) = \sum_{\ell}^{\ell_{max}} \zeta_\ell(r_{12}, r_{13}) \, \mathcal{L}_\ell(\mu) \,, \tag{3.1}$$

where r_{12} and r_{13} are two sides of a triangle while μ is the cosine of the angle between them. The length of the third side is set by the vector relation $\mathbf{r}_{12} + \mathbf{r}_{13} + \mathbf{r}_{23} = 0$. $\mathcal{L}_\ell(\mu)$ is the ℓ-th Legendre polynomial. The SHD algorithm measures the coefficients of the expansion $\zeta_\ell(r_{12}, r_{13})$ up to a maximum multipole ℓ_{max}. Therefore eq. 3.1 provides an approximated estimate of the halo 3PCF. However, for a typical galaxy catalog, $\ell_{max} = 10$ is sufficient to converge to the correct solution [35].

For the multipoles measurements $\hat{\zeta}_\ell(r_{12}, r_{23})$, we consider linear bins of size $\Delta r = 10 \, h^{-1} \, \mathrm{Mpc}$ up to a maximum separation $r_{max} = 150 \, h^{-1} \, \mathrm{Mpc}$. The standard 3PCF is then obtained as

$$\hat{\zeta}(r_{12}, r_{13}; r_{23}) = \sum_{\ell}^{\ell_{max}} \hat{\zeta}_\ell(r_{12}, r_{13}) \, \widetilde{\mathcal{L}}_\ell(r_{12}, r_{13}, r_{23}) \,, \tag{3.2}$$

where $\widetilde{\mathcal{L}}_\ell(r_{12}, r_{13}, r_{23})$ is the Legendre polynomial of order ℓ weighted over the triangle

$$\widetilde{\mathcal{L}}_\ell(r_{12}, r_{13}, r_{23}) = \frac{\int_{r_{12}-\Delta_r/2}^{r_{12}+\Delta_r/2} \mathrm{d}p p^2 \int_{r_{13}-\Delta_r/2}^{r_{13}+\Delta_r/2} \mathrm{d}q q^2 \int_{r_{23}-\Delta_r/2}^{r_{23}+\Delta_r/2} \mathrm{d}s s^2 \Theta(p,q,s) \mathcal{L}_\ell(p,q,s)}{\int_{r_{12}-\Delta_r/2}^{r_{12}+\Delta_r/2} \mathrm{d}p p^2 \int_{r_{13}-\Delta_r/2}^{r_{13}+\Delta_r/2} \mathrm{d}q q^2 \int_{r_{23}-\Delta_r/2}^{r_{23}+\Delta_r/2} \mathrm{d}s s^2 \Theta(p,q,s)} \tag{3.3}$$

with

$$\Theta(r_{12}, r_{13}, r_{23}) = \begin{cases} 1 & \text{if } \left| \frac{r_{12}^2 + r_{13}^2 - r_{23}^2}{2 r_{12} r_{13}} \right| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

and it accounts for the binning of the third side [35]. The estimator cannot properly recover the 3PCF corresponding to isosceles configurations ($r_{12} = r_{13}$) with a finite and manageable number of multipoles. In fact, this intrinsic limitation extends to nearly isosceles triangles where $r_{13} - r_{12} \simeq \Delta r$. For this reason, we characterise a given set of measurements not only

in terms of the minimum separation r_{min}, but also by a minimum value η_{min} for the relative difference

$$\eta \equiv \frac{r_{13} - r_{12}}{\Delta r} \, . \qquad (3.5)$$

In other terms, fixed Δr, r_{min} and r_{max}, setting $\eta_{min} = 0$ would correspond to measure all available triangles with sides within the given range in steps of Δr. $\eta_{min} = 1$ would exclude configurations with $r_{13} = r_{12}$, while higher values would further reduce the triangle set. As we will see our ability to properly model the 3PCF measured with the SHD estimator depends significantly on this parameter. In Sec. 5 we perform several tests to assess the sensitivity of the 3PCF analysis to the choice of both r_{min} and η_{min}.

Input to the estimator are a catalog of halos and a catalog of random objects, that we created, which are distributed in the same volume but lacking clustering properties. We've performed the measurements with the random splitting method. With this technique, we measure the 3PCF with many small random samples and then take their average. We impose the data-random ratio after splitting $N_R = 1$. Since the samples have different numbers of objects, we change the random accordingly. In the MINERVA 3PCF measurements the number of random objects is 50 times larger than that of the halos. In the PINOCCHIO case, for computational constraints, we use 10 times as many objects. For this reason, even for the low density randoms, we do not expect a biased 3PCF estimation but only an additional contribution to its variance. The effect on the total error budget, however, is negligible. We notice that it's not straightforward to change the estimator to exploit the periodicity of the box, thus avoiding the use of a random sample. First, the Szapudi-Szalay estimator directly measure the connected part of the 3PCF, internally subtracting the disconnected part. Second, since we use [29] for the triplet counts, we measure the multipoles of the 3PCF. We've tried to estimate the full 3PCF as DDD/N_D^3 normalized by the analytic triplet counts, with DDD the data triplets and N_D is the total number of triplets. We then need to subtract the disconnected part, that is $1 + \xi(r_{12}) + \xi(r_{13}) + \xi(r_{23})$. We've verified that this procedure is not sufficiently precise for our setting.

3.2 Covariance matrix

The numerical covariance matrix is estimated from the 3PCFs measured in each mock catalog as follows:

$$\widehat{C}_{i,j} = \frac{1}{N_{mocks} - 1} \sum_{n=1}^{N_{mocks}} \left(\hat{\zeta}_i^n - \overline{\zeta}_i \right) \left(\hat{\zeta}_j^n - \overline{\zeta}_j \right) \qquad (3.6)$$

where the indexes i and j identify any two triangles in the sample with sides $\{r_{12}, r_{13}, r_{23}\}$ and $\{r'_{12}, r'_{13}, r'_{23}\}$. $\hat{\zeta}^n$ indicates the 3PCF measured in the n-th mock and $\overline{\zeta}$ the average among the mocks

$$\overline{\zeta}_i = \frac{1}{N_{mocks}} \sum_{n=1}^{N_{mocks}} \hat{\zeta}_i^n. \qquad (3.7)$$

Then, to compare the MINERVA and PINOCCHIO estimates we consider as well the correlation matrix, defined as

$$c_{ij} \equiv \frac{\widehat{C}_{ij}}{\sqrt{\sigma_i \sigma_j}} \qquad (3.8)$$

where $\widehat{C}_{i,j}$ is the covariance matrix element and $\sigma_i \equiv \sqrt{\widehat{C}_{i,i}}$ represents the dispersion of the measurements for triangle i.
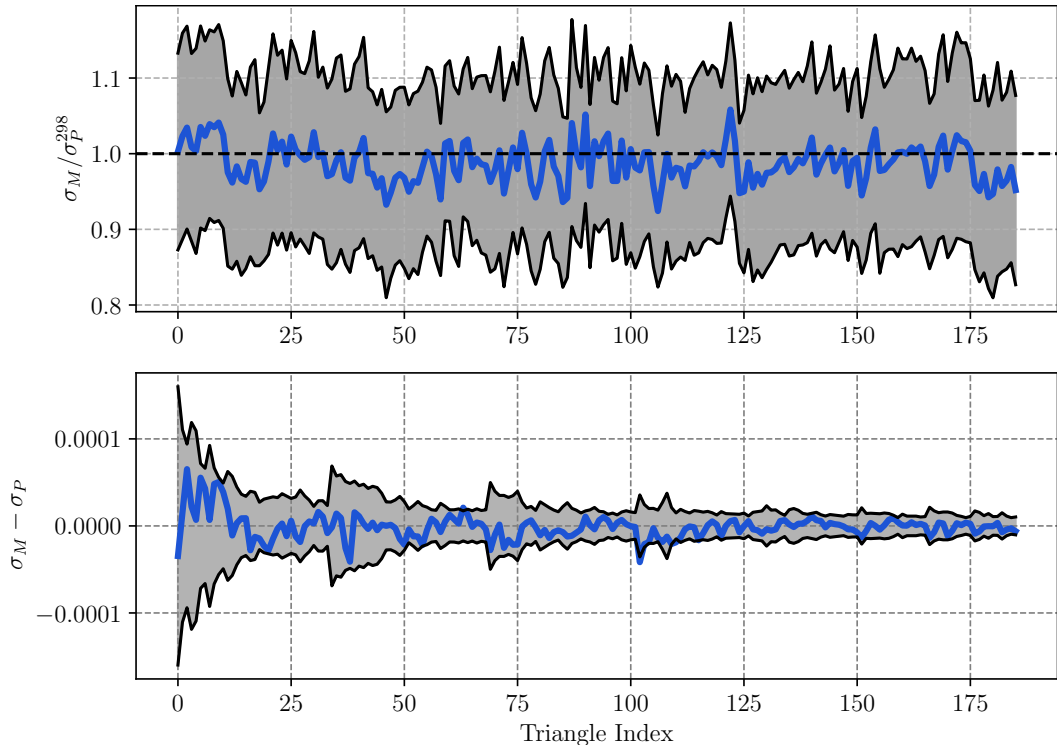
Figure 1. *Top panel.* Blue curve: ratio between the 3PCF *rms* scatter measured in the MINERVA, σ_M, and measured in the first 298 PINOCCHIO mock catalogs, σ_P^{298}, shown as a function of the triangle type, identified by an index. Larger indexes identify larger triangles. Grey band: $2-\sigma$ scatter of σ_M/σ_P^{298}, obtained propagating the variance over single *rms*. *Bottom panel.* Blue curve: difference between the 3PCF *rms* scatter measured the MINERVA, σ_M, and measured in the PINOCCHIO mock catalogs, σ_P, shown as a function of the triangle type, identified by an index. Grey band: the reference $2-\sigma$ scatter was obtained from subsets of 298 realisations extracted from the 10,000 PINOCCHIO mocks. See text for details.

3.3 3PCF covariance: MINERVA vs PINOCCHIO

In this section we provide a first comparison between the numerical covariance estimated from the MINERVA N-body simulations and the one estimated from the PINOCCHIO approximate mocks.

The top panel of Fig. 1 shows the ratio between the MINERVA variance and the variance from the first 298 PINOCCHIO mocks sharing the same initial conditions. Each measurement corresponds to all triangles with $r_{min} = 20h^{-1}\,\text{Mpc}, r_{max} = 130h^{-1}\,\text{Mpc}, \eta = 1$ ordered assuming $r_{12} < r_{13} \le r_{23}$. The shaded area shows the 2-σ scatter obtained from subsets of 298 realisations extracted from the 10,000 PINOCCHIO mocks. The bottom panel shows instead the difference between the MINERVA variance and the PINOCCHIO variance this time estimated from the full 10,000 mocks set. We notice that the difference between the two estimates is of the order of 5% and it is well within the estimated scatter for the variance based on 298 measurements.

In Fig. 2 we compare instead the two correlation matrices $c_{i,j}$, defined in eq. (3.8), with the MINERVA one in the upper-left triangle and the PINOCCHIO one from the 10,000 mocks in the lower-right triangle. Small differences in the most off-diagonal elements are barely visible.

A more quantitative comparison is shown in Fig. 3 in terms of cuts across the correlation
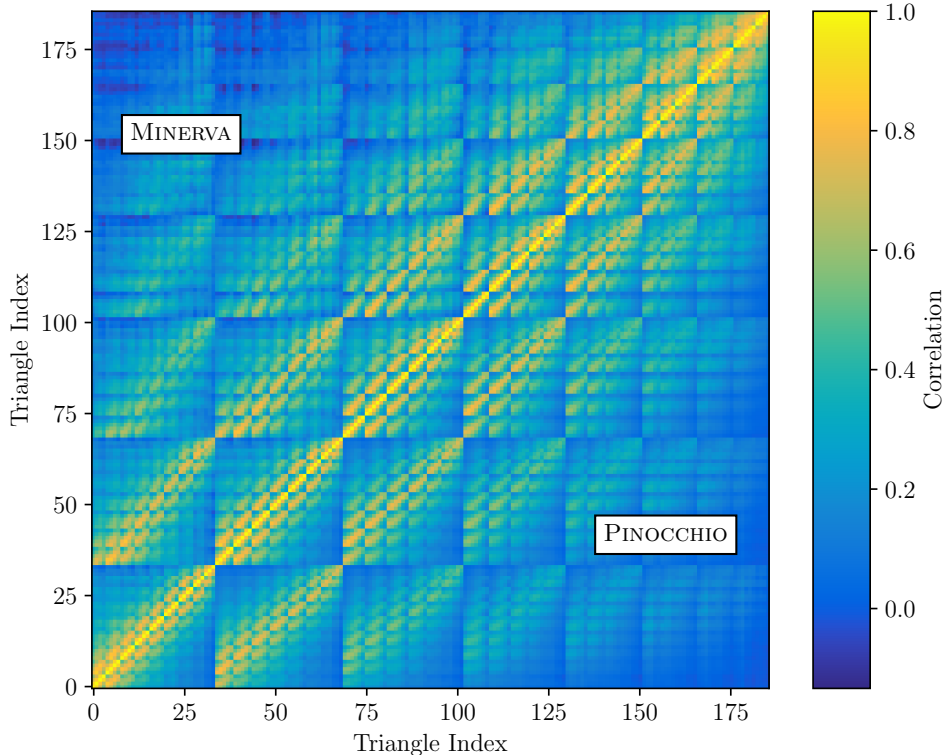
Figure 2. 3PCF correlation matrix measured from the MINERVA mock catalogs (upper triangular matrix) and from the 10000 PINOCCHIO mocks (lower triangular matrix). Indexes identify different triangle configurations. Larger indexes indicate larger scales. The correlation amplitude is color-coded as indicated in the color bar.

matrix: the black dots show a row of the MINERVA correlation matrix, with the error bars estimated from the scatter of subsets of 298 realisations extracted from the 10,000 PINOCCHIO mocks, while the yellow curve connects the elements from the same row in the PINOCCHIO matrix. The 3PCF covariance matrix estimated from the 10000 PINOCCHIO mocks and the one from the 298 MINERVA simulations agree with each other within the uncertainty estimated for the latter. We also quantify the distribution of residuals of the elements of the two covariance matrices, in units of their standard deviation $\equiv \left(C_{i,j}^M - C_{i,j}^P \right) / \sigma_{C_{i,j}}$. The residuals are normally distributed, with mean ~ 0.2 and standard deviation ~ 0.95, close to a standard normal distribution. This deviation is indeed small with respect to the variance and can be safely ignored. We shall consider the PINOCCHIO covariance as a reference and will use it to perform all likelihood analyses presented in what follows.

3.4 3PCF covariance: MINERVA vs theory

We now consider an analytical model for the covariance matrix, in the Gaussian approximation, as described in [29]. An analytic expression has the obvious advantage of being noiseless and flexible to changes in the fiducial cosmology. Most importantly it does not require large sets of mock catalogs and the numerically expensive measurements of the 3PCF. On the other hand, the simple analytic model we consider here makes assumption, as the Gaussianity of the halo distribution, that could lead to potential systematic errors and does not take into account the specific geometry of the sample (but see [48] for possible improvements).
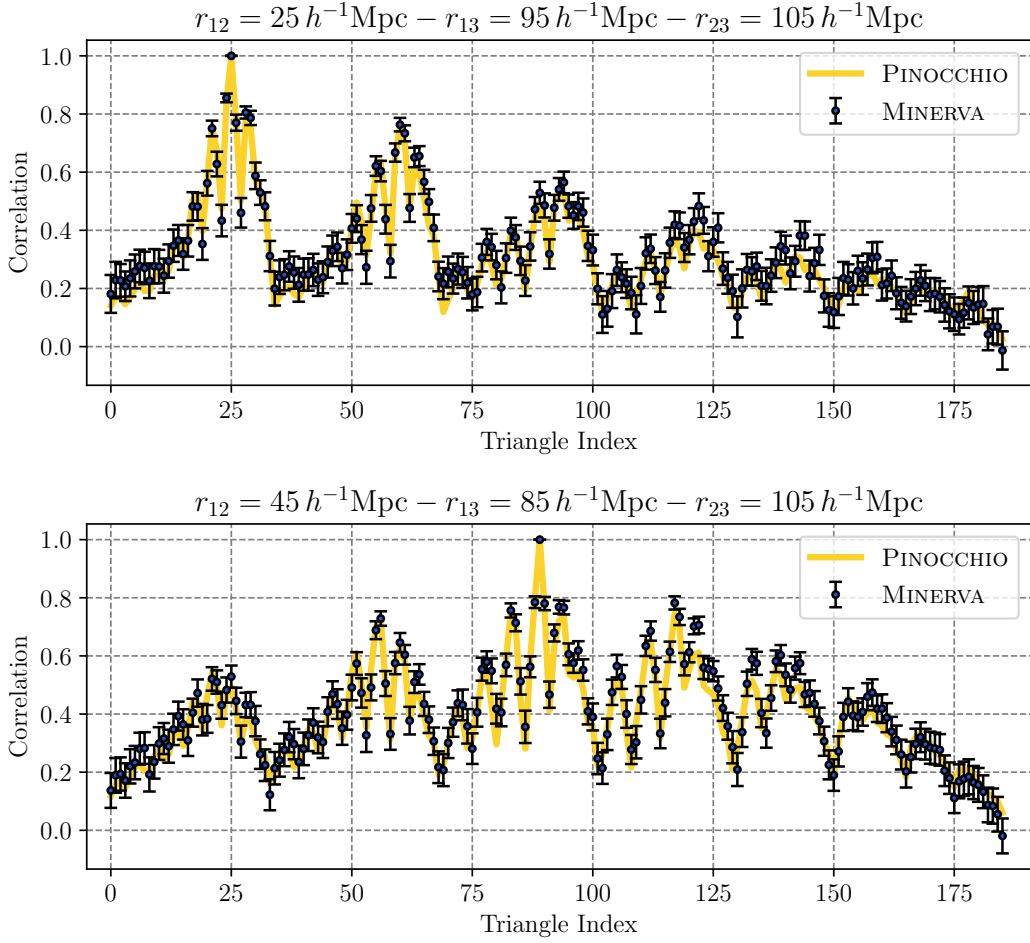
Figure 3. *Top panel*: correlation amplitude of the elements along a cut through correlation matrix in correspondence of a triangle with sides $25\,h^{-1}\,\mathrm{Mpc}, 95\,h^{-1}\,\mathrm{Mpc}, 105\,h^{-1}\,\mathrm{Mpc}$. Black dots: MINERVA estimates. Yellow curve: PINOCCHIO estimates. Error bars: scatter off subsets of 298 realisations extracted from the 10,000 PINOCCHIO mocks. *Bottom panel*: same quantity estimated for a larger triangle with sides $45\,h^{-1}\,\mathrm{Mpc}, 85\,h^{-1}\,\mathrm{Mpc}, 105\,h^{-1}\,\mathrm{Mpc}$.
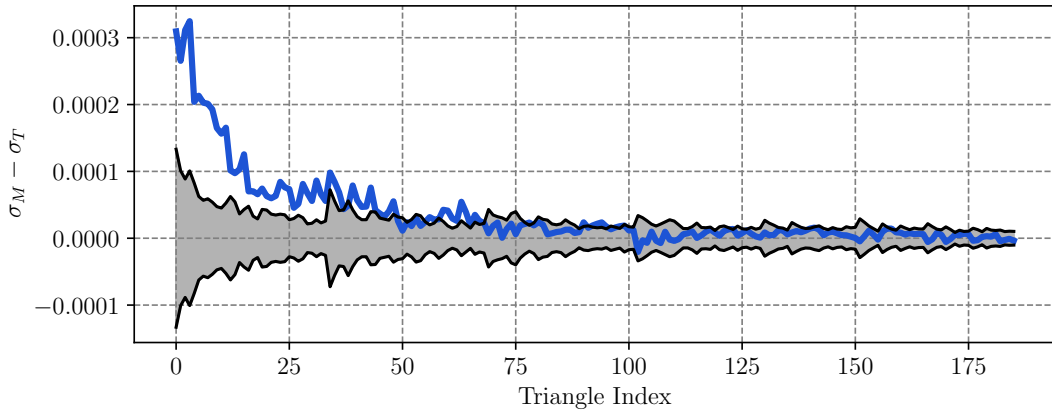
To compute the analytic covariance matrix we need to assume a fiducial, non-linear

power spectrum, the covered volume and a shot noise contribution to the power spectrum. The former is provided by the best-fit model to the MINERVA power spectrum measurements[1]. The latter is set equal to the inverse of the mean halo number density in the mocks, i.e. $1./\bar{n} = 4500\,h^{-3}\mathrm{Mpc}^3$.

To check if the analytical, Gaussian errors are consistent with those obtained from the mock catalogs, we show in fig. 3.4 the difference between the two sets of errors (blue curve), compared with the 2-σ error measured in the MINERVA mocks (grey band) obtained as in Fig. 1. We notice that the theory errors are systematically smaller than those measured from the mocks, with the mismatch decreasing with the size of the triangles. This is expected since the adequacy of the Gaussian hypothesis increases with the scale, while it is expected to break down on small scales due to the non-linear evolution inducing non-Gaussian contributions to the covariance. At larger scales the deviation is smaller, of the order of few percent. In order to reabsorb this small difference, one can exploit a small set of measured 3PCFs to tune the theoretical covariance parameters [29, 49].

4 The halo 3PCF Model

Since one of the goals of this work is to compare the results of the 3PCF analysis to the one of the bispectrum analyses, we adopt the model of [15] and obtain the 3PCF model via inverse-FFT as for example in [24, 25, 50]. The model corresponds to the simple tree-level expression in Perturbation Theory including the contribution from gravitational instability plus the one from non-linear, local and tidal bias [51–53]. Here we follow [50] to obtain the *pre-cyclic* 3PCF part of the prediction as

$$
\begin{aligned}
\zeta_{pc}(r_{12}, r_{13}, \mu_{23}) =& b_1^3 \left(\frac{34}{21} + \frac{b_2}{b_1} - \frac{4}{3}\frac{\gamma_2}{b_1} \right) \xi_0(r_{12})\xi_0(r_{13}) + \\
& - b_1^3 \left[\xi_1^+(r_{12})\xi_1^-(r_{13}) + \xi_1^-(r_{12})\xi_1^+(r_{13}) \right] \mathcal{L}_1(\mu_{23}) \\
& + b_1^3 \left(\frac{8}{21} + \frac{4}{3}\frac{\gamma_2}{b_1} \right) \xi_2(r_{12})\xi_2(r_{13})\mathcal{L}_2(\mu_{23}) ,
\end{aligned}
\tag{4.1}
$$

where the triangle sides r_{12}, r_{13}, r_{23} and the cosine angle μ_{23} are related through $r_{23}^2 = r_{12}^2 + r_{13}^2 - 2\mu_{23}r_{12}r_{13}$ and j_ℓ are the spherical Bessel functions.

The two-point correlation function moments in eq. (4.1)

$$
\begin{aligned}
\xi_\ell(r) &= \frac{1}{2\pi^2} \int_0^\infty \mathrm{d}k k^2 P(k) j_\ell(kr) , \\
\xi_\ell^{[\pm]}(r) &= \frac{1}{2\pi^2} \int_0^\infty \mathrm{d}k k^2 k^{\pm 1} P(k) j_\ell(kr)
\end{aligned}
\tag{4.2}
$$

depend on the linear matter power spectrum $P(k)$, that we model using CAMB [54].

The full galaxy 3PCF model is then obtained by summing permutations of the ζ_{pc} function:

$$
\zeta_h(r_{12}, r_{13}, r_{23}) = \zeta_{pc}(r_{12}, r_{13}, \mu_{23}) + \zeta_{pc}(r_{13}, r_{23}, \mu_{12}) + \zeta_{pc}(r_{23}, r_{12}, \mu_{13}).
\tag{4.3}
$$

---

[1]The best-fit power spectrum has been derived and kindly provided by Kevin Pardede using the PT model and assumptions of [17].

To speed up the sampling of the parameter space in the likelihood analysis of Sec. 5 we first numerically evaluate and tabulate the integrals in eq. (4.2) in the reference ΛCDM cosmology and then maximize the likelihood with respect to the bias parameters.

This model, however, cannot be compared directly to the quantities measured by our estimator based on the Spherical Harmonics Decomposition. We should therefore, in the first place, expand it in multipoles. In addition, we need to account for the finite size of separation bins and we do that at the multipoles level, computing

$$\widetilde{\zeta}_\ell(r_{12}, r_{13}) = \frac{2l+1}{2} \int_{-1}^{1} \mathrm{d}\mu \mathcal{L}_\ell(\mu) \int_{r_{12}-\Delta r/2}^{r_{12}+\Delta r/2} \mathrm{d}q\, q^2 \int_{r_{13}-\Delta r/2}^{r_{13}+\Delta r/2} \mathrm{d}p\, p^2\, \zeta(p, q, \mu), \qquad (4.4)$$

where r_{12}, r_{13} represent the center of the bins, Δr is the bin size and $\zeta(p, q, \mu)$ is the model, eq. (4.3). In the end, the binned 3PCF model that we compare with the data is

$$\widetilde{\zeta}(r_{12}, r_{13}, r_{23}) = \sum_{l=0}^{\ell_{max}} \widetilde{\zeta}_\ell(r_{12}, r_{13})\, \widetilde{\mathcal{L}}_\ell(r_{12}, r_{13}, r_{23}). \qquad (4.5)$$

We notice that the comparison of the model to an accurate estimation of the full 3PCF takes advantage of a clear dependence on the physical scales represented by the three sides. This allows in turn to a proper definition of the minimal separation r_{min} at which the model fails. The accuracy of the estimation of the full 3PCF $\zeta(r_{12}, r_{13}, r_{23})$ from the equation above, however, is limited by the finite value of $\ell_{max} = 10$, required if we want to preserve the efficiency of the implementation of [29]. In section 6 we will explore in detail the subset of triangular configurations affected by this approximation and therefore not properly described by the tree-level expression.

5 Parameter inference

We validate the galaxy 3PCF model, in terms of the posterior distribution on the three bias parameters, testing various assumptions on the likelihood analysis.

For each catalog, identified by the α index, we define the χ^2 function

$$\chi_\alpha^2(\mu, \mathbb{C}) = (\mathbf{d}_\alpha - \boldsymbol{\mu})^T \mathbb{C}^{-1} (\mathbf{d}_\alpha - \boldsymbol{\mu}), \qquad (5.1)$$

where μ and $\mathbf{d}_\alpha$ represent, respectively, the model and data vector while $\mathbb{C}$ is the covariance matrix with elements $C_{i,j}$. In the case of the (noiseless) analytic covariance matrix we adopt, for the single realization, the Gaussian likelihood function

$$\mathcal{L}_\alpha \propto \exp\left(-\frac{\chi_\alpha^2}{2}\right). \qquad (5.2)$$

We will also consider covariance matrices, $\widehat{\mathbb{C}}$, estimated numerically from a limited number of N_m halo catalogs, whose statistical noise can bias the determination of the precision matrix $\mathbb{C}^{-1}$. We then include the correction suggested in [55] and given by

$$\mathbb{C}^{-1} = \frac{N_m - n_d - 2}{N_m - 1} \widehat{\mathbb{C}}^{-1} \qquad (5.3)$$

where N_m is the number of independent measurements used to estimate $\widehat{\mathbb{C}}$ while n_d is the size of the data vector. The correction inflates the errors in order to encompass any possible

Name	Binning	Covariance	b_1	b_2	γ_2	Sec.
Reference	Integrated	Pinocchio	$\mathcal{U}(0.5, 5)$	$\mathcal{U}(-10, 10)$	$\mathcal{U}(-10, 10)$	6.1
Benchmark	Integrated	Minerva	$\mathcal{U}(0.5, 5)$	$\mathcal{U}(-10, 10)$	$\mathcal{U}(-10, 10)$	6.2
Gaussian	Integrated	Theoretical	$\mathcal{U}(0.5, 5)$	$\mathcal{U}(-10, 10)$	$\mathcal{U}(-10, 10)$	6.3
Shrinkage	Integrated	Shrinkage	$\mathcal{U}(0.5, 5)$	$\mathcal{U}(-10, 10)$	$\mathcal{U}(-10, 10)$	6.4
Bin-center	Center	Pinocchio	$\mathcal{U}(0.5, 5)$	$\mathcal{U}(-10, 10)$	$\mathcal{U}(-10, 10)$	6.5
$\gamma_2(b_1)$	Integrated	Pinocchio	$\mathcal{U}(0.5, 5)$	$\mathcal{U}(-10, 10)$	Eq. 6.7	6.6
$b_2(b_1)$	Integrated	Pinocchio	$\mathcal{U}(0.5, 5)$	Eq. 6.8	$\mathcal{U}(-10, 10)$	6.6

Table 1. Summary of test case and their main characteristics.

systematics on $\mathbb{C}^{-1}$. In the case of the covariance estimated from the full set of 10,000 Pinocchio mocks, given a data vector comprising at most 184 triangular configurations, the correction is rather small and negligible. This is not the case, naturally, for the covariance estimated from the more limited set of Minerva simulations, where the factor can be as low as 0.35. We considered as well the prescription of [56] finding marginal differences in our applications.

The total likelihood adopted for each evaluation is obtained as the product of all individual likelihood, or

$$\log \mathcal{L}_{tot} = \sum_{\alpha=1}^{N_m} \log \mathcal{L}_\alpha \, . \tag{5.4}$$

Finally, we use the correction factor in [57] to fully take into account the propagation of covariance matrix uncertainty on the parameter posterior distributions. The factor is

$$m_1 = \frac{1 + B(n_d - n_p)}{1 + A + B(n_p + 1)}; \tag{5.5}$$

with n_p the size of the model parameter vector, and

$$A = \frac{2}{(N_m - n_d - 1)(N_m - n_d - 4)}, \tag{5.6}$$

$$B = \frac{N_m - n_d - 2}{(N_m - n_d - 1)(N_m - n_d - 4)} \, . \tag{5.7}$$

We adopt flat, uninformative priors, specified in Table 1. To sample the likelihood function we use a Monte Carlo Markov Chain (MCMC) approach as implemented in the EMCEE software [58].

6 Results

In this section we will consider the effect of different triangle selections, covariance matrices and binning schemes, as well as the inclusion of relations among bias parameters to reduce the parameter space. Table 1 lists the characteristics of the various tests and summarizes the methodological choices adopted in each of them. All runs share the same Minerva halo

$$r_{min}\,[h^{-1}\,\mathrm{Mpc}]$$

		10	20	30	40	50	60	70
	1	216	186	152	117	84	56	35
η_{min}	2	170	143	113	83	56	35	20
	3	131	107	81	56	35	20	10

Table 2. Number of triangles as a function of η_{min} and r_{min}. r_{max} is fixed to $130\,h^{-1}\,\mathrm{Mpc}$.

catalogs at $z = 1$ where the 3PCF is measured with the SHD estimator using $\ell_{max} = 10$, $\Delta r = 10 h^{-1}\,\mathrm{Mpc}$ and $\eta_{min} = 1$. The Gaussian likelihood is maximized with respect to the three halo bias parameters b_1, b_2 and γ_2. All other parameters are kept fixed.

6.1 Sensitivity to triangle size and shape

To assess the sensitivity of the 3PCF analysis to the size and shapes of the triangles we performed different likelihood analyses in which we fix η_{min}, hence selecting the triangle types, and progressively increase the value of r_{min}. The number of triangles that remain after a specific selection is shown in Tab. 2.

The results are shown in Fig. 4. In each panel we plot the best fit bias parameters (b_1, b_2, γ_2, from top to bottom) as a function of r_{min}. Shaded areas represent 1-σ uncertainty. We also plot, for reference, the best fit values (and uncertainties) obtained by [15] from the bispectrum analysis of the same dataset with Fourier space triangles smaller than $k_{max} = 0.082\,h\,\mathrm{Mpc}^{-1}$ (dashed black lines and grey shaded areas). Each panel shows the results obtained with $\eta_{min} = 1, 2$ and 3 (red, green and blue curves, respectively), corresponding to excluding those triangles with no or a small difference between the sides r_{12} and r_{13}, see eq. (3.5).

Let us focus first on the cases with $\eta_{min} = 2$ and 3. In these cases, both parameters b_2 and γ_2, for $r_{min} \geq 20\,h^{-1}\,\mathrm{Mpc}$, converge to a constant value consistent with the one obtained in the bispectrum analysis. Only the linear bias parameters b_1 exhibits a scale dependence, particularly significant for r_{min} up to $\sim 40 h^{-1}\,\mathrm{Mpc}$ and for $\eta_{min} = 2$.

The case of $\eta_{min} = 1$ is different. In this case γ_2 and b_1 show some dependence on r_{min}; the former converges to the correct value, but we should notice how the behavior of b_2 shows no clear trend and it is inconsistent with the bispectrum results for all values of r_{min}. This can be explained recalling that the case $\eta_{min} = 1$ includes the *quasi-isosceles* configurations, that is those with $r_{12} \simeq r_{13}$, with the third side r_{23} approaching zero. When this happens, the 3PCF tree-level model that we adopted in our analysis, specific to the fast estimator described in section 3.1, fails [29, 35]. In fact, both the estimator and the model are defined by a maximum value of $\ell_{max} = 10$ in the Legendre expansion, eq. (3.1), and probably a larger value of ℓ_{max} would have reduced the dependence on η_{min}. In the rest of this work we adopt $\eta_{min} = 3$ for all our results.

To inspect the degeneracy between the bias parameters and how this depends on r_{min}, we plot their 1D and 2D marginalized distributions in the triangle plot of Fig. 5. We compare the three cases of $r_{min} = 20, 30$ and $40 h^{-1}\,\mathrm{Mpc}$ (respectively in blue green and red) with the result of the bispectrum analysis (grey) from [15], where a scale cut of $k_{max} = 0.09\,h\,\mathrm{Mpc}^{-1}$ is assumed.

For $r_{min} \leq 30 h^{-1}\,\mathrm{Mpc}$ the constraints from the 3PCF analysis are slightly tighter and less degenerate than in the bispectrum case, with the b_1 value recovered from the 3PCF 1-σ
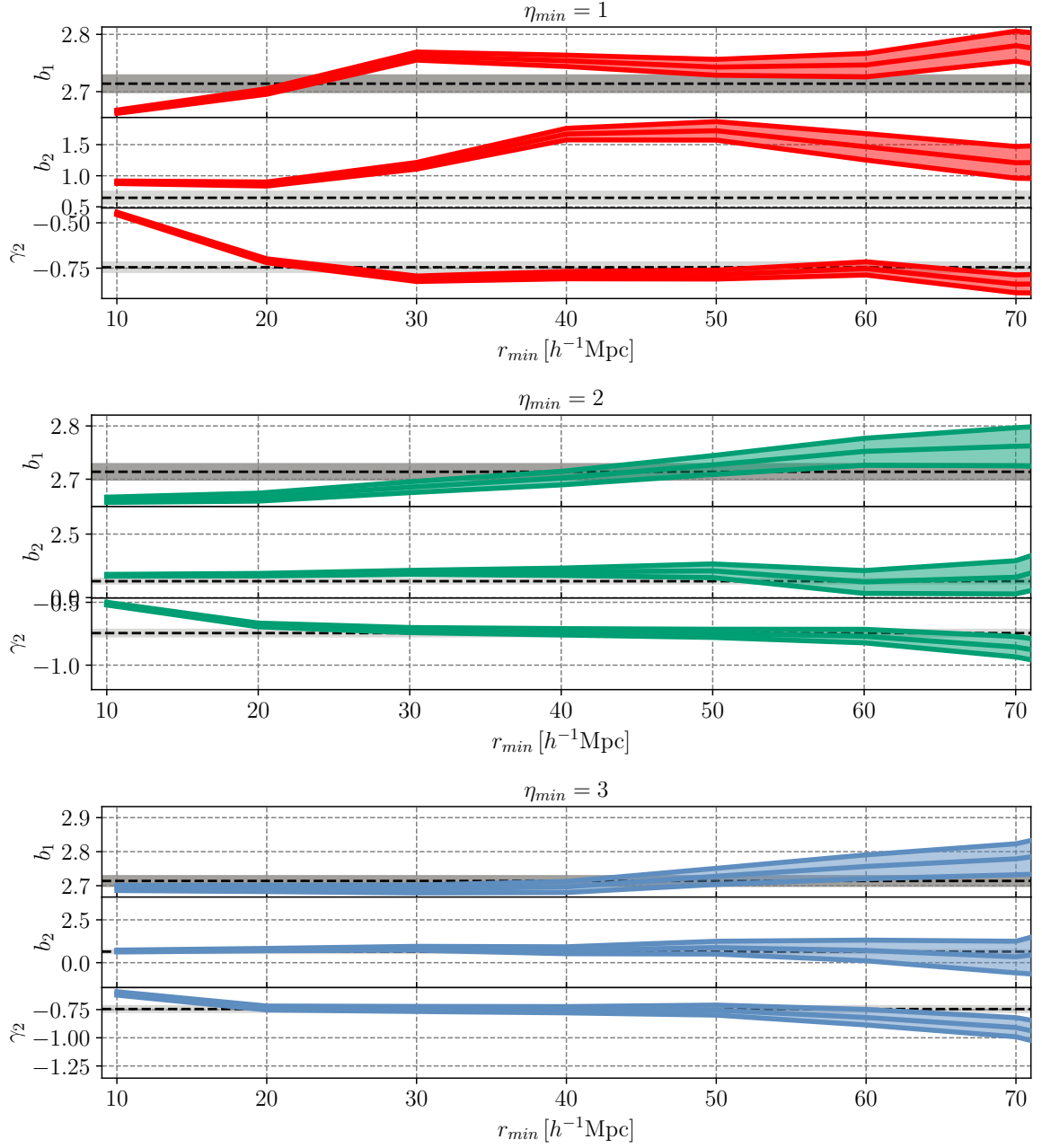
Figure 4. Inferred bias parameters as a function of the minimum triangle side r_{min} (thick red curve) with their 1σ uncertainty (shaded area) for the Reference case. Dashed black lines with grey shaded region represent the bias values and their 1σ errors estimated from the bispectrum analysis of [15]. *Top, central and bottom panels*: all results are for the reference case using triangles with $\eta \geq 1$, 2 and 3, respectively.

lower than the one from the bispectrum. On the other hand, the error on b_2 inferred from the bispectrum is smaller. The posteriors width is a non-trivial combination of two aspects: the ability of the different probes to break degeneracies among parameters and the information content as a function of the scale. It is indeed evident, in Fig. 5, that the two probes are
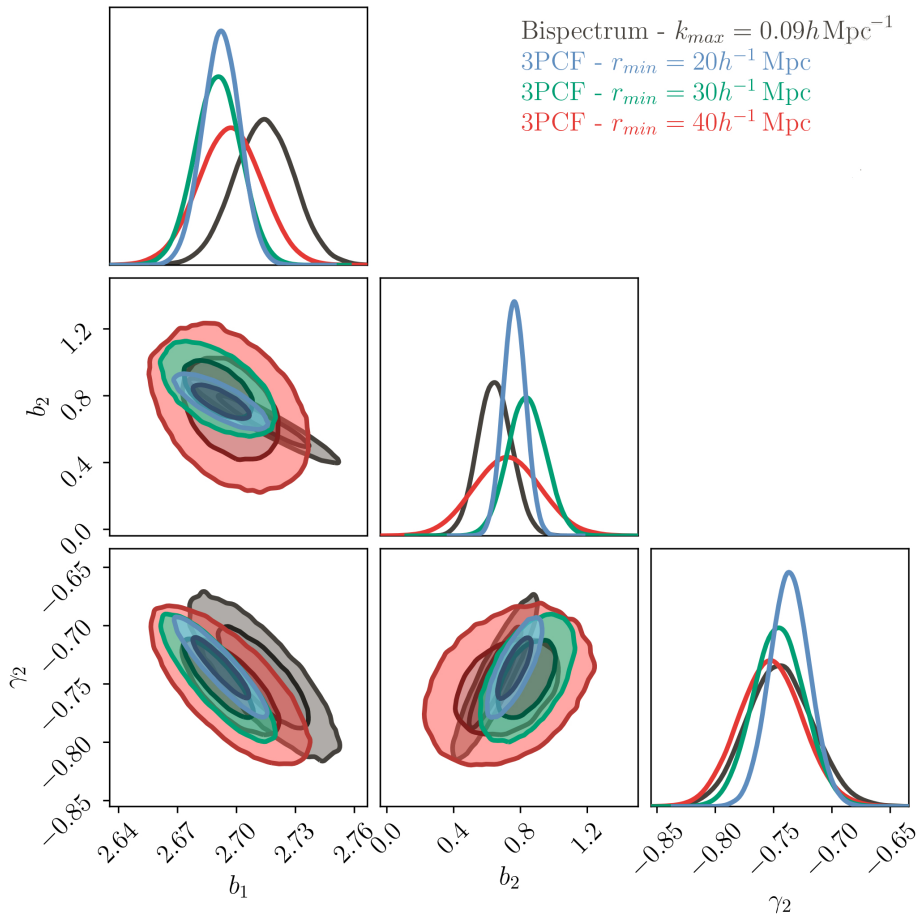
Figure 5. $1D - 2D$ posterior probability distributions of the bias parameters inferred from the likelihood analysis in the reference case. The green, blue and red contours (curves) are used for the cases the cases $r_{min} = 20, 30, 40\,h^{-1}\,\mathrm{Mpc}$ and $\eta_{min} = 3$. Reference dark grey contours (curves) are from the bispectrum analysis of [15].

not equally able to constrain bias parameters, particularly b_2, with the difference showing a notable dependence on the scales-cut. Yet, the overall agreement is an indication that our understanding of halo 3-point correlation in terms of local and tidal bias along with the control of systematic in both the Fourier-space as the configuration-space estimations and analyses are on rather solid grounds. A more detailed comparisons deserve a dedicated discussion that is beyond the aims of this paper.

6.2 Noisy estimates of the numerical covariance

In Section 3.3 we found no significant difference between the 3PCF covariance matrices estimated from the MINERVA simulations and the PINOCCHIO approximate mocks. We can nevertheless expect a bias, particularly on the inverse covariance matrix, when the numerical estimation is based on a small number of catalogs. As described in Sec. 5, we account for this enlarging the error bars to encompass possible systematic effects.

In Fig. 6 we compare the marginalized posteriors on the bias parameters obtained assuming the covariance matrix estimated from the small sets of 298 MINERVA simulations (red curves) and the one estimated from the same number of PINOCCHIO mocks (green) to the
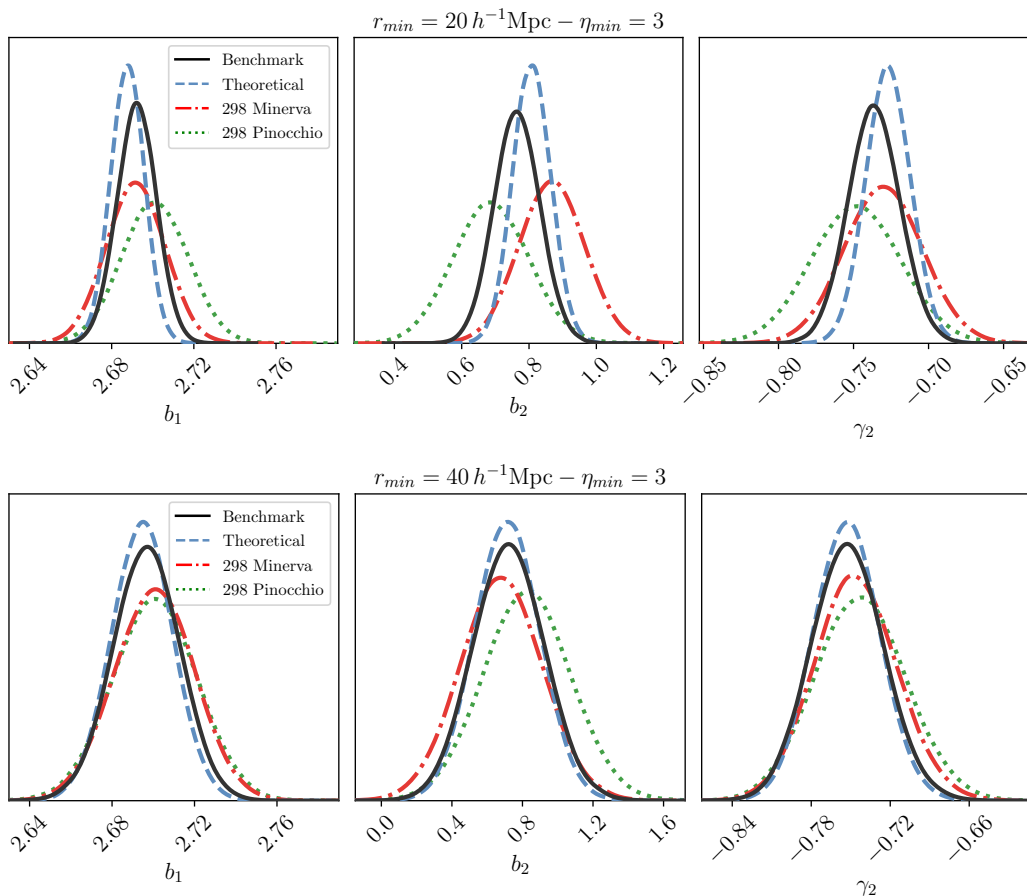
Figure 6. Marginalized posterior distribution of three bias parameters assuming $\eta_{min} = 3$, $r_{min} = 20$ (top panels) and $r_{min} = 40$ (bottom panels), obtained from different covariance matrices: numerical estimates from the full PINOCCHIO set (black continuous curve), small set of 298 catalogs from the MINERVA simulations (red dot-dashed), 298 catalogs from the PINOCCHIO mocks (green dotted) and the analytic Gaussian model (blue dashed).

reference case of the covariance from the full PINOCCHIO set. (black). The top panels assume $r_{min} = 20h^{-1}$ Mpc, while for the lower ones the data vector is limited to $r_{min} = 40h^{-1}$ Mpc. For $r_{min} = 20h^{-1}$ Mpc the main difference is clearly due to the correction factor applied to the poorly estimated covariances. In this case we notice, in addition, a shift of about 1-σ in the mean of the b_2, sign of a systematic error induced by a poor estimation of the covariance. We also notice a shift in the mean values between the 298 MINERVA and 298 PINOCCHIO cases, despite they share the same initial conditions. This suggests that noisy covariance obtained with approximated methods should be taken with care. This discrepancy disappears when excluding triangles with sides smaller that 40 h^{-1} Mpc from the analysis. Similar results are obtained with different values for η_{min} values and for larger r_{min} sizes and for this reason they are not shown here.

6.3   Analytical covariance model

We turn now to the effect of the analytical covariance on the on the likelihood analysis. The resulting posterior distributions of the bias parameters are shown as blue dashed curves in
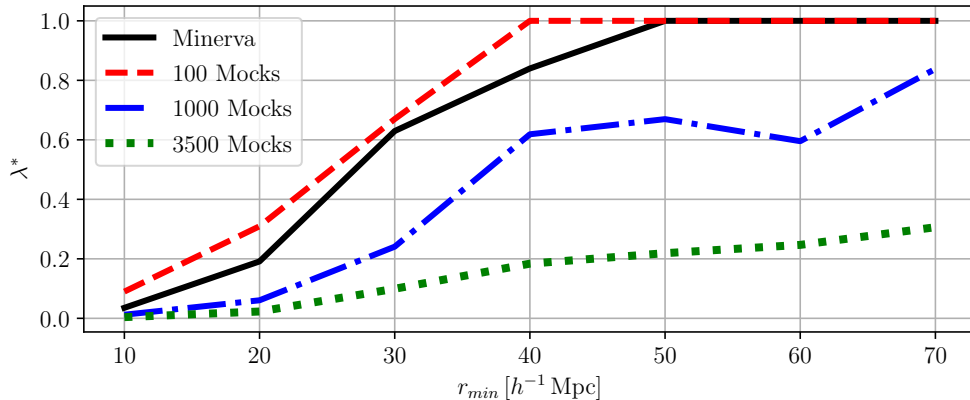
Figure 7. Value of the shrinkage intensity λ^* as a function of the minimum triangle scale. Different curves correspond to different choices of the numerical matrix $C_{i,j}$ estimated from 298 MINERVA simulations (black, continuous) and 100, 1000 and 3500 PINOCCHIO mocks (respectively red dashed, blue dot-dashed and green dotted). For this figure we select only triangles with $\eta_{min} = 3$.

Fig. 6. Not surprisingly, in the $r_{min} = 20 \, h^{-1}\,\mathrm{Mpc}$ case (top panels) the marginalized uncertainties are underestimated w.r.t. the case of the reference numerical covariance. However, the difference is not very large and we do not notice a significant bias on the central values of the parameters. The comparison improves significantly for $r_{min} = 40 \, h^{-1}\,\mathrm{Mpc}$ (bottom panels), indicating that the Gaussian hypothesis is rather adequate for 3PCF analyses above this scale.

6.4 Shrinkage covariance matrix

Several methods have been proposed to estimate the covariance matrix with a limited number of mocks while keeping a good accuracy [44, 59]. We consider here the shrinkage technique of [44], based on the combination of a noisy but unbiased estimate of the covariance (e.g. from N-body simulations) with a possibly biased but more precise estimate or noiseless model (e.g. from theory). The benefit of this combination is that of reducing the number of mocks but also, in our case, the non-negligible computational cost of estimating their 3PCFs.

In practice, we consider in the linear combination the (biased but noiseless) theoretical, Gaussian covariance $T_{i,j}$ with the numerical covariance, $\widehat{C}_{i,j}$, that is

$$S_{i,j} = \lambda^* T_{i,j} + (1 - \lambda^*)\widehat{C}_{i,j}, \tag{6.1}$$

where the parameter λ^* represent the *shrinkage intensity*. For $\lambda^* = 0$ one recover the numerical covariance while for $\lambda^* = 1$ we have the theory model. The optimal choice for λ is found by minimizing a mean-square difference cost function giving, in our case

$$\lambda^\star = \frac{\sum_{i,j} \widehat{\mathrm{Var}}\left(\widehat{C}_{ij}\right)}{\sum_{i,j} \left(T_{ij} - \widehat{C}_{ij}\right)^2}, \tag{6.2}$$

where $\widehat{\mathrm{Var}}\left(\widehat{C}_{ij}\right)$ is the variance associated with each $\widehat{C}_{i,j}$, that we compute as [44]

$$\widehat{\mathrm{Var}}\left(\widehat{C}_{ij}\right) = \frac{N_m}{(N_m - 1)^3} \sum_{\alpha=1}^{N_m} \left(W_{ij}^{(\alpha)} - \widehat{W}_{ij}\right)^2, \tag{6.3}$$
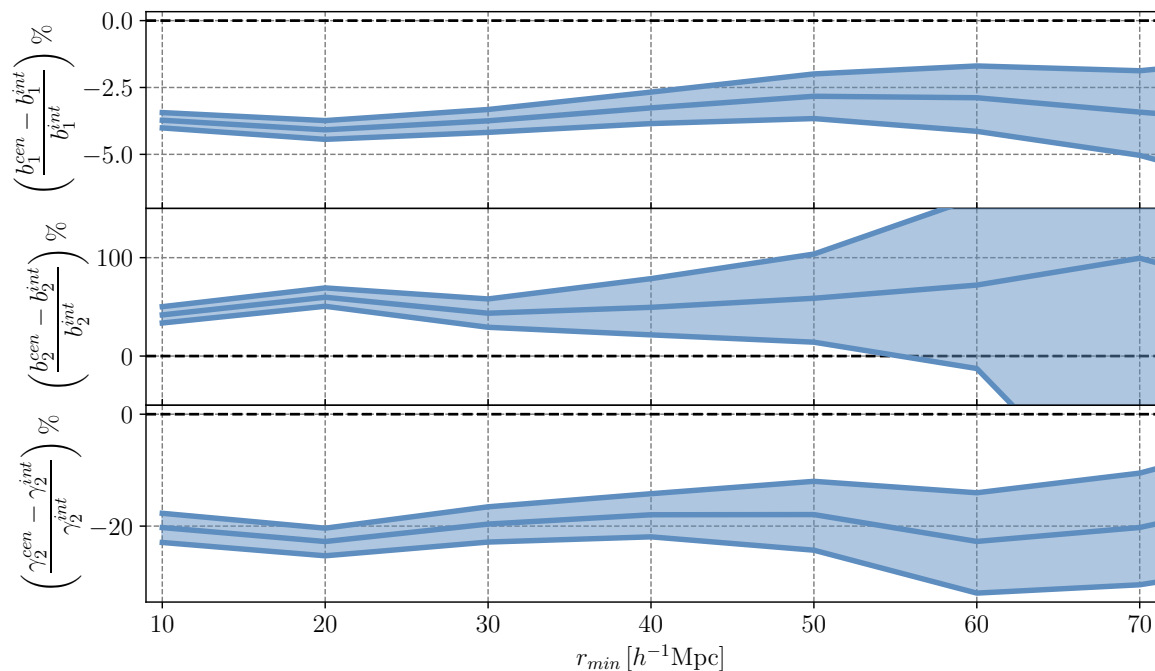
Figure 8. Percentage difference between bias constraints obtained with and without averaging over the triangle side bin, as a function of r_{min} for $\eta \geq 3$. The black dashed line drawn at the 0 value (perfect match) indicates when the two hypothesis give consistent constraints.

with

$$W_{ij}^{(\alpha)} = \left(\hat{\zeta}_i^{(\alpha)} - \overline{\zeta}_i \right) \left(\hat{\zeta}_j^{(\alpha)} - \overline{\zeta}_j \right) \tag{6.4}$$

computed in terms of the 3PCF measurement $\hat{\zeta}_i^{(\alpha)}$ for the i-th triangle in the α-th realization and the average $\overline{\zeta}_i$ at triangle i, and with

$$\widehat{W}_{ij} = \frac{1}{N_m} \sum_{\alpha=1}^{N_m} W_{ij}^{(\alpha)} . \tag{6.5}$$

We show our estimates of λ^* as a function of r_{min}, assuming $\eta_{min} = 3$, in fig. 7. Different curves correspond to different choices of the numerical matrix $\widehat{C}_{i,j}$. As expected, on large scales $\lambda^* \to 1$ since the covariance is approximately Gaussian, although when 1000 or more mocks are available, the parameter λ^* never reaches unity, showing relevant differences between the numerical estimate and analytical predictions at all scale. At small scales, where deviations from Gaussianity are larger, the numerical covariance is always favoured. In the MINERVA case, we also run the likelihood analysis with the shrinkage covariance. Unsurprisingly, the same trend found for λ^* reflects on the constraints on the bias parameters, which stay close to the constraints shown in Fig. 6.

6.5   Sensitivity to binning

In all the results presented so far we have used the 3PCF multipole model predictions obtained by performing the three-dimensional integral in eq. (4.4). This step significantly contributes to the total computing budget and becomes particularly burdensome when performed at each
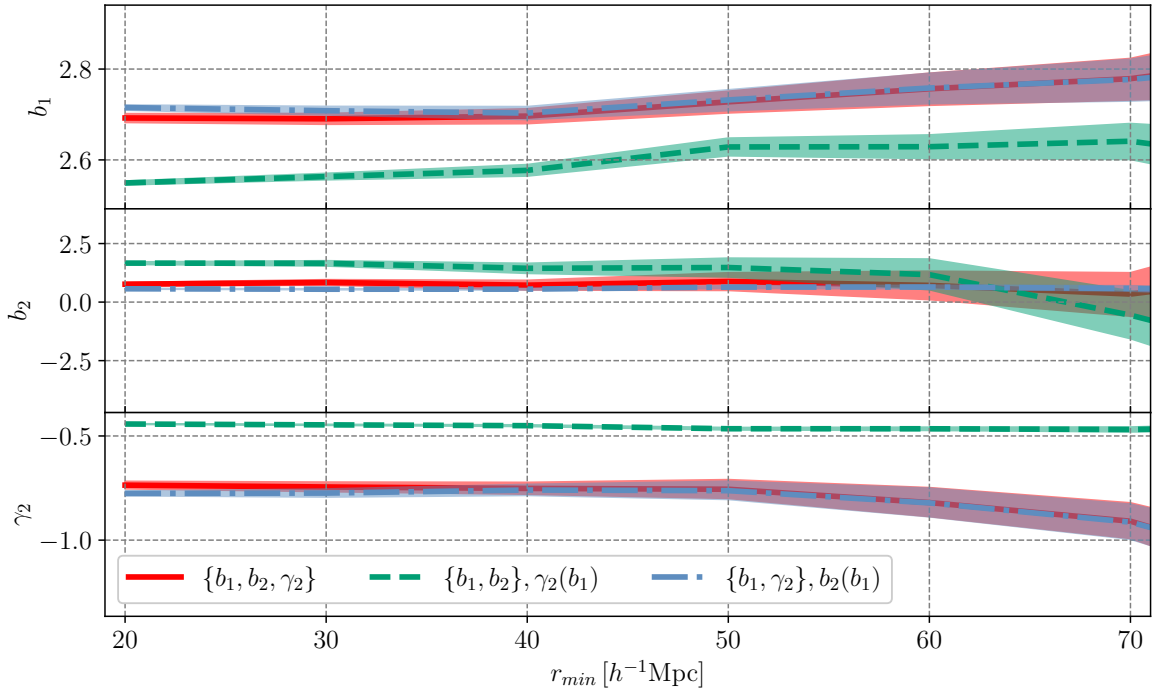
Figure 9. Best fit values of the bias parameters as a function of r_{min} obtained in the reference case (red continuous curve), assuming the LLB hypothesis (dashed green curves) and using the N-body calibrated fitting function (blue, dot-dashed curve). Dashed areas represent the 1-σ uncertainty interval. All results are obtained for triangles with $\eta \geq 3$.

likelihood evaluation in a Markov Chain exploring cosmological parameters. An alternative, cheaper option could be the simple angle integration

$$\zeta_\ell(r_{12}, r_{13}) = \frac{2\ell + 1}{2} \int_{-1}^{1} d\mu \mathcal{L}_\ell(\mu) \zeta(r_{12}, r_{13}, r_{23}), \qquad (6.6)$$

with r_{12} and r_{13} evaluated at the center of the separation bin.

In Fig. 8 we compare the results of the likelihood analysis obtained in these two cases. We show, for each bias parameters, the percent difference relative to the reference case in which the binning effect has been fully modelled, along with the relative marginalized uncertainty. The simplification given by the evaluation at the bin center can induce significant deviations from the reference values, as large as 50% in the case of b_2, and for all values of r_{min} up to $70 h^{-1}$ Mpc.

We acknowledge the fact that the effect may be particularly large due to the bin size of $\Delta r = 10 \, h^{-1}$ Mpc, and that using finer binning will reduce the mismatch. That said, our results show that the analysis is sensitive to the binning adopted and that an alternative definition eq. (6.6) to speed up the analysis should be considered carefully.

6.6 Bias relations

So far we assume all three bias parameters free to vary. However, these parameters are not fully independent, and several relations among them can be found either from theoretical considerations or as fits to numerical simulations. Here we consider specifically the local

Lagrangian bias (LLB) prescription for the tidal parameter [52, 53]

$$\gamma_2(b_1) = -\frac{2}{7}(b_1 - 1) \,, \tag{6.7}$$

and the numerical fit [60]:

$$b_2(b_1) = 0.412 - 2.143 \, b_1 + 0.929 \, b_1^2 + 0.008 \, b_1^3 + \frac{4}{3}\gamma_2 \,. \tag{6.8}$$

Fig. 9 shows the mean over marginalized posterior of bias parameters of the best fit and their 1-σ uncertainties as a function of r_{min} for the case of $\eta_{min} = 3$. The results obtained assuming the LLB prescription (dashed green curves) are not compatible with those obtained in the reference case (red continuous curves), especially for the derived γ_2 parameter, confirming previous results [15, 61, 62]. On the other hand, the posteriors obtained assuming the N-body-calibrated relation in eq. (6.8) (dot-dashed blue curves) are in very good agreement with the reference ones for all r_{min} values. A similarly good agreement has been found in the bispectrum analysis of [15].

7 Discussion and conclusions

This work extends to the 3-point correlation function the detailed analysis, previously presented for the power spectrum and bispectrum in [15, 17, 21, 23, 40, 63], of a large set of 298 halo catalogs from the MINERVA simulations [45], corresponding to a total volume of about 1,000 $h^{-3}\,\mathrm{Gpc}^3$. These are supplemented with an even larger set of 10,000 approximate halo catalogs obtained with the PINOCCHIO code on the same box and cosmology and reproducing the large-scale amplitude of the simulations' halo power spectrum. While not directly corresponding to any galaxy population from actual surveys, these large data-sets allow for rigorous test of several methodological choices in the analysis of galaxy clustering, taking advantage of a robust estimate of their covariance properties, even in the case of higher-order statistics.

We measure the real-space 3PCF from all the MINERVA halo catalogs as well as from the PINOCCHIO mocks. Each measurement includes all measurable triangular configurations with sides between 20 and $130h^{-1}\,\mathrm{Mpc}$, with only a small subset excluded due to the limitations of the fast 3PCF estimator of [29]. This is crucial to explore the potential of the 3PCF in constraining bias and cosmological parameters. In our tests, we focus on the determination of the three bias parameters required by the tree-level model in Perturbation Theory.

- We explored how the recovered bias parameter depend on the value of the smallest separation r_{min} included in the analysis as well as on the minimal, relative difference η_{min} between the sides r_{12} and r_{13}, defined in eq. (3.5). For $\eta_{min} = 1$ the sum of eq. (4.5) limited to $\ell_{max} = 10$ does not provide an accurate estimate of the full 3PCF ζ and consequently is not properly described by the tree-level model. We find that setting $\eta_{min} = 3$ leads to a consistent determination of the linear bias b_1 and the non-linear parameters b_2 and γ_2, with values independent of r_{min} for $r_{min} \geq 20\,h^{-1}\,\mathrm{Mpc}$. Some dependence of b_1 on r_{min} is evident instead for $\eta_{min} = 2$, while $\eta_{min} = 1$ leads to largely inconsistent results for all bias parameters.

- We compare our constraints on the bias parameters for $\eta_{min} = 3$ (assumed as reference value in all that follows) to the same constraints obtained from the bispectrum

analysis of [15]. We find a remarkable agreement, particularly for the scale cuts given by $r_{min} = 40h^{-1}\,\mathrm{Mpc}$ and $k_{max} = 0.09\,h\,\mathrm{Mpc}^{-1}$ for the 3PCF and bispectrum, respectively. This validates both the theoretical description of the large-scale 3-point statistics in terms of local and tidal bias contributions as well as our accurate estimation and modelling. This is, as far as we are aware, the most balanced comparison between the two statistics, based, on both sides, on all measurable configurations and on a robust numerical estimate of the covariance from a very large set of mocks.

- We compare the results obtained from covariance matrix estimated from the PINOC-CHIO mocks to those assuming a covariance obtained from more limited set of MINERVA N-body simulations (which also provide our data-vector), accounting for possible systematic errors due to poor statistics in terms of the factor proposed by [55]. The results are very similar for $r_{min} = 40h^{-1}\,\mathrm{Mpc}$, while we notice some systematic shift in the determination of b_2 for $r_{min} = 20h^{-1}\,\mathrm{Mpc}$, possibly due to small-scale suppression of power characteristic of the Lagrangian PT on which PINOCCHIO, as many other similar approximate methods, are based.

- We show a comparison of the same results with those obtained using an analytical model for the covariance matrix, assuming the Gaussian approximation [29]. Again we find largely consistent results for $r_{min} \geq 40\,h^{-1}\,\mathrm{Mpc}$, while including smaller scales we can notice a slight underestimate of the marginalized uncertainty for all parameters, w.r.t. the reference numerical covariance. This is expected also from a direct comparison in terms of the variance alone, with the analytical model providing lower values at small scales.

- We investigate how the shrinkage technique [44] can provide an hybrid estimate of the covariance based on the possibly biased analytical model and on a noisy numerical estimate, determining the shrinkage parameter λ^* as a function of r_{min} for different choices of the numerical matrix. We find that in all cases the numerical result is preferred when $r_{min} \lesssim 30h^{-1}\,\mathrm{Mpc}$, but also that a significant numerical component is retained when the number of available mocks is sufficiently large (~ 1000).

- We compare our reference analysis, fully accounting for the size of the separation bins of the 3PCF estimator, to a more efficient evaluation that neglects binning effects, finding large systematic errors on the bias parameters. Although such errors can be reduced by using a smaller bin size than the one adopted here ($10h^{-1}\,\mathrm{Mpc}$), a careful assessment has to be made before adopting this approximation, also considering that a smaller bin implies a much larger covariance matrix, with all its significant numerical requirements.

- Finally, we consider relations among the bias parameters in order to reduce the parameter space. We confirmed that the simple local Lagrangian expression for the tidal bias leads to significant systematic errors in the 3PCF case as well. The relation between b_2 and b_1 obtained by [60] as a fit to numerical simulations, instead leads to results consistent with our reference case where all three parameters are left free.

With this work we have provided a solid test of 3-point correlation function as a useful tool in the determination of second-order bias parameters, showing constraints consistent to those from a bispectrum analysis. A fully numerical estimates of the 3PCF covariance matrix, however, is always a very numerically demanding task. We discussed here possible strategies

to tackle this problem and provide ideal tests based on periodic-box simulations. We notice that we could consider other techniques to compress the information, avoiding the explicit need of a large set of mocks to analyse [see e.g. 29, 49, 64, 65]. We will explore these techniques in more details and more realistic scenarios in future analyses, quantifying the effect of specific survey geometry, as well as test a full likelihood analysis in redshift space. Other possible extensions of this work go in the direction of testing more sophisticated models (e.g. [66]) to extract as much information as possible on both bias and cosmological parameters.

Acknowledgments

References

[1] DESI Collaboration, A. Aghamousa, J. Aguilar, S. Ahlen, S. Alam, L. E. Allen et al., *The DESI Experiment Part I: Science,Targeting, and Survey Design*, *arXiv e-prints* (2016) arXiv:1611.00036 [1611.00036].

[2] DESI Collaboration, A. Aghamousa, J. Aguilar, S. Ahlen, S. Alam, L. E. Allen et al., *The DESI Experiment Part II: Instrument Design*, *arXiv e-prints* (2016) arXiv:1611.00037 [1611.00037].

[3] R. Laureijs, J. Amiaux, S. Arduini, J. L. Auguères, J. Brinchmann, R. Cole et al., *Euclid Definition Study Report*, *arXiv e-prints* (2011) arXiv:1110.3193 [1110.3193].

[4] R. Akeson, L. Armus, E. Bachelet, V. Bailey, L. Bartusek, A. Bellini et al., *The Wide Field Infrared Survey Telescope: 100 Hubbles for the 2020s*, *arXiv e-prints* (2019) arXiv:1902.05569 [1902.05569].

[5] H. Gil-Marín, W. J. Percival, L. Verde, J. R. Brownstein, C.-H. Chuang, F.-S. Kitaura et al., *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies*, *Mon. Not. R. Astron. Soc.* **465** (2017) 1757 [1606.00439].

[6] G. d'Amico, J. Gleyzes, N. Kokron, K. Markovic, L. Senatore, P. Zhang et al., *The cosmological analysis of the SDSS/BOSS data from the Effective Field Theory of Large-Scale Structure*, *Journal of Cosmology and Astro-Particle Physics* **2020** (2020) 005 [1909.05271].

[7] G. D'Amico, M. Lewandowski, L. Senatore and P. Zhang, *Limits on primordial non-Gaussianities from BOSS galaxy-clustering data*, *arXiv e-prints* (2022) arXiv:2201.11518 [2201.11518].

[8] O. H. E. Philcox and M. M. Ivanov, *BOSS DR12 full-shape cosmology: Λ CDM constraints from the large-scale galaxy power spectrum and bispectrum monopole*, *Phys. Rev. D* **105** (2022) 043517 [2112.04515].

[9] G. Cabass, M. M. Ivanov, O. H. E. Philcox, M. Simonović and M. Zaldarriaga, *Constraints on Single-Field Inflation from the BOSS Galaxy Survey*, *arXiv e-prints* (2022) arXiv:2201.07238 [2201.07238].

[10] G. Cabass, M. M. Ivanov, O. H. E. Philcox, M. Simonović and M. Zaldarriaga, *Constraints on Multi-Field Inflation from the BOSS Galaxy Survey*, *arXiv e-prints* (2022) arXiv:2204.01781 [2204.01781].

[11] E. Sefusatti, M. Crocce, S. Pueblas and R. Scoccimarro, *Cosmology and the bispectrum*, *Phys. Rev. D* **74** (2006) 023522 [arXiv: astro-ph/0604505].

[12] K. C. Chan and L. Blot, *Assessment of the information content of the power spectrum and bispectrum*, *Phys. Rev. D* **96** (2017) 023528 [1610.06585].

[13] M. Colavincenzo, E. Sefusatti, P. Monaco, L. Blot, M. Crocce, M. Lippich et al., *Comparing approximate methods for mock catalogues and covariance matrices - III: bispectrum*, *Mon. Not. R. Astron. Soc.* **482** (2019) 4883 [1806.09499].

[14] C. Hahn, F. Villaescusa-Navarro, E. Castorina and R. Scoccimarro, *Constraining M_ν with the bispectrum. Part I. Breaking parameter degeneracies*, *Journal of Cosmology and Astro-Particle Physics* **2020** (2020) 040 [1909.11107].

[15] A. Oddo, E. Sefusatti, C. Porciani, P. Monaco and A. G. Sánchez, *Toward a robust inference method for the galaxy bispectrum: likelihood function and model selection*, *Journal of Cosmology and Astro-Particle Physics* **2020** (2020) 056 [1908.01774].

[16] D. Gualdi, H. Gil-Marín and L. Verde, *Joint analysis of anisotropic power spectrum, bispectrum and trispectrum: application to N-body simulations*, *Journal of Cosmology and Astro-Particle Physics* **2021** (2021) 008 [2104.03976].

[17] A. Oddo, F. Rizzo, E. Sefusatti, C. Porciani and P. Monaco, *Cosmological parameters from the likelihood analysis of the galaxy power spectrum and bispectrum in real space*, *Journal of Cosmology and Astro-Particle Physics* **2021** (2021) 038 [2108.03204].

[18] T. Baldauf, M. Garny, P. Taule and T. Steele, *Two-loop bispectrum of large-scale structure*, *Phys. Rev. D* **104** (2021) 123551 [2110.13930].

[19] A. Moradinezhad Dizgah, M. Biagetti, E. Sefusatti, V. Desjacques and J. Noreña, *Primordial non-Gaussianity from biased tracers: likelihood analysis of real-space power spectrum and bispectrum*, *Journal of Cosmology and Astro-Particle Physics* **2021** (2021) 015 [2010.14523].

[20] M. M. Ivanov, O. H. E. Philcox, T. Nishimichi, M. Simonović, M. Takada and M. Zaldarriaga, *Precision analysis of the redshift-space galaxy bispectrum*, *arXiv e-prints* (2021) arXiv:2110.10161 [2110.10161].

[21] D. Alkhanishvili, C. Porciani, E. Sefusatti, M. Biagetti, A. Lazanu, A. Oddo et al., *The reach of next-to-leading-order perturbation theory for the matter bispectrum*, *arXiv e-prints* (2021) arXiv:2107.08054 [2107.08054].

[22] A. Eggemeier, R. Scoccimarro, R. E. Smith, M. Crocce, A. Pezzotta and A. G. Sánchez, *Testing one-loop galaxy bias: Joint analysis of power spectrum and bispectrum*, *Phys. Rev. D* **103** (2021) 123550 [2102.06902].

[23] F. Rizzo, C. Moretti, K. Pardede, A. Eggemeier, A. Oddo, E. Sefusatti et al., *The Halo Bispectrum Multipoles in Redshift Space*, *arXiv e-prints* (2022) arXiv:2204.13628 [2204.13628].

[24] Y. P. Jing and G. Boerner, *Three-point correlation function in the quasilinear regime.*, *Astron. Astrophys.* **318** (1997) 667 [astro-ph/9606122].

[25] J. Barriga and E. Gaztañaga, *The three-point function in large-scale structure - I. The weakly non-linear regime in N-body simulations*, *Mon. Not. R. Astron. Soc.* **333** (2002) 443 [astro-ph/0112278].

[26] J. Bel, K. Hoffmann and E. Gaztañaga, *Non-local bias contribution to third-order galaxy correlations*, *Mon. Not. R. Astron. Soc.* **453** (2015) 259 [1504.02074].

[27] K. Hoffmann, E. Gaztañaga, R. Scoccimarro and M. Crocce, *Testing the consistency of three-point halo clustering in Fourier and configuration space*, *Mon. Not. R. Astron. Soc.* **476** (2018) 814 [1708.08941].

[28] J. Kuruvilla and C. Porciani, *The n-point streaming model: how velocities shape correlation functions in redshift space*, *Journal of Cosmology and Astro-Particle Physics* **2020** (2020) 043 [2005.05331].

[29] Z. Slepian and D. J. Eisenstein, *Computing the three-point correlation function of galaxies in $O(N^2)$ time*, *Mon. Not. R. Astron. Soc.* **454** (2015) 4142 [1506.02040].

[30] Z. Slepian and D. J. Eisenstein, *A practical computational method for the anisotropic redshift-space three-point correlation function*, *Mon. Not. R. Astron. Soc.* **478** (2018) 1468 [1709.10150].

[31] N. S. Sugiyama, S. Saito, F. Beutler and H.-J. Seo, *A complete FFT-based decomposition formalism for the redshift-space bispectrum*, *Mon. Not. R. Astron. Soc.* **484** (2019) 364 [1803.02132].

[32] O. Umeh, *Optimal computation of anisotropic galaxy three point correlation function multipoles using 2DFFTLOG formalism*, *Journal of Cosmology and Astro-Particle Physics* **2021** (2021) 035 [2011.05889].

[33] Z. Slepian, D. J. Eisenstein, J. R. Brownstein, C.-H. Chuang, H. Gil-Marín, S. Ho et al., *Detection of baryon acoustic oscillation features in the large-scale three-point correlation function of sdss boss dr12 cmass galaxies*, *Mon. Not. R. Astron. Soc.* **469** (2017) 1738 [1607.06097].

[34] M. Moresco, A. Veropalumbo, F. Marulli, L. Moscardini and A. Cimatti, *C³: Cluster Clustering Cosmology. II. First Detection of the Baryon Acoustic Oscillations Peak in the Three-point Correlation Function of Galaxy Clusters*, *Astrophys. J.* **919** (2021) 144 [2011.04665].

[35] A. Veropalumbo, I. Sáez Casares, E. Branchini, B. R. Granett, L. Guzzo, F. Marulli et al., *A joint 2- and 3-point clustering analysis of the VIPERS PDR2 catalogue at z 1: breaking the degeneracy of cosmological parameters*, *Mon. Not. R. Astron. Soc.* **507** (2021) 1184 [2106.12581].

[36] Z. Slepian and D. J. Eisenstein, *Modelling the large-scale redshift-space 3-point correlation function of galaxies*, *Mon. Not. R. Astron. Soc.* **469** (2017) 2059 [1607.03109].

[37] N. S. Sugiyama, S. Saito, F. Beutler and H.-J. Seo, *Towards a self-consistent analysis of the anisotropic galaxy two- and three-point correlation functions on large scales: application to mock galaxy catalogues*, *Mon. Not. R. Astron. Soc.* **501** (2021) 2862 [2010.06179].

[38] I. Szapudi and A. S. Szalay, *A New Class of Estimators for the N-Point Correlations*, *Astrophys. J. Lett.* **494** (1998) L41.

[39] O. H. E. Philcox, *Cosmology without window functions. II. Cubic estimators for the galaxy bispectrum*, *Phys. Rev. D* **104** (2021) 123529 [2107.06287].

[40] K. Pardede, F. Rizzo, M. Biagetti, E. Castorina, E. Sefusatti and P. Monaco, *Bispectrum-window convolution via Hankel transform*, *arXiv e-prints* (2022) arXiv:2203.04174 [2203.04174].

[41] P. Monaco, *Approximate Methods for the Generation of Dark Matter Halo Catalogs in the Age of Precision Cosmology*, *Galaxies* **4** (2016) 53 [1605.07752].

[42] P. Monaco, T. Theuns and G. Taffoni, *The pinocchio algorithm: pinpointing orbit-crossing collapsed hierarchical objects in a linear density field*, *Mon. Not. R. Astron. Soc.* **331** (2002) 587 [astro-ph/0109323].

[43] E. Munari, P. Monaco, E. Sefusatti, E. Castorina, F. G. Mohammad, S. Anselmi et al., *Improving fast generation of halo catalogues with higher order Lagrangian perturbation theory*, *Mon. Not. R. Astron. Soc.* **465** (2017) 4658 [1605.04788].

[44] A. C. Pope and I. Szapudi, *Shrinkage estimation of the power spectrum covariance matrix*, *Mon. Not. R. Astron. Soc.* **389** (2008) 766 [0711.2509].

[45] J. N. Grieb, A. G. Sánchez, S. Salazar-Albornoz and C. Dalla Vecchia, *Gaussian covariance matrices for anisotropic galaxy clustering measurements*, *Mon. Not. R. Astron. Soc.* **457** (2016) 1577 [1509.04293].

[46] M. Lippich, A. G. Sánchez, M. Colavincenzo, E. Sefusatti, P. Monaco, L. Blot et al., *Comparing approximate methods for mock catalogues and covariance matrices - I. Correlation function*, *Mon. Not. R. Astron. Soc.* **482** (2019) 1786 [1806.09477].

[47] V. Springel, *The cosmological simulation code GADGET-2*, *Mon. Not. R. Astron. Soc.* **364** (2005) 1105 [astro-ph/0505010].

[48] O. H. E. Philcox and D. J. Eisenstein, *Estimating covariance matrices for two- and three-point correlation function moments in Arbitrary Survey Geometries*, *Mon. Not. R. Astron. Soc.* **490** (2019) 5931 [1910.04764].

[49] A. Fumagalli, M. Biagetti, A. Saro, E. Sefusatti, A. Slosar, P. Monaco et al., *Fitting covariance matrices models to simulations*, *arXiv e-prints* (2022) arXiv:2206.05191 [2206.05191].

[50] Z. Slepian, D. J. Eisenstein, F. Beutler, C.-H. Chuang, A. J. Cuesta, J. Ge et al., *The large-scale three-point correlation function of the SDSS BOSS DR12 CMASS galaxies*, *Mon. Not. R. Astron. Soc.* **468** (2017) 1070 [1512.02231].

[51] J. N. Fry and E. Gaztañaga, *Biasing and hierarchical statistics in large-scale structure*, *Astrophys. J.* **413** (1993) 447 [astro-ph/9302009].

[52] K. C. Chan, R. Scoccimarro and R. K. Sheth, *Gravity and large-scale nonlocal bias*, *Phys. Rev. D* **85** (2012) 083509 [1201.3614].

[53] T. Baldauf, U. Seljak, V. Desjacques and P. McDonald, *Evidence for quadratic tidal tensor bias from the halo bispectrum*, *Phys. Rev. D* **86** (2012) 083540 [1201.4827].

[54] A. Lewis, A. Challinor and A. Lasenby, *Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models*, *Astrophys. J.* **538** (2000) 473 [astro-ph/9911177].

[55] J. Hartlap, P. Simon and P. Schneider, *Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix*, *Astron. Astrophys.* **464** (2007) 399 [astro-ph/0608064].

[56] E. Sellentin and A. F. Heavens, *Parameter inference with estimated covariance matrices*, *Mon. Not. R. Astron. Soc.* **456** (2016) L132 [1511.05969].

[57] W. J. Percival, A. J. Ross, A. G. Sánchez, L. Samushia, A. Burden, R. Crittenden et al., *The clustering of Galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: including covariance matrix errors*, *Mon. Not. R. Astron. Soc.* **439** (2014) 2531 [1312.4841].

[58] D. Foreman-Mackey, D. W. Hogg, D. Lang and J. Goodman, *emcee: The MCMC Hammer*, *Publications of the Astronomical Society of the Pacific* **125** (2013) 306 [1202.3665].

[59] D. J. Paz and A. G. Sánchez, *Improving the precision matrix for precision cosmology*, *Mon. Not. R. Astron. Soc.* **454** (2015) 4326 [1508.03162].

[60] T. Lazeyras, C. Wagner, T. Baldauf and F. Schmidt, *Precision measurement of the local bias of dark matter halos*, *Journal of Cosmology and Astro-Particle Physics* **2016** (2016) 018 [1511.01096].

[61] T. Lazeyras and F. Schmidt, *Beyond LIMD bias: a measurement of the complete set of third-order halo bias parameters*, *Journal of Cosmology and Astro-Particle Physics* **2018** (2018) 008 [1712.07531].

[62] M. M. Abidi and T. Baldauf, *Cubic halo bias in Eulerian and Lagrangian space*, *Journal of Cosmology and Astro-Particle Physics* **2018** (2018) 029 [1802.07622].

[63] J. Byun, A. Oddo, C. Porciani and E. Sefusatti, *Towards cosmological constraints from the compressed modal bispectrum: a robust comparison of real-space bispectrum estimators*, *Journal of Cosmology and Astro-Particle Physics* **2021** (2021) 105 [2010.09579].

[64] D. Gualdi, H. Gil-Marín, M. Manera, B. Joachimi and O. Lahav, *GEOMAX: beyond linear compression for three-point galaxy clustering statistics*, *Mon. Not. R. Astron. Soc.* **497** (2020) 776 [1912.01011].

[65] O. H. E. Philcox, M. M. Ivanov, M. Zaldarriaga, M. Simonović and M. Schmittfull, *Fewer mocks and less noise: Reducing the dimensionality of cosmological observables with subspace projections*, *Phys. Rev. D* **103** (2021) 043508 [2009.03311].

[66] A. Eggemeier, R. Scoccimarro and R. E. Smith, *Bias loop corrections to the galaxy bispectrum*, *Phys. Rev. D* **99** (2019) 123514 [1812.03208].

[67] D. Gualdi, H. Gil-Marín, M. Manera, B. Joachimi and O. Lahav, *Geometrical compression: a new method to enhance the BOSS galaxy bispectrum monopole constraints*, *Mon. Not. R. Astron. Soc.* (2019) L1 [1901.00987].

[68] F. Bernardeau, S. Colombi, E. Gaztañaga and R. Scoccimarro, *Large-scale structure of the Universe and cosmological perturbation theory*, Phys. Rep. **367** (2002) 1 [astro-ph/0112551].

[69] L. Blot, M. Crocce, E. Sefusatti, M. Lippich, A. G. Sánchez, M. Colavincenzo et al., *Comparing approximate methods for mock catalogues and covariance matrices II: power spectrum multipoles*, *Mon. Not. R. Astron. Soc.* **485** (2019) 2806 [1806.09497].

[70] E. Sefusatti, M. Crocce, R. Scoccimarro and H. M. P. Couchman, *Accurate estimators of correlation functions in Fourier space*, *Mon. Not. R. Astron. Soc.* **460** (2016) 3624 [1512.07295].

[71] V. Yankelevich, I. G. McCarthy, J. Kwan, S. G. Stafford and J. Liu, *The halo bispectrum as a sensitive probe of massive neutrinos and baryon physics*, *arXiv e-prints* (2022) arXiv:2202.07680 [2202.07680].

[72] S. Avila, S. G. Murray, A. Knebe, C. Power, A. S. G. Robotham and J. Garcia-Bellido, *HALOGEN: a tool for fast generation of mock halo catalogues*, *Mon. Not. R. Astron. Soc.* **450** (2015) 1856 [1412.5228].

[73] A. Izard, M. Crocce and P. Fosalba, *ICE-COLA: towards fast and accurate synthetic galaxy catalogues optimizing a quasi-N-body method*, *Mon. Not. R. Astron. Soc.* **459** (2016) 2327 [1509.04685].

[74] F. S. Kitaura, G. Yepes and F. Prada, *Modelling baryon acoustic oscillations with perturbation theory and stochastic halo biasing.*, *Mon. Not. R. Astron. Soc.* **439** (2014) L21 [1307.3285].

[75] G. Stein, M. A. Alvarez and J. R. Bond, *The mass-Peak Patch algorithm for fast generation of deep all-sky dark matter halo catalogues and its N-body validation*, *Mon. Not. R. Astron. Soc.* **483** (2019) 2236 [1810.07727].

[76] V. Desjacques, D. Jeong and F. Schmidt, *Large-scale galaxy bias*, Phys. Rep. **733** (2018) 1 [1611.09787].