Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **497,** 776–792 (2020) doi:10.1093/mnras/staa1941
Advance Access publication 2020 July 6

# GEOMAX: beyond linear compression for three-point galaxy clustering statistics

Davide Gualdi [1,2]★ Héctor Gil-Marín [1,2] Marc Manera,[3,4] Benjamin Joachimi[5] and Ofer Lahav[5]

[1]*ICC, University of Barcelona, IEEC-UB, Martí i Franquès, 1, E-08028 Barcelona, Spain*
[2]*Institute of Space Studies of Catalonia (IEEC), E-08034 Barcelona, Spain*
[3]*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra (Barcelona), Spain*
[4]*Centre for Mathematical Sciences, DAMTP, Cambridge University, Wilberforce Rd, Cambridge CB3 0WA, UK*
[5]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*

Accepted 2020 July 1. Received 2020 June 2; in original form 2019 December 4

## ABSTRACT

We present the GEOMAX algorithm and its PYTHON implementation for a two-step compression of bispectrum measurements. The first step groups bispectra by the geometric properties of their arguments; the second step then maximizes the Fisher information with respect to a chosen set of model parameters in each group. The algorithm only requires the derivatives of the data vector with respect to the parameters and a small number of mock data, producing an effective, non-linear compression. By applying GEOMAX to bispectrum monopole measurements from BOSS DR12 CMASS redshift-space galaxy clustering data, we reduce the 68 per cent credible intervals for the inferred parameters ($b_1$, $b_2$, $f$, $\sigma_8$) by 50.4, 56.1, 33.2, and 38.3 per cent with respect to standard MCMC on the full data vector. We run the analysis and comparison between compression methods over 100 galaxy mocks to test the statistical significance of the improvements. On average, GEOMAX performs ∼15 per cent better than geometrical or maximal linear compression alone and is consistent with being lossless. Given its flexibility, the GEOMAX approach has the potential to optimally exploit three-point statistics of various cosmological probes like weak lensing or line-intensity maps from current and future cosmological data sets such as DESI, *Euclid*, PFS, and SKA.

**Key words:** methods: analytical – cosmological parameters – large-scale structure of Universe.

## 1 INTRODUCTION

Recent applications of three-point (3pt) statistics to cosmological data sets, such as the one provided by the Sloan Digital Sky Survey[1] (Eisenstein et al. 2011), have shown their potential in improving the current constraints on cosmological parameters (Gil-Marín et al. 2017; Slepian et al. 2017a). For example, the physics of baryonic acoustic oscillations has been investigated via 3pt statistics (Slepian et al. 2017b; Pearson & Samushia 2018) and new methods have been developed to better capture it (Child et al. 2018), improving the results obtained via standard 2pt methods.

In preparation for future game-changing data sets like DESI[2] (Levi et al. 2013), *Euclid*[3] (Laureijs et al. 2011), PFS[4] (Ellis et al. 2014), SKA[5] (Bacon et al. 2018), and LSST[6] (Abell et al. 2009), the scientific community has been working on improving the modelling of 3pt statistics and extending their applications.

Bertacca et al. (2018), Clarkson et al. (2019), and Di Dio et al. (2019) studied the full-sky angular galaxy bispectrum beyond the flat-sky approximation including relativistic effects and corrections, Yamamoto, Nan & Hikage (2017), Nan, Yamamoto & Hikage (2018), and Sugiyama et al. (2019) proposed a complete multipole decomposition for the galaxy bispectrum. Sabiu et al. (2019) proposed a technique to speed up the computation of 3pt and higher order statistics, while D'Amico et al. (2020) applied the effective field theory formalism for the galaxy bispectrum on data.

In the weak lensing field, progress has been made since the first seminal works (Takada & Jain 2004; Kilbinger & Schneider 2005), Rizzato et al. (2019) investigated the information content of the bispectrum, while Kayo, Takada & Jain (2013) forecasted parameters constraints using a covariance matrix with terms beyond the Gaussian approximation. 3pt statistic have also been applied to line intensity map probes (Beane & Lidz 2018; Hoffmann et al. 2019; Schmit, Heavens & Pritchard 2019; Watkinson et al. 2019).

Studies on the bispectrum covariance and the bispectrum information content have also been realized in the recent past (Barreira 2019; Colavincenzo et al. 2019; Yankelevich & Porciani 2019; Oddo et al. 2020), while Ruggeri et al. (2018), Hahn et al. (2020), and Coulton et al. (2019) proved that the bispectrum can also help with improving the sum of the neutrino masses constraints. The accuracy of the bispectrum modelling was extended towards non-linear scales by including loop corrections (Hashimoto, Rasera & Taruya 2017; Desjacques, Jeong & Schmidt 2018; Castiblanco et al. 2019; Eggemeier, Scoccimarro & Smith 2019). Loop corrections for the bispectrum (Sefusatti 2009; Sefusatti, Crocce &

★ E-mail: dgualdi@icc.ub.edu
[1]http://www.sdss3.org/surveys/boss.php.
[2]http://desi.lbl.gov.
[3]http://sci.esa.int/euclid/.
[4]http://pfs.ipmu.jp.
[5]https://www.skatelescope.org.
[6]https://www.lsst.org/.

Desjacques 2012) must indeed be included if one wants to study primordial non-Gaussianity using mildly non-linear scales to lift the degeneracy with the other cosmological parameters (Scoccimarro, Sefusatti & Zaldarriaga 2004; Bose & Taruya 2018; Karagiannis et al. 2018).

Compression plays an essential role in the analysis of 3pt statistics. Indeed, the difficulties to perform these analyses include the dimension of the associated data vectors due to the limited number of simulations available to estimate the covariance matrix (Hartlap, Simon & Schneider 2007; Taylor & Joachimi 2014) and the computational challenge linked with the handling of large data vectors. In particular, when the estimators of these data vectors are expressed in terms of multipole expansion, even including the quadrupole term becomes prohibitive. An alternative approach to compression would be to accurately model the analytic covariance matrix as done by Sugiyama et al. (2020), which helps with cross-validating the obtained results.

In a previous paper, we introduced two methods achieving 'maximal' compression for the redshift space galaxy bispectrum (Gualdi et al. 2018), and tested them on bispectrum monopole measurements from Baryon Oscillation Spectroscopic Survey (BOSS) DR12 data (Gualdi et al. 2019b). These maximal compression methods transform the original data vector into a new one with dimension equal to the number of parameters considered in the analysis. In order to work, they require an approximate analytical expression of the original data vector covariance matrix.

More recently, in a second paper, we presented the geometrical compression algorithm (Gualdi et al. 2019a), which is based on averaging the bispectra of wavenumber triangle configurations having similar geometrical properties. Geometrical compression does not require an analytical expression for the covariance matrix but also does not take into account correlation between different triangle configurations.

In this work, we present an algorithm along with its PYTHON implementation[7] that combines maximal and geometrical compression, accounting for the correlation between bispectra without needing an analytical expression for the covariance matrix. This is achieved in two steps. First, we define triangle sets using the geometrical criteria. Secondly, we apply the maximal compression separately to each of the triangles sets. A limited number of simulations for the maximal compression step is still needed. However, by maximally compressing each triangle set, the required number is much lower than what usually required to estimate the covariance matrix for the full data vector.

We perform our analysis on measurements from both BOSS DR12 CMASS data (Dawson et al. 2013) and on 100 realizations of the relative set of mock data (Kitaura et al. 2016).

In the main analysis, we sample the joint posterior distribution for four model parameters: matter-galaxy bias parameters $b_1$ and $b_2$, the growth rate $f$, and the amplitude of dark matter oscillations $\sigma_8$. The main result of this work is a further improvement for the 68 per cent credible intervals of the inferred parameters when using the joint data vector including power spectrum (monopole and quadrupole) together with the bispectrum (monopole).

We run a series of tests to verify the added value of GEOMAX with respect to the previous methods. First, we consider alternative ways to first regroup triangles before applying the maximal compression step. Secondly, we run the analysis for alternative parameter sets. In one case, we only add the local primordial non-Gaussianity parameter $f_{\rm NL}$. The bispectrum has indeed, especially for future data set, the potential to produce constraints on $f_{\rm NL}$ of similar order (Verde et al. 2000; Jeong & Komatsu 2009; Sefusatti 2009) to the ones obtained by Planck (Akrami et al. 2019).

In the second set, we use $A_{\rm s}$, the amplitude of scalar perturbations, and the matter density parameter, $\Omega_{\rm m}$. These are of interest in order to obtain complementary late-time estimates on quantities very well constrained by cosmic microwave background (CMB) experiments (Aghanim et al. 2018).

This paper is organized as follows: In Section 2, we briefly recap the maximal and geometrical compression methods. Section 3 explains how to optimally combine maximal and geometrical compression methods, together with presenting the code structure. The analysis results for the main four parameters case are reported in Section 4 for both galaxy mocks and data. In Section 5, we consider two alternatives to the geometrical compression step, while in Section 6, we repeat the analysis for two larger parameter sets. We conclude in Section 7. In Appendix A, the analytical expression for the used data vectors are reported, while in Appendix B, we re-derive the primordial non-Gaussianity leading correction terms for the power spectrum and bispectrum. Before starting with the main part of this paper, we summarize below the analysis setup.

### 1.1 Analysis setup

For a fair comparison with our previous papers, we once more apply the compression to the measurements of the galaxy bispectrum monopole from the DR12 CMASS sample ($0.43 \leq z \leq 0.70$) of the BOSS (Dawson et al. 2013), which is part of the Sloan Digital Sky Survey III (Eisenstein et al. 2011).

We use 1400 realizations of the MultiDark Patchy galaxy catalogues for the BOSS DR12 data set by Kitaura et al. (2016). These mocks were realized having a fiducial cosmology with parameters $\Omega_\Lambda(z=0) = 0.693$, $\Omega_{\rm m}(z=0) = 0.307$, $\Omega_{\rm b}(z=0) = 0.048$, $\sigma_8 = 0.829$, $n_{\rm s} = 0.96$, and $h_0 = 0.678$.

The data vector used for the parameter constraints analysis is given by joining the galaxy tree-level power spectrum monopole and quadrupole measurements to the bispectrum monopole ones. The analytical expressions are given in Appendix A. For the power spectrum part, we use a bin size of $\Delta k = 0.01\,h\,{\rm Mpc}^{-1}$, while for the bispectrum, we use two different sizes proportional to the fundamental frequency $k_{\rm f} = \frac{(2\pi)^3}{V_{\rm s}}$. $V_{\rm s} = (3500\,{\rm Mpc}\,h^{-1})^3$ is the survey volume for the cubic box used to generate the mock catalogues. For the bispectrum, we then consider the two cases $\Delta k_{6,2} = 6,\, 2 \times k_{\rm f}$ corresponding to 116 and 2734 triangle configurations, respectively.

As done in the previous works (Gualdi et al. 2019a, b), we use $0.03 \leq k \leq 0.09\,h\,{\rm Mpc}^{-1}$ for the power spectrum terms and $0.02 \leq k \leq 0.12\,h\,{\rm Mpc}^{-1}$ for the bispectrum monopole. The bispectrum has a larger $k$-range because we adopted the effective kernel calibrated on simulations used also for the BOSS DR11 and DR12 analysis (Gil-Marín et al. 2015, 2017), which allows to safely extend the analysis to mildly non-linear scales (Gil-Marin et al. 2012).

All the MCMC samplings (both on original and compressed data vectors) have been run with the same settings as in previous works (Gualdi et al. 2019a,b).

Finally we use a flat Lambda cold dark matter cosmology to compute the linear matter power spectrum, with parameters close to the results from the Planck analysis (Akrami et al. 2019), in particular $\Omega_{\rm m}(z=0) = 0.31$, $\Omega_{\rm b}(z=0) = 0.049$, $A_{\rm s} = 2.21 \times 10^{-9}$, $n_{\rm s} = 0.9624$, $h_0 = 0.6711$, and $\sum m_\nu = 0.06$ eV.

[7] https://github.com/davidegua/max_geo_compression.git.

## 2 PREVIOUS COMPRESSING METHODS

### 2.1 Maximal compression

Maximal linear compression derives from the MOPED method (Heavens, Jimenez & Lahav 2000), which compresses the original data vector by extending to the multiple parameters case in the algorithm introduced by Tegmark, Taylor & Heavens (1997). As a result, an originally arbitrarily large data vector $\boldsymbol{x}$ can be transformed into a much shorter one $\boldsymbol{y}$ having a dimension equal to the number of model parameters considered in the analysis. This is achieved by taking the scalar product of $\boldsymbol{x}$ with a set of weights $\boldsymbol{b}_i$ for each of the model parameters $\theta_i$:

$$y_i \;=\; \langle \boldsymbol{x} \rangle_i^{\intercal} \cdot \mathbf{Cov}_{\boldsymbol{x}}^{-1} \cdot \boldsymbol{x} \;\equiv\; \boldsymbol{b}_i^{\intercal} \cdot \boldsymbol{x}, \tag{1}$$

where $\mathbf{Cov}$ is the covariance matrix for the original data vector $\boldsymbol{x}$, while $\langle \boldsymbol{x} \rangle_{,i}$ are the derivatives of the mean of the modelled data vector with respect to the model parameters $\theta_i$. The maximal compression method presented in Gualdi et al. (2018) that we consider in this paper consists in running an MCMC sampling on the compressed bispectrum data vector.

### 2.2 Geometrical compression

The main idea in Gualdi et al. (2019a) is to group together into new bins the triangles with similar geometrical properties. Each of these bins will correspond to an element of the compressed data vector. The value of the data vector's element is then given by the average bispectra of the triangle configurations belonging to the same bin.

The standard parametrization of each triangle configuration is given in terms of the three sides $(k_1, k_2, k_3)$. The new parameters are chosen using the physical intuition regarding which quantities most influence the bispectrum value. These are as follows:

(i) the square root of the triangle's area: ℵ ('aleph');
(ii) the cosine of the largest internal angle, ℸ = cos $\psi_{\rm max}$ ('daleth');
(iii) the ratio between the cosines of the intermediate and smallest angles, ג = cos $\psi_{\rm int}$/cos $\psi_{\rm min}$ ('gimel').

We rewrite the galaxy bispectrum monopole data vector as a function of the three new variables (ℵ, ℸ, ג):

$$B_{\rm g}^{(0)}(k_1, k_2, k_3) \implies B_{\rm g}^{(0)}(\aleph, \daleth, \gimel)\,. \tag{2}$$

By choosing large enough bins for the new variables, triangles with similar (ℵ, ℸ, ג) coordinates will be grouped together.

The new data vector $\boldsymbol{g}$ is then obtained by averaging over all the bispectra in the new bins defined by different sets of the coordinates (ℵ, ℸ, ג):

$$g_k(\aleph, \daleth, \gimel)_k = \frac{1}{N_k^{\rm tr.}} \sum_{j\,:\,(k_1,k_2,k_3)_j \in (\aleph, \daleth, \gimel)_k}^{N_k^{\rm tr.}} \mathrm{B}_{\rm g}^{(0)}(k_1, k_2, k_3)_j\,. \tag{3}$$

Each new data vector element has been normalized dividing by $N_k^{\rm tr.}$, the number of triangles belonging to the same bin obtained from a particular choice of $(\aleph, \daleth, \gimel)_k$.

## 3 ENHANCED GEOMETRICAL COMPRESSION

We combine geometrical and maximal compression, labelling this new method GEOMAX, in the following way. First, we regroup triangles as described in Section 2.2. Secondly, instead of averaging over all the bispectra of the triangle configurations belonging to each bin, we separately apply the maximal compression to each bin. Therefore, for every bin defined by a set of new coordinates $(\aleph, \daleth, \gimel)_k$, we will obtain, for each of the model parameters $\theta_i$, a compressed data vector element $g_{ik}^{\rm opt.}$ defined by

$$g_{ik}^{\rm opt.}(\aleph, \daleth, \gimel)_k = \boldsymbol{b}_{ik} \cdot \boldsymbol{B}_{{\rm g},k}^{(0)}\,. \tag{4}$$

$\boldsymbol{B}_{{\rm g},k}^{(0)}$ is the reduced data vector formed by the bispectra $B_{\rm g}^{(0)}(k_1, k_2, k_3)_j$ of the triangles belonging to the bin defined by $(\aleph, \daleth, \gimel)_k$, such that $j \;:\; (k_1, k_2, k_3)_j \in (\aleph, \daleth, \gimel)_k$. The weight vector $\boldsymbol{b}_{ik}$ for the $k$-bin, according to the definition given in equation (1), is given by

$$\boldsymbol{b}_{ik} \;=\; \left( \frac{\partial \langle \boldsymbol{B}_{{\rm g},k}^{(0)} \rangle}{\partial \theta_i} \right)^{\intercal} \cdot \mathbf{Cov}_{\boldsymbol{B}_{{\rm g},k}^{(0)}}^{-1}\,, \tag{5}$$

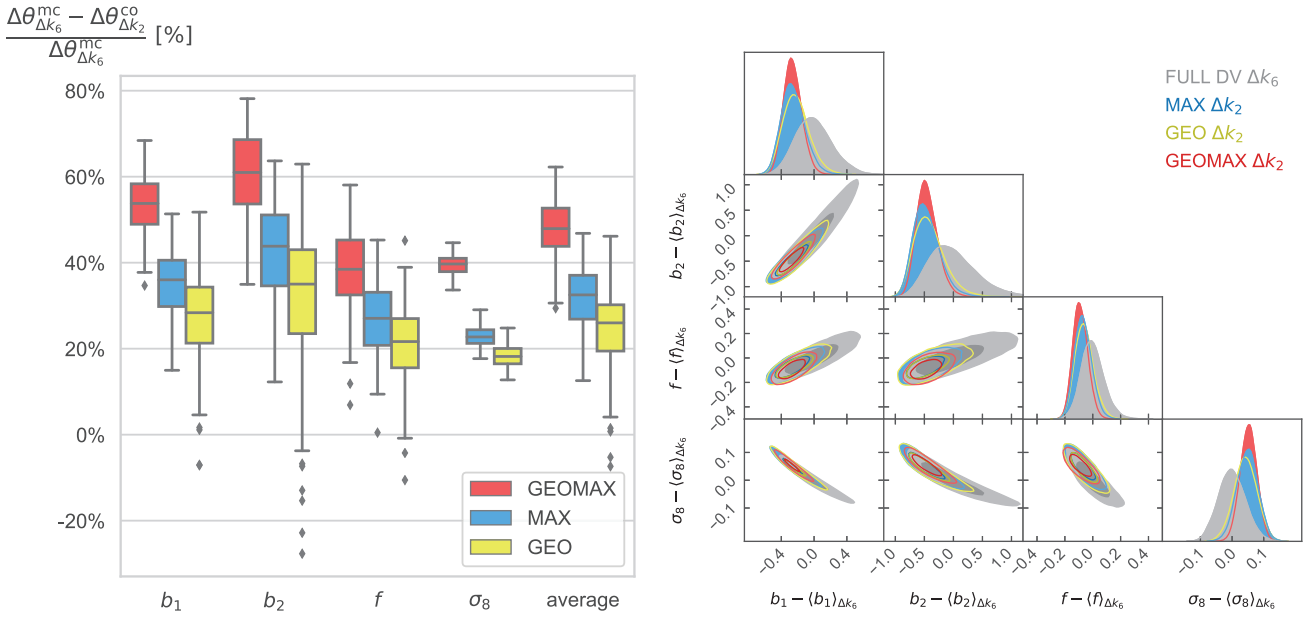
where $\mathbf{Cov}_{\boldsymbol{B}_{{\rm g},k}^{(0)}}^{-1}$ is the covariance matrix for the reduced data vector $\boldsymbol{B}_{{\rm g},k}^{(0)}$ of the bin $k$ computed using the available simulations or galaxy mock catalogues. In our case, we used 1400 realizations of the MultiDark Patchy BOSS DR12 mocks. We adopt the conservative approach of ensuring that the number of triangle configurations belonging to each new bin is less than half the number of available mocks. This is to reduce to a reasonable level the bias induced by estimating the covariance matrix from a limited number of realizations (Hartlap et al. 2007). In any case, this bias would be a constant factor and therefore not affecting the compression weights. We can then assume that $\mathbf{Cov}_{\boldsymbol{B}_{{\rm g},k}^{(0)}}^{-1}$ is a good approximation of the covariance matrix of the reduced data vector bin $\boldsymbol{B}_{{\rm g},k}^{(0)}$. Therefore, in order to estimate the covariance matrix needed to maximally compress the largest bin (which includes ∼500 triangle configurations), we need at least ∼1000 mock catalagoues. This requirement is more than five times lower than the number of mocks needed when the full data vector (2734 triangle configurations) is considered (≃ 5500 simulations).

In the algorithm selecting the best bin number settings, it is possible to impose a maximum number of triangle configurations per bin. In this way, one can arbitrarily reduce the number of simulations required to estimate the covariance matrix for each bin. However, excessively reducing the number of maximum triangles per bin would also decrease the performance of the compression. In the same way, increasing the maximum number of triangle configurations per bin up to the number of simulations available would decrease the overall performance. This is because the covariance matrices used for the maximal compression step would no longer be accurate enough, hence reducing the efficacy of the compression weights.

### 3.1 Updated optimal binning choice selection criteria

In order to choose the number of bins for the new data vector, or in other words, the triplets of $(\aleph, \daleth, \gimel)_k$ defining the elements of the new data vector, in Gualdi et al. (2019a), we estimated the sensitivity of the potential compressed data vector with respect to the considered cosmological parameters. This was done by computing a summary statistic for each possible choice of bins numbers. This number is obtained by averaging together the derivatives of the compressed data vector $\boldsymbol{g}$. For the geometrical compression, it was fundamental to divide each compressed data vector element's derivative by the number of triangle configurations corresponding to the bin $k$ before averaging

$$S_{ij} \equiv \sum_{k=1}^{N_{\boldsymbol{g}}(n_\aleph, n_\daleth, n_\gimel)_j} \frac{1}{N_k^{\rm tr.}} \left| \frac{\partial g_k}{\partial \theta_i} \right|, \tag{6}$$




(a) Mocks (b) Data

**Figure 1.** Joint data vector $\left[P_g^{(0)}, P_g^{(2)}, B_g^{(0)}\right]$ posteriors: four parameters case. Panel (a): The above boxplot summarizes the statistical distribution of the parameter constraints improvements for the new geometric-maximal method (GEOMAX) and the two individual maximal (MAX) and geometrical (GEO) ones. The coloured boxes show the central quartiles of each distribution while the individual dots are automatically considered outliers by the plotting routine. Each boxplot is obtained comparing the different methods (compressing 2734 triangles, $\Delta k_2$ case) with the MCMC on the full data vector (116 triangles, $\Delta k_6$ case) for 100 realizations of the Patchy Mocks (Kitaura et al. 2016). The last column showing the average improvement has the purpose to explain the reason for which certain mocks have some of the parameters with negative improvements (in particular for the GEO compression). The always positive average improvement (beside for two mocks out of 100) shows that the individual negative ones are a statistical fluctuation due the above average improvements for the other parameters. Panel (b): compression performance; 2D 68 per cent and 95 per cent credible regions are shown for the standard MCMC sampling on the full data vector using the triangles from the $\Delta k_6$ case (116 triangles, MCMC). For the $\Delta k_2$ case (2734 triangles) are shown the contours obtained by compressing the bispectrum part of the data vector using maximal (Gualdi et al. 2018, 2019b, MAX), geometrical (Gualdi et al. 2019a, GEO), and enhanced geometrical compression methods (GEOMAX). The combination of geometrical and maximal compressing methods (GEOMAX) further improves the parameter constraints as quantitatively described in Table 1. These marginalized posterior distributions have been derived using measurements from BOSS DR12 CMASS data. We subtracted to all the distributions the central value obtained for the $\Delta k_6$ standard MCMC case. This because our goal is to test whether we would observe on data, which usually include unknown systematics, the same improvements statistically observed on galaxy mock catalogues.

where $S_{ij}$ is an estimator of the sensitivity of $\boldsymbol{g}$ when varying the model parameter $\theta_i$, defined for a particular choice of number of bins $(n_\aleph, n_\beth, n_\gimel)_j$.

For GEOMAX, this no longer works since the new data vector elements are derived from a linear combination of the original bispectra, where the weights are given by the maximal compression applied to each bin as shown in equation (5). Therefore, we need to define a new summary statistic that normalizes each compressed data vector's element derivative dividing by the sum of the weights used to compute it:

$$S_{ij}^{\rm gm} \equiv \sum_{k=1}^{N_g(n_\aleph,n_\beth,n_\gimel)_j} \sum_{\ell=1}^{N_{\rm par}} \left(\sum_{m=1}^{N_k^{\rm tr.}} b_{\ell k}^m\right)^{-1} \frac{\partial g_{\ell k}}{\partial\theta_i} = \sum_{k=1}^{N_g(n_\aleph,n_\beth,n_\gimel)_j} \sum_{\ell=1}^{N_{\rm par}} \left(\sum_{m=1}^{N_k^{\rm tr.}} b_{\ell k}^m\right)^{-1} \boldsymbol{b}_{\ell k} \cdot \frac{\partial \boldsymbol{B}_k^{(0)}}{\partial\theta_i}, \quad (7)$$

where the sum over $k$ accounts for all the elements of the compressed data vector. The sum over $\ell$ covers the number of linear combinations (equal to the number of model parameters) obtained from each triangle configurations group defined by a set $(\aleph, \beth, \gimel)_k$. The self-contained sum over $m$ inside the curved brackets acts as a normalization factor specific for each of the compressed data vector derivative's elements.

We can then proceed as in the case of the geometrical compression where a single number can be obtained by

$$\bar{s}_j^{\rm gm} \equiv \sum_{i=1}^{N_\theta} s_{ij}^{\rm gm} = \sum_{i=1}^{N_\theta} \frac{S_{ij}^{\rm gm}}{\max\left[S_{ij}^{\rm gm}\right]_{\forall j}}, \quad (8)$$

Again, we choose the set of $(n_\aleph, n_\beth, n_\gimel)_j$ that maximizes $\bar{s}_j^{\rm gm}$.

### 3.2 Code structure

In order to derive the compressing function, three ingredients are required:

(i) the triangle configurations in terms of the sides length $(k_1, k_2, k_3)$;
(ii) the derivatives of the 3pt data vector with respect to the model parameters;
(iii) measurements of the 3pt data vector from the available simulations.

Moreover, one can set the maximum number of triangles per bin and the range to check for the number of bins for each of the geometrical parameters.

The code[8] consists of creating an object with several functions needed to find the optimal compression as well as to convert the analytical model and measurements of the data vector into their compressed form. First of all, the code computes the optimal binning in terms of geometrical compression, using the updated selection criteria presented in the previous section. Afterwards, bins with less triangles than the number of model parameters are merged together into a single bin. Finally, the weights to maximally compress each group of triangles are obtained using equation (1) where the covariance matrix is estimated using the available simulations (whose number has to be larger than the number of triangles in each group).

The created object contains all the information required to convert the data vectors and the measurements into their compressed form. This can be done by using the object's methods. For example, in our analysis pipeline, this is done at every step of MCMC sampling: We first compute the full data vector that is subsequently compressed (together with the covariance matrix) before the likelihood evaluation.

## 4 RESULT ANALYSIS

We compare the enhanced geometrical compression (GEOMAX) with the previous works including the MCMC on the full data vector obtained by considering the same $k$-bin size ($\Delta k_6$), as in the BOSS analysis (Gil-Marín et al. 2015, 2017), which, for our $k$-range choice, corresponds to 116 triangle configurations. For all the compression methods, we use 2734 triangle configurations obtained by choosing a three times smaller $k$-bin size ($\Delta k_2$). We cannot use the full data vector with 2734 triangle configurations since there are not enough galaxy mock catalogues available to numerically estimate an invertible covariance matrix (Hartlap et al. 2007). Maximal compression (MAX) returns a compressed data vector with a number of elements equal to the number of model parameters. For the geometrical (GEO) and enhanced geometrical (GEOMAX) compression methods we impose, for a fair comparison with standard MCMC, a maximum number of bins equal to the dimension of the full data vector (116 triangles) plus one. The two algorithms return a compressed data vector with a dimension equal to 116 (GEO) and 115 (GEOMAX), respectively .

### 4.1 Statistical significance from galaxy mocks

We validate our method by studying the distribution of the improvements, over the 1D 68 per cent credible regions on the parameter constraints, by both running MCMC samplings on the full data vector ($\Delta k_6$) and on the compressed one for MAX, GEO, and GEOMAX methods ($\Delta k_2$) on 100 realizations of the Patchy Mocks.

In the left-hand panel of Fig. 1, we can appreciate how GEOMAX outperforms, considering the means of the distributions, the MAX and GEO methods by obtaining approximately, on average, 15 per cent tighter constraints. With respect to the standard MCMC, the 1D 68 per cent credible regions are 40–60 per cent tighter when using GEOMAX. Moreover, the improvement value scatter for GEOMAX is, on average, smaller than the ones for both MAX and GEO methods.

The column showing the average in the left-hand panel of Fig. 1 explains why for certain mocks some parameters have negative improvements. Since the average is always positive, we can explain the individual negative improvements as a statistical fluctuation due to the other parameters above average improvements. In Appendix C, we test for one of these mocks that the negative improvements for some parameters are not due to the choice of fiducial cosmology used to derive the compression.

[8] https://github.com/davidegua/max_geo_compression.git.

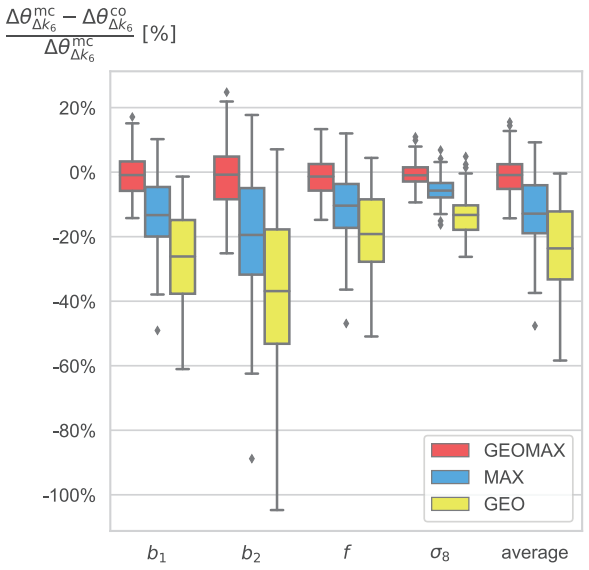



**Figure 2.** Loss of information test: Using the bispectra values of the 116 triangle configurations corresponding to the $\Delta k_6$ binning case, we check whether the compressed data vectors retain the same information of the original full one. We show for the three different methods, enhanced geometrical c. (GEOMAX), maximal c. (MAX), and geometrical c. (GEO), the ratio between the width of the 1D 68 per cent credible regions with the ones obtained running an MCMC sampling on the full data vector. The distributions for the different parameters are obtained by computing the ratio for 100 galaxy catalogue measurements. From the above boxplots, we see that only GEOMAX is statistically consistent with zero loss of information.

For the different methods, we also test the loss of information associated with the compression of the data vector. In Fig. 2, we show once more the ratio of the 1D per cent credible regions; however, in this case, we compress the same 116 triangle configurations long data vector used for the standard MCMC ($\Delta k_6$). For GEOMAX and GEO methods, we set a maximum number of elements for the compressed data vector equal to 60. We chose this threshold since for the GEOMAX method, it corresponds (in the case of four parameters) to a maximum number of bins for the geometrical step equal to 15. Fig. 2 highlights that, even in this 'few-bins' possible scenario for the geometrical step, GEOMAX compression is statistically consistent with zero loss of information with respect to the MCMC on the full data vector. The geometrical compression suffers more from the few bins available, since, in each bin, the bispectra values are averaged, whilst in GEOMAX, they are weighted by the coefficients given by the maximal compression step given by equation (5).

### 4.2 Test on BOSS DR12 CMASS data

We apply our method on data to test whether we find a performance similar to what statistically observed using the measurements from the galaxy mocks. Qualitative results for the data can be observed in the right-hand panel of Fig. 1. Since we are mainly interested in the improvements on the parameters constraints, we present our results with the central value of each parameter obtained through standard MCMC sampling subtracted. We will perform a full parameter constraints analysis, including also Finger of God and Alcock–

**Table 1.** The improvements in parameter constraints shown are the relative change of the 68 per cent credible intervals for the $\Delta k_2$ $k$-binning case with respect to the $\Delta k_6$.

| | $\Delta\theta^{\rm mc}_{\Delta k_6}$ | $\frac{\Delta\theta^{\rm mc}_{\Delta k_6} - \Delta\theta^{\rm comp.}_{\Delta k_2}}{\Delta\theta^{\rm mc}_{\Delta k_6}}$ [%] | | |
|---|---|---|---|---|
| | MCMC $N_{\rm tr} = 116$ | MAX $N_{\rm el.} = 4$ | GEO $N_{\rm g} = 116$ | GEOMAX $N_{\rm g} = 116$ |
| $\Delta b_1$ | 0.22 | 36.1 | 28.6 | 50.4 |
| $\Delta b_2$ | 0.40 | 45.5 | 36.5 | 56.1 |
| $\Delta f$ | 0.08 | 23.9 | 18.5 | 33.2 |
| $\Delta\sigma_8$ | 0.04 | 21.6 | 15.1 | 38.3 |
| $\left\langle \frac{\Delta\theta^{\rm mc}_{\Delta k_6} - \Delta\theta^{\rm comp.}_{\Delta k_2}}{\Delta\theta^{\rm mc}_{\Delta k_6}} \, [\%] \right\rangle$ | | 31.8 | 24.7 | 44.5 |

*Note*. While maximal and geometrical methods achieve similar results, their combination, GEOMAX, on average performs better by ∼15 per cent when applied to BOSS DR12 CMASS data.

Paczynski effects, in a following paper, extending the modelling in order to include smaller scales.

In Table 1, we can see the improvements obtained by the compression methods for each of the parameter constraints derived using the MCMC on the full data vector, maximal compression (MAX), geometrical compression (GEO), and enhanced geometrical compression (GEOMAX).

For what concerns the model parameter 1D 68 per cent credible intervals, GEOMAX performs on data, on average, 45.5 per cent better than standard MCMC and improves the results obtained by maximal and geometrical compression by approximately ∼15 per cent. This shows very good agreement with the improvements observed in the case of the galaxy catalogues shown in the left-hand panel of Fig. 1.

We find that maximal compression is suboptimal compared to the enhanced geometrical one. We speculate that the main reason behind this limitation is the linear limit implicit in the compression scheme we developed in Gualdi et al. (2018, 2019a). With a non-linearly degenerate parameter space, even if these linear compression techniques achieve an improvement of the parameter constraints by allowing the employment of larger data vectors, they still miss part of the available information. That could be the reason why combining the maximal compression with a complementary approach, such as the geometrical one, produces an effective non-linear compression that returns tighter parameters constraints once applied to an originally longer (but not redundant) data vector.

This effective non-linearity feature is achieved by adopting a motivated criteria that defines which triangles are linearly combined together. The non-linearity indeed lies in the method (equation 8), used to define the triangle bins. In other words, grouping the bispectra is a non-linear operation in the cosmological parameters.

With respect to the original data vector, GEOMAX achieves a similar compression factor to the geometrical compression ∼23, but a much lower one than the maximal one (∼683). From the results displayed in Fig. 1, we conclude that the reduction of the compressing factor is a fair price for the increased constraining power obtained by exploiting the physical insight granted by the geometrical compression step.

## 5 ALTERNATIVES TO THE GEOMETRICAL COMPRESSION STEP

We show in Fig. 3 two different ways to define the triangles sets, before applying the maximal compression in order to obtain the final data vector. Since the full bispectrum for the standard MCMC $\Delta k_6$ case has 116 triangles, we group the 2734 triangles of the $\Delta k_2$ case into 29 sets. Maximal compression is then applied on each of these 29 groups. In this way, since we consider four parameters, the final data vector will also have 116 elements.

This is the largest allowed dimension in order to fairly compare the compressed data vector (given by the $\Delta k_2$ triangles) with the original full one (given by the $\Delta k_6$ triangles).

### 5.1 Random regrouping

The most naive way to group together the triangles before the maximal compression step is to randomly distribute them into $N$-groups (as equally populated as possible). This immediately raises the concern that different random allocations of the triangles into $N$-groups can, in principle, produce wider/tighter posterior distributions.

The performance cannot be predicted in advance unless it is applied an a priori criteria to choose whether a random allocation is optimal or not. Even if such a criteria could be devised, the random choice factor would make its application very inefficient. In contrast, the geometrical step possesses a precise criteria to define the optimal way to group triangles together and also in how many bins.

From the marginalized contours in Fig. 3, we can immediately deduce that the geometrical compression algorithm outperforms this alternative.

### 5.2 Reference triangles

A more sophisticated approach consists of defining 'reference' triangles and to assign each of the $\Delta k_2$ case 2734 triangles to the bin defined by the most similar 'reference' triangle. We then assign a triangle $a$ characterized by the sides $(k^a_1, k^a_2, k^a_3)$ to the bin $j$ having as reference triangle the configuration $(p^b_1, p^b_2, p^b_3)$, such that

$$\sum_{i=1,2,3} \frac{\left|k^a_i - p^b_i\right|}{k^a_i} < \sum_{i=1,2,3} \frac{\left|k^a_i - p^\ell_i\right|}{k^a_i} \;\forall\; \ell \neq b\,. \tag{9}$$

The similarity criteria is then simply the minimal sum of the normalized absolute difference between the triangles sides. The reference triangles are chosen among the original 2734 ones. Considering the algorithm that generated them, the selection is done by taking, in terms of the position in the array, equidistant configurations in the data vector (one every 95 configurations in this particular case).

Fig. 3 shows that also this slightly more sophisticated criteria does not reach the improvements achieved by the enhanced geometrical compression. As in the case of the 'random regrouping' approach, the marginalized posterior distributions are not significantly tighter than the one given by just using maximal compression.

## 6 ALTERNATIVE PARAMETER SETS

In order to check that the improvement achieved by the enhanced geometrical compression is not dependent on the chosen parameter set, we use GEOMAX to derive the constraints for two additional parameter sets from the same galaxy catalogues and data.

### 6.1 Local primordial non-Gaussianity

In order to distinguish between different models of inflation, one of the key observables is the deviation from a Gaussian distribution of the primordial density fluctuations (Bartolo et al. 2004). In the case of

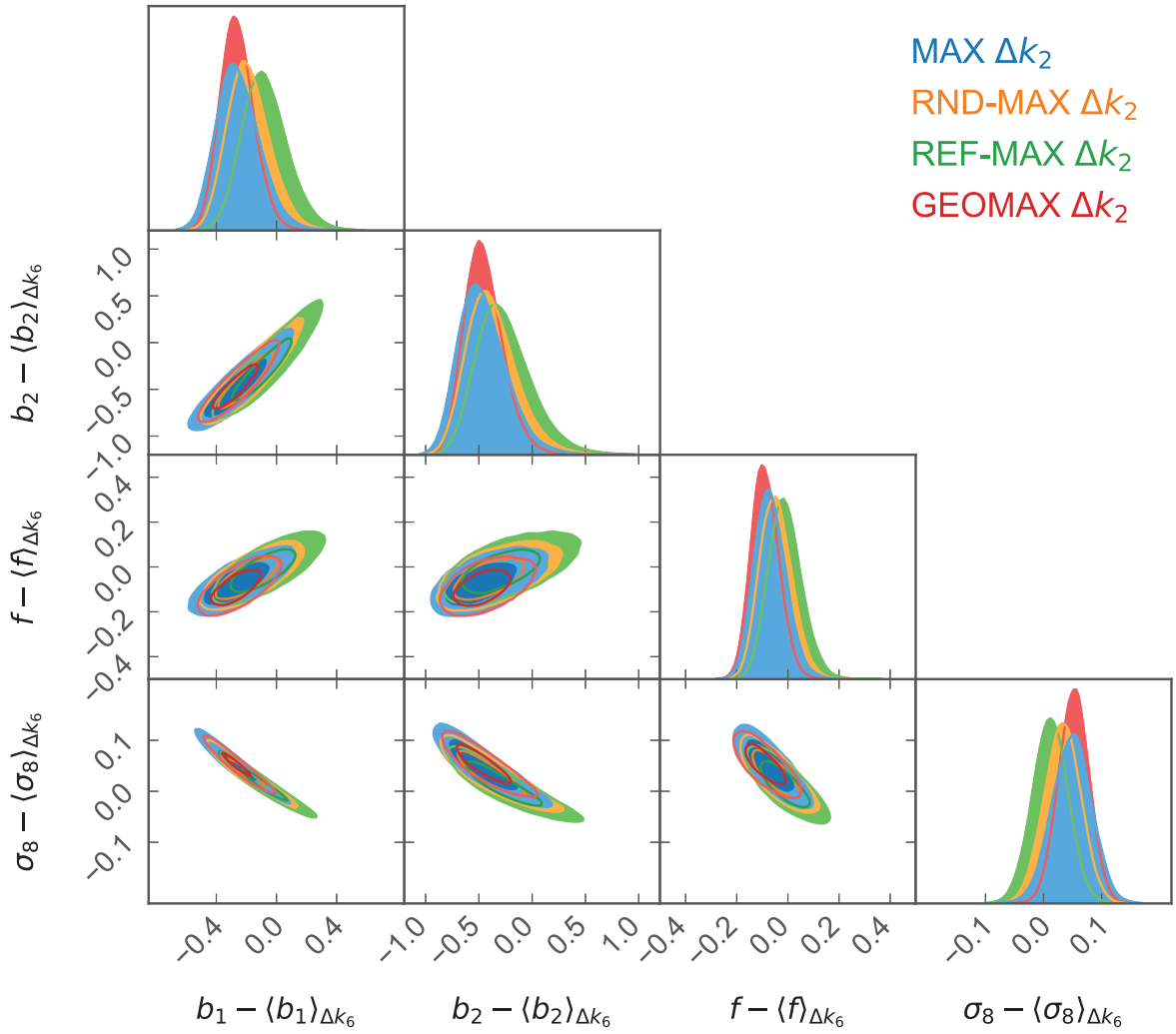



**Figure 3.** Compression performance of the alternatives to the geometrical compression step: 2D 68 and 95 per cent credible regions are shown for data vectors obtained by grouping together the 2734 $\Delta k_2$ triangles in the two alternative ways described in Section 5. RND-MAX corresponds to the random assignation (triangles equally distributed) to 29 bins before the maximal compression. For REF-MAX, the triangles are associated with the bin whose 'reference' triangle is closer in terms of perimeter. Comparing the contours of RND-MAX and REF-MAX with the ones for GEOMAX, we can see the importance of the geometrical compression step. Without it, there is no significant improvement with respect maximal compression (MAX) alone. The different 1D marginalized posterior distributions shifts with respect to the central values obtained by the MCMC sampling are due to the different reduction of the parameter space degeneracy achieved by the methods shown. The smaller is the improvement with respect to the MCMC constraints, the smaller is the shift. This was previously discussed in Gualdi et al. (2019a,b).

local primordial non-Gaussianity, this deviation can be parametrized through an expansion of the Bardeen's gravitational potential $\Phi$ (Bardeen 1980) in terms of a Gaussian field $\phi$ and a costant $f_{\rm NL}$ acting as the amplitude of the deviation from linearity at first order:

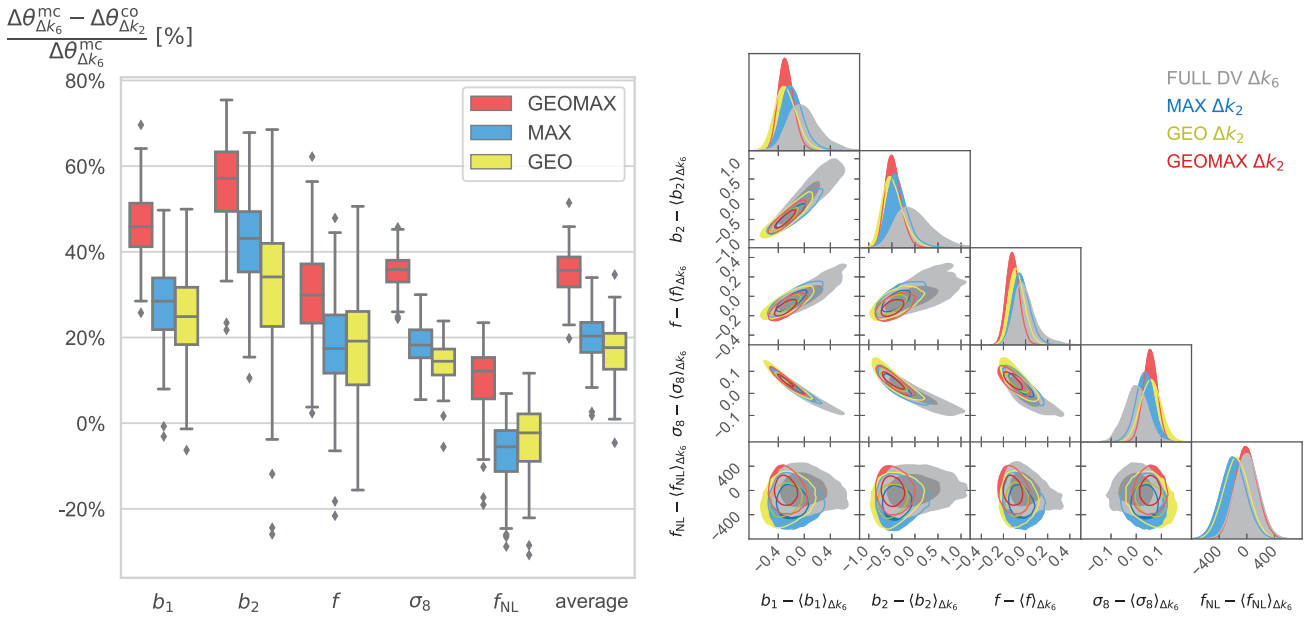
$$\Phi = \phi + \frac{f_{\rm NL}^2}{c^2}\left[\phi^2 - \langle\phi^2\rangle\right] + \ldots \tag{10}$$

In our analysis, we only consider a local type of primordial non-Gaussianity (see Byrnes & Choi 2010 for a review). There are also other types of primordial non-Gaussianities (for simulations on these, see Scoccimarro et al. 2012).

Late-time large-scale-structure analyses have the potential of reaching, with the next generation of surveys, constraints on $f_{\rm NL}$ of similar order to the ones obtained by Planck (Akrami et al. 2019). The modelling and constraints forecasts for the matter and galaxy bispectrum have already been widely studied in the literature (Verde et al. 2000; Jeong & Komatsu 2009; Sefusatti 2009). We proceed as in Scoccimarro et al. (2004) to derive the correction for the galaxy power spectrum for a local primordial non-Gaussianity:

$$\begin{aligned} P^{\rm g}_{\rm NG}(\boldsymbol{k}) &= P_{11} + P_{12} \approx P_{12} \\ &= \frac{4 f_{\rm NL}}{c^2} F_k^{(1)} \beta^2 k^4 T_k^2 \int \frac{{\rm d}\boldsymbol{p}_a^3}{(2\pi)^3} F^{(2)}_{a|\boldsymbol{k}+\boldsymbol{p}_a|} P^{\phi}_{|\boldsymbol{k}+\boldsymbol{p}_a|} \left[P^{\phi}_a + 2P^{\phi}_k\right] . \end{aligned} \tag{11}$$

We only use $P_{12}$ ($\propto f_{\rm NL}/c^2$) since $P_{11}$ is expected to be negligible given that is proportional to $f_{\rm NL}^2/c^4$. In the above expression, $F_k^{(1)}$ and $F^{(2)}_{a|\boldsymbol{k}+\boldsymbol{p}_a|}$ are the standard first- and second-order perturbation theory kernels in redshift space (known also as $Z$ in the literature). $T_k$ is the matter transfer function normalized to one for $k \to 0$. $P^{\phi}$ is the primordial power spectrum for the Gaussian part of the Bardeen's potential $\phi$. $\beta = \frac{3}{5} D_1(z)/\left(\Omega_{\rm m} H_0\right)$, where $D_1$ is the linear growth factor at redshift $z$, while $\Omega_{\rm m}$ and $H_0$ are the matter density




(a) Mocks (b) Data

**Figure 4.** Joint data vector $\left[P_g^{(0)}, P_g^{(2)}, B_g^{(0)}\right]$ posteriors: five-parameter case including local primordial non-Gaussianity. Same as Fig. 1 when considering the additional parameter $f_{NL}$.

**Table 2.** Same as Table 1 for the first additional parameter set test case.

| | $\Delta\theta^{mc}_{\Delta k_6}$ | $\frac{\Delta\theta^{mc}_{\Delta k_6} - \Delta\theta^{comp.}_{\Delta k_2}}{\Delta\theta^{mc}_{\Delta k_6}}$ [%] | | |
|---|---|---|---|---|
| | MCMC $N_{tr} = 116$ | MAX $N_{el.} = 4$ | GEO $N_g = 116$ | GEOMAX $N_g = 115$ |
| $\Delta b_1$ | 0.22 | 36.1 | 28.6 | 50.4 |
| $\Delta b_2$ | 0.40 | 45.5 | 36.5 | 56.1 |
| $\Delta f$ | 0.08 | 23.9 | 18.5 | 33.2 |
| $\Delta\sigma_8$ | 0.04 | 21.6 | 15.1 | 38.3 |
| $\Delta f_{NL}$ | 171.5 | − 5.8 | − 6.1 | 6.8 |
| | | 21.1 | 19.9 | 38.8 |
| $\left\langle \frac{\Delta\theta^{mc}_{\Delta k_6} - \Delta\theta^{comp.}_{\Delta k_2}}{\Delta\theta^{mc}_{\Delta k_6}} \text{ [%]} \right\rangle$ | | | | |

parameter and the Hubble constant, respectively. We fixed $H_0$ to the fiducial value used to compute the linear matter power spectrum.

For the bispectrum, we have the additional term

$$B^g_{NG}(\boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{k}_3) = B^g_{111} = F^{(1)}_{k_1} F^{(1)}_{k_2} F^{(1)}_{k_3} \beta^{-1} k_1^2 k_2^{-2} k_3^{-2} \frac{T_{k_1}}{T_{k_2} T_{k_3}} \frac{2 f_{NL}}{c^2} P^m_{k_2} P^m_{k_3} + \text{cyc.}, \quad (12)$$

where $P^m$ is the linear matter power spectrum. The derivation of both power spectrum and bispectrum primordial non-Gaussianity corrections is described in Appendix B. The relevance of the primordial non-Gaussianity terms with respect to the gravitational ones is shown in Fig. B1 for the power spectrum and Fig. B2 for the bispectrum.

Fig. 4 displays the results relative to the addition of the $f_{NL}$ parameter to the analysis for both mocks and data measurements. While MAX and GEO compression return larger posterior distributions for $f_{NL}$ than the standard MCMC on the full data vector with less triangles, GEOMAX returns 1D 68 per cent credible regions tighter

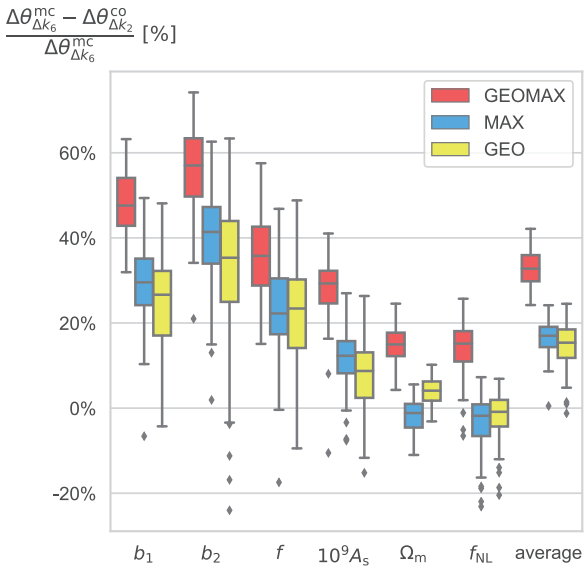



**Figure 5.** Same as the right-hand panel of Fig. 1 for the six parameters case. Also, in this case, the improvement of GEOMAX with respect to MAX and GEO compression methods is statistically significant.

than the MCMC ones by ∼10 per cent, on average, on mocks (left-hand panel Fig. 4) and 6.8 per cent for data (right-hand panel of Fig. 4 and Table 2), respectively.

### 6.2 Cosmological parameters

In the second alternative parameter set, we substitute $\sigma_8$ with the amplitude of scalar perturbations, $A_s$ , and the matter density parameter, $\Omega_m$. The results for this case are displayed in Figs 5 and 6.

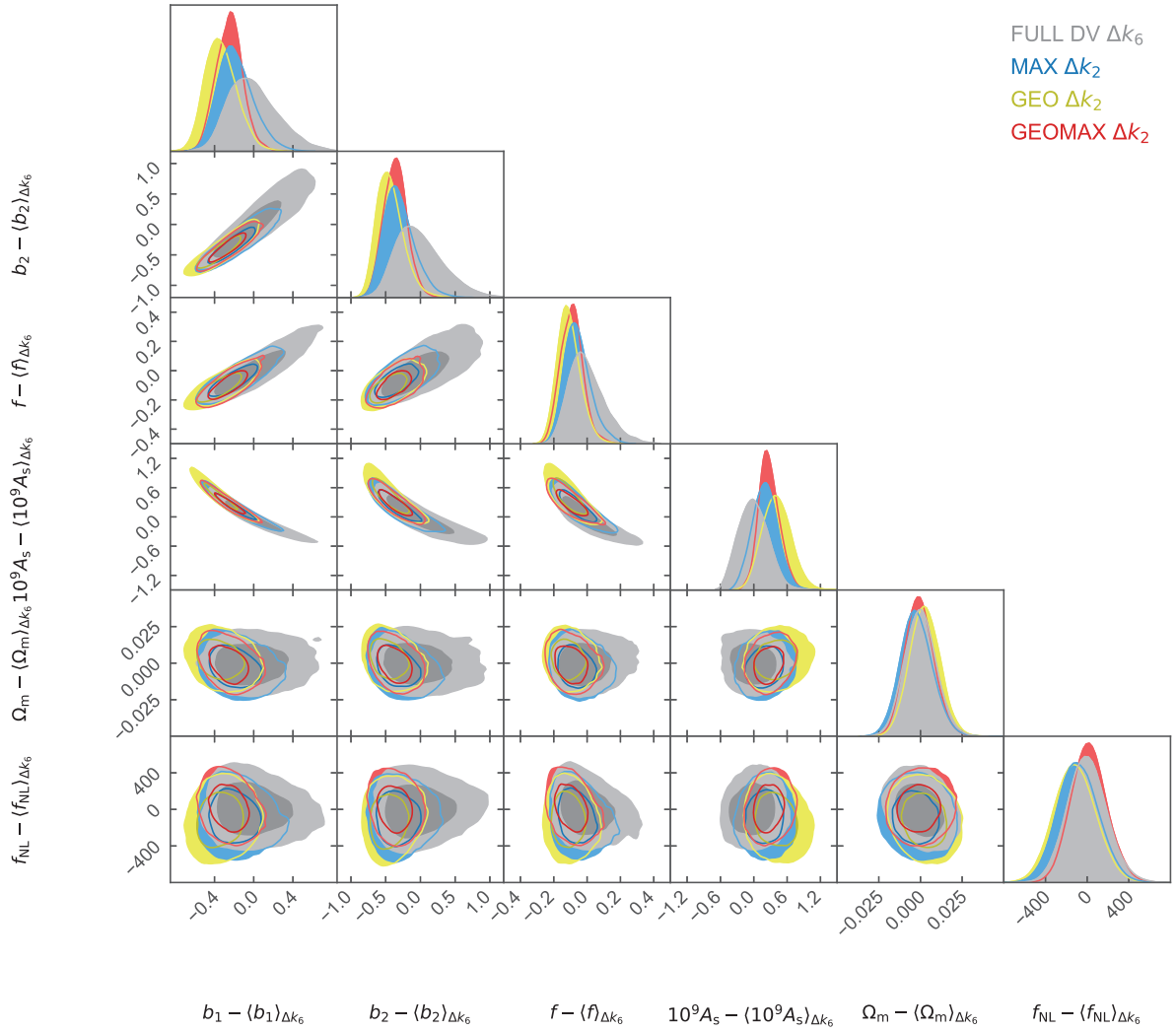



**Figure 6.** Same as the left-hand panel of Fig. 1 for the six parameters case. The enhanced geometrical compression still outperforms the individual maximal and geometrical ones. In particular, it also improves the constraints for those parameters where the other methods performs as well as the MCMC on the full data vector with less triangles.

**Table 3.** Same as Table 1 for the second additional parameter set test case.

| | $\Delta\theta^{\rm mc}_{\Delta k_6}$ | $\frac{\Delta\theta^{\rm mc}_{\Delta k_6} - \Delta\theta^{\rm comp.}_{\Delta k_2}}{\Delta\theta^{\rm mc}_{\Delta k_6}}$ [%] | | |
|---|---|---|---|---|
| | MCMC $N_{\rm tr} = 116$ | MAX $N_{\rm el.} = 4$ | GEO $N_{\rm g} = 116$ | GEOMAX $N_{\rm g} = 114$ |
| $\Delta b_1$ | 0.25 | 30.2 | 36.8 | 47.8 |
| $\Delta b_2$ | 0.33 | 36.0 | 44.7 | 50.9 |
| $\Delta f$ | 0.11 | 27.0 | 35.9 | 37.7 |
| $10^9 \Delta A_s$ | 0.27 | 15.0 | 4.7 | 32.0 |
| $\Delta\Omega_m$ | 0.01 | − 0.8 | 3.6 | 8.8 |
| $\Delta f_{NL}$ | 183 | − 4.8 | − 6.8 | 7.9 |
| | | 17.1 | 19.8 | 30.8 |
| $\left\langle \frac{\Delta\theta^{\rm mc}_{\Delta k_6} - \Delta\theta^{\rm comp.}_{\Delta k_2}}{\Delta\theta^{\rm mc}_{\Delta k_6}} \text{ [%]} \right\rangle$ | | | | |

Once more, GEOMAX outperforms both MAX and GEO methods over all the parameters, improving, on average, by ∼30 per cent the constraints given by standard MCMC on the full data vector (Table 3). In particular, this case again shows that the enhanced compression is able to obtain noticeable improvements for those parameters ($\Omega_m$, $f_{NL}$) where maximal and geometrical compression do not surpass the results of standard MCMC sampling.

## 7 CONCLUSIONS

We introduced an optimized compression method for the galaxy bispectrum and made the code publicly available.[9] This enhanced geometrical compression (GEOMAX) can, in principle, be easily applied to any 3pt statistics in cosmology.

The first requirement is a sufficient large number of simulations. These however are normally far fewer than the ones that would be needed to estimate the covariance matrix for a 3pt statistics data vector without compression. The second input are the derivatives of the data vector model with respect to the model parameters that one wishes to constrain.

[9] https://github.com/davidegua/max_geo_compression.git.

The geometrical compression (GEO; Gualdi et al. 2019a) splits the triangles into groups based on similarity of their geometrical properties. This allows us to use the available simulations to estimate the covariance matrix for each of these groups, which can then be used to maximally compress (MAX, Gualdi et al. 2018, 2019b) the bispectrum in each of them.

We tested the GEOMAX algorithm on both a set of galaxy bispectrum monopole measurements from 100 Patchy mocks (Kitaura et al. 2016) and the measurement from BOSS DR12 CMASS sample (Gil-Marín et al. 2017). Through the galaxy catalogues, we studied the statistical significance of the improvements on parameter constraints (left-hand panel in Figs 1, 4, and 5). In Fig. 2, we also show that GEOMAX is statistically consistent with being 'information-lossless' with respect to the MCMC on the full data vector (while GEO and MAX methods are not). It would be interesting to compare this method's performance with other information lossless compression algorithms (Alsing & Wandelt 2018; Charnock, Lavaux & Wandelt 2018), which are usually directly applied to data through likelihood-free inference analyses (Alsing, Wandelt & Feeney 2018; Alsing et al. 2019).

We used the DR12 CMASS data measurements to check that known systematics (for example the choice of a fiducial cosmology for the analysis) and unknown ones did not affect the compression results. With respect to the standard MCMC on the full data vector, the enhanced geometrical compression returns 1D 68 per cent credible regions tighter by a factor of 50.4, 56.1, 33.2, and 38.3 per cent for the parameters $(b_1, b_2, f, \sigma_8)$. With respect to the individual maximal and geometrical compression methods, the constraints are ∼15 per cent smaller (see Fig. 1 and Table 1 for details).

Two alternative pre-maximal compression steps have been considered in order to test the importance of the geometrical method. These alternatives do not produce the same improvements as the geometrical method when combined with the maximal compression (Fig. 3). Moreover, they do not show significant differences from maximal compression alone.

To strengthen our case, we also run the analysis for two larger parameter sets, proving that the benefits of this new method are not parameter-set dependent (Figs 4–6 and Tables 2 and 3). In particular, GEOMAX improves the MCMC 1D 68 per cent credible regions also for those parameters where, instead, MAX and GEO methods return larger marginalized 1D posterior distribution. For these two additional parameter sets, the average improvement observed on data of GEOMAX with respect to MAX and GEO methods varies between 10 and 20 per cent.

In order to maximize the extraction of cosmological information from 3pt statistics, we conclude with the expectation that this flexible method will be employed for the analysis of the forthcoming cosmological data sets such as DESI, *Euclid*, PFS, and SKA.

## ACKNOWLEDGEMENTS

DG thanks Prof. Alan Heavens, Prof. Andrew Pontzen, and Prof. Licia Verde for the useful discussions. DG is also grateful to Pérez Forcadell Gabriel for the help in using the ICCC-UB computer cluster. DG acknowledges support from European Union's Horizon 2020 research and innovation programme ERC (BePreSySe, grant agreement 725327). HGM acknowledges the support from la Caixa Foundation (ID 100010434) with code LCF/BQ/PI18/11630024. MM acknowledges support from the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie grant agreement No. 6655919y.

The linear matter power spectrum has been computed using the CLASS code (Lesgourgues 2011). C (Kernighan 1988) and PYTHON 2.7 (Rossum 1995) have been used together with many packages like IPYTHONS (Perez & Granger 2007), NUMPY (van der Walt, Colbert & Varoquaux 2011), SCIPY (Jones, Oliphant & Peterson 2001), and MATPLOTLIB (Hunter 2007). The corner plots have been realized using PYGTC developed by Bocquet & Carter (2016). We used EMCEE (Foreman-Mackey et al. 2013) as MCMC sampler.

## DATA AVAILABILITY

The data underlying this paper are available at https://data.sdss.org/sas/dr12/boss/lss/ from the public domain source at https://www.sdss.org/dr12/.

## REFERENCES

Abell P. A. et al., 2009, preprint (arXiv:e-prints)
Aghanim N. et al., 2018, preprint (arXiv:1807.06209)
Akrami Y. et al., 2019, preprint (arXiv:1905.05697)
Alsing J., Wandelt B., 2018, MNRAS, 476, L60
Alsing J., Wandelt B., Feeney S., 2018, MNRAS, 477, 2874
Alsing J., Charnock T., Feeney S., Wandelt B., 2019, MNRAS, 488, 4440
Bacon D. J. et al., 2018, Publ. Astron. Soc. Aust., 37, e007
Bardeen J. M., 1980, Phys. Rev. D, 22, 1882
Barreira A., 2019, J. Cosmol. Astropart. Phys., 1903, 008
Bartolo N., Komatsu E., Matarrese S., Riotto A., 2004, Phys. Rep., 402, 103
Beane A., Lidz A., 2018, ApJ, 867, 26
Bertacca D., Raccanelli A., Bartolo N., Liguori M., Matarrese S., Verde L., 2018, Phys. Rev. D, 97, 023531
Bocquet S., Carter F. W., 2016, J. Open Source Softw., 1, 46
Bose B., Taruya A., 2018, J. Cosmol. Astropart. Phys., 1810, 019
Byrnes C. T., Choi K.-Y., 2010, Adv. Astron., 2010, 724525
Castiblanco L., Gannouji R., Noreña J., Stahl C., 2019, J. Cosmol. Astropart. Phys., 1907, 030
Charnock T., Lavaux G., Wandelt B. D., 2018, Phys. Rev. D, 97, 083004
Child H. L., Takada M., Nishimichi T., Sunayama T., Slepian Z., Habib S., Heitmann K., 2018, Phys. Rev. D, 98, 123521
Clarkson C., de Weerd E. M., Jolicoeur S., Maartens R., Umeh O., 2019, MNRAS, 486, L101
Colavincenzo M. et al., 2019, MNRAS, 482, 4883
Coulton W. R., Liu J., Madhavacheril M. S., B'ohm V., Spergel D. N., 2019, J. Cosmol. Astropart. Phys., 1905, 043
D'Amico G., Gleyzes J., Kokron N., Markovic D., Senatore L., Zhang P., Beutler F., Gil-Marín H., 2020, JCAP, 05, 005
Dawson K. S. et al., 2013, AJ, 145, 10
Desjacques V., Jeong D., Schmidt F., 2018, J. Cosmol. Astropart. Phys., 1812, 035
Di Dio E., Durrer R., Maartens R., Montanari F., Umeh O., 2019, J. Cosmol. Astropart. Phys., 1904, 053
Eggemeier A., Scoccimarro R., Smith R. E., 2019, Phys. Rev. D, 99, 123514
Eisenstein D. J. et al., 2011, AJ, 142, 72
Ellis R. et al., 2014, PASJ, 66, R1
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Gil-Marin H., Wagner C., Fragkoudi F., Jimenez R., Verde L., 2012, J. Cosmol. Astropart. Phys., 1202, 047
Gil-Marín H., Noreña J., Verde L., Percival W. J., Wagner C., Manera M., Schneider D. P., 2015, MNRAS, 451, 539
Gil-Marín H., Percival W. J., Verde L., Brownstein J. R., Chuang C.-H., Kitaura F.-S., Rodríguez-Torres S. A., Olmstead M. D., 2017, MNRAS, 465, 1757

Gualdi D., Manera M., Joachimi B., Lahav O., 2018, MNRAS, 476, 4045
Gualdi D., Gil-Marín H., Manera M., Joachimi B., Lahav O., 2019a, MNRAS, 484, L29
Gualdi D., Gil-Marín H., Schuhmann R. L., Manera M., Joachimi B., Lahav O., 2019b, MNRAS, 484, 3713
Hahn C., Francisco V.-N., Emanuele C., Roman S., 2020, JCAP, 03, 040
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hashimoto I., Rasera Y., Taruya A., 2017, Phys. Rev. D, 96, 043526
Heavens A., Jimenez R., Lahav O., 2000, MNRAS, 317, 965
Hoffmann K., Mao Y., Xu J., Mo H., Wandelt B. D., 2019, MNRAS, 487, 3050
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Jeong D., Komatsu E., 2009, ApJ, 703, 1230
Jones E., Oliphant T., Peterson P., 2001, SciPy: Open Source Scientific Tools for Python
Karagiannis D., Lazanu A., Liguori M., Raccanelli A., Bartolo N., Verde L., 2018, MNRAS, 478, 1341
Kayo I., Takada M., Jain B., 2013, MNRAS, 429, 344
Kernighan B. W., 1988, The C Programming Language, 2nd edn. Prentice Hall, Inc., Upper Saddle River, NJ
Kilbinger M., Schneider P., 2005, A&A, 442, 69
Kitaura F.-S. et al., 2016, MNRAS, 456, 4156
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Lesgourgues J., 2011, preprint (arXiv:1104.2932)
Levi M. et al., 2013, preprint (arXiv:1308.0847)
Nan Y., Yamamoto K., Hikage C., 2018, J. Cosmol. Astropart. Phys., 1807, 038
Oddo A., Sefusatti E., Porciani C., Monaco P., Sánchez A. G., 2020, JCAP, 03, 056
Pearson D. W., Samushia L., 2018, MNRAS, 478, 4500
Perez F., Granger B. E., 2007, Comput. Sci. Eng., 9, 21
Rizzato M., Benabed K., Bernardeau F., Lacasa F., 2019, MNRAS, 490, 4688
Rossum G., 1995, Python Reference Manual Technical Report. Centre for Mathematics and Computer Science, Amsterdam
Ruggeri R., Castorina E., Carbone C., Sefusatti E., 2018, J. Cosmol. Astropart. Phys., 1803, 003
Sabiu C. G., Hoyle B., Kim J., Li X.-D., 2019, ApJS, 242, 29
Schmit C. J., Heavens A. F., Pritchard J. R., 2019, MNRAS, 483, 4259
Scoccimarro R., Sefusatti E., Zaldarriaga M., 2004, Phys. Rev. D, 69, 103513
Scoccimarro R., Hui L., Manera M., Chan K. C., 2012, Phys. Rev. D, 85, 083002
Sefusatti E., 2009, Phys. Rev. D, 80, 123002
Sefusatti E., Crocce M., Desjacques V., 2012, MNRAS, 425, 2903
Slepian Z. et al., 2017a, MNRAS, 468, 1070
Slepian Z. et al., 2017b, MNRAS, 469, 1738
Sugiyama N. S., Saito S., Beutler F., Seo H.-J., 2019, MNRAS, 484, 364
Sugiyama N. S., Saito S., Beutler F., Seo H.-J., 2020, MNRAS, 00, 00
Takada M., Jain B., 2004, MNRAS, 348, 897
Taylor A., Joachimi B., 2014, MNRAS, 442, 2728
Tegmark M., Taylor A., Heavens A., 1997, ApJ, 480, 22
van der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Verde L., Wang L.-M., Heavens A., Kamionkowski M., 2000, MNRAS, 313, L141
Watkinson C. A., Giri S. K., Ross H. E., Dixon K. L., Iliev I. T., Mellema G., Pritchard J. R., 2019, MNRAS, 482, 2653
Yamamoto K., Nan Y., Hikage C., 2017, Phys. Rev. D, 95, 043528
Yankelevich V., Porciani C., 2019, MNRAS, 483, 2078

## APPENDIX A: DATA VECTOR MODELS

In this section, we list the analytical expressions we used for computing the various terms of the data vector. The monopole and quadrupole of the galaxy power spectrum are given by

$$P_{\rm g}^{(\ell)}(k) = \frac{2\ell+1}{2}\int_{-1}^{+1} {\rm d}\mu\, P_{\rm g}^{(s)}(k,\mu)\, L_\ell(\mu)\ , \tag{A1}$$

where $L_\ell(\mu)$ is the $\ell$-order Legendre polynomial and $P_{\rm g}^{(s)}(k,\mu)$ is the redshift space galaxy power spectrum at tree level (Gualdi et al. 2019b). For the bispectrum monopole, we adopt the effective formula given by Gil-Marin et al. (2012), which was calibrated on simulations

$$\begin{aligned} B_{\rm g}^{(0)}(k_1,k_2,k_3) &= \frac{1}{4}\int_{-1}^{1}{\rm d}\mu_1\int_{-1}^{1}{\rm d}\mu_2\, B_{\rm g}^{(s)}(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3) \\ &= \frac{1}{4\pi}\int_{-1}^{1}{\rm d}\mu_1\int_{0}^{2\pi}{\rm d}\phi\, B_{\rm g}^{(s)}(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3)\ , \end{aligned} \tag{A2}$$

where $\mu_i$ is the angle between the $\boldsymbol{k}_i$ vector and the line of sight. The angle $\phi$ is defined as $\mu_2 \equiv \mu_1 x_{12} - \sqrt{1-\mu_1^2}\sqrt{1-x_{12}^2}\cos\phi$, where $x_{12}$ is the cosine of the angle between $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$.

## APPENDIX B: PRIMORDIAL NON-GAUSSIANITY EXPANSION

Following what was done in Scoccimarro et al. (2004), we can compute for the power spectrum and the bispectrum, the contribution due to the presence of a primordial non-Gaussian component in the potential field. In order to do so, we assume a local type of non-Gaussianity that, in terms of the primordial potential, can be parametrized as

$$\Phi_{\rm p}(\boldsymbol{x}) = \phi_{\rm p}(\boldsymbol{x}) + \frac{f_{\rm NL}}{c^2}\left[\phi_{\rm p}^2(\boldsymbol{x}) - \langle\phi_{\rm p}^2(\boldsymbol{x})\rangle\right] + \frac{g_{\rm NL}}{c^4}\left[\phi_{\rm p}^3(\boldsymbol{x}) - 3\phi_{\rm p}(\boldsymbol{x})\langle\phi_{\rm p}^2(\boldsymbol{x})\rangle\right] + \ldots, \tag{B1}$$

where $\phi_{\rm p}$ represents a Gaussian field, while $f_{\rm NL}$ and $g_{\rm NL}$ are the constant parameters of the expansion up to third order in $\phi_{\rm p}$. In Fourier space, it translates into (dropping the 'p' index for $\phi$)

$$\Phi_{\rm p}(\boldsymbol{k}) = \phi_k + \frac{f_{\rm NL}}{c^2}\left[I^k_{ab}\phi_a\phi_b - \delta_D(\boldsymbol{k})\langle\phi^2\rangle\right] + \frac{g_{\rm NL}}{c^4}\left[I^k_{abc}\phi_a\phi_b\phi_c - \frac{3}{(2\pi)^3}\phi_k\langle\phi^2\rangle\right]\ , \tag{B2}$$

where, in Fourier space, $\langle\phi^2\rangle = \int {\rm d}\boldsymbol{q}^3 P_\phi(\boldsymbol{q}) = (2\pi)^3\sigma_\phi^2$ and $\delta_{\rm D}(\boldsymbol{k})$ is the Dirac's delta function. We introduced the short notation for the integral over the wavevectors:

$$\begin{aligned} I^k_{ab} &= \int \frac{{\rm d}\boldsymbol{q}_a^3{\rm d}\boldsymbol{q}_b^3}{(2\pi)^3}\,\delta_{\rm D}(\boldsymbol{k}-\boldsymbol{q}_a-\boldsymbol{q}_b) \\ I^k_{abc} &= \int \frac{{\rm d}\boldsymbol{q}_a^3{\rm d}\boldsymbol{q}_b^3{\rm d}\boldsymbol{q}_c^3}{(2\pi)^6}\,\delta_{\rm D}(\boldsymbol{k}-\boldsymbol{q}_a-\boldsymbol{q}_b-\boldsymbol{q}_c)\,. \end{aligned} \tag{B3}$$

The primordial potential is related to the late-time one by

$$\Phi_{\rm l.t.}(a) = \frac{9}{10}\frac{D_+}{a}T(k)\Phi_{\rm p}\,, \tag{B4}$$

where $D_+(a)$ is the growth factor from the linear perturbation theory as a function of the scale factor $a$. $T(k)$ is the transfer function normalized to unity for $k\to 0$. At late times, the potential field is related to the density perturbation variable by the Poisson equation:

$$\nabla^2\Phi_{\rm l.t.}(\boldsymbol{x},a) = \frac{3}{2}\frac{\Omega_{\rm m}H_0^2}{a}\delta(\boldsymbol{x},a)\,. \tag{B5}$$

This allows to link the primordial potential with the late-time matter density perturbation:

$$\delta_k = \frac{3}{5}\frac{D_+}{\Omega_{\rm m}H_0^2}k^2T_k\Phi_{\rm p} = \beta k^2 T_k\Phi_{\rm p}\,. \tag{B6}$$

### B1 Power spectrum

Let us start with the two-point correlation function in Fourier space. We will consider all the terms up to order $\phi^4$:

$$\langle\delta_{\rm s}\delta_{\rm s}\rangle = \langle(\delta_{\rm m}^{(1)}+\delta_{\rm m}^{(2)}+\delta_{\rm m}^{(3)}+O(\delta_{\rm m}^{(4)}))(\delta_{\rm m}^{(1)}+\delta_{\rm m}^{(2)}+\delta_{\rm m}^{(3)}+O(\delta_{\rm m}^{(4)}))\rangle\,, \tag{B7}$$

where the upper index represents the order in terms of $\delta_{\rm m}$ given by the expansion done in the previous section. Up to the considered order, we then have

$$\begin{aligned} \langle\delta_{\rm s}\delta_{\rm s}\rangle &= \langle\delta_{\rm m}^{(1)}\delta_{\rm m}^{(1)}\rangle + 2\langle\delta_{\rm m}^{(1)}\delta_{\rm m}^{(2)}\rangle + \langle\delta_{\rm m}^{(2)}\delta_{\rm m}^{(2)}\rangle + 2\langle\delta_{\rm m}^{(1)}\delta_{\rm m}^{(3)}\rangle \\ &= P_{11}+P_{12}+P_{22}+P_{13}\,. \end{aligned} \tag{B8}$$

*B1.1* $P_{11}$

Expanding in term of the primordial Gaussian potential $\phi$:

$$\begin{aligned}\langle\delta_{\rm m}^{(1)}(\boldsymbol{k})\delta_{\rm m}^{(1)}(\boldsymbol{q})\rangle = \Bigg\langle & F_k^{(1)}F_q^{(1)}\beta^2k^2q^2T_kT_q \\ & \times\left\{\phi_k+\frac{f_{\rm NL}}{c^2}\left[I^k_{ab}\phi_a\phi_b-\delta_{\rm D}(\boldsymbol{k})\langle\phi^2\rangle\right]+\frac{g_{\rm NL}}{c^4}\left[I^k_{def}\phi_d\phi_e\phi_f-\frac{3}{(2\pi)^3}\phi_k\langle\phi^2\rangle\right]\right\} \\ & \times\left\{\phi_q+\frac{f_{\rm NL}}{c^2}\left[I^q_{gh}\phi_g\phi_h-\delta_{\rm D}(\boldsymbol{q})\langle\phi^2\rangle\right]+\frac{g_{\rm NL}}{c^4}\left[I^q_{ilm}\phi_i\phi_l\phi_m-\frac{3}{(2\pi)^3}\phi_q\langle\phi^2\rangle\right]\right\}\Bigg\rangle\,. \end{aligned} \tag{B9}$$

Recalling that all odd moments of a Gaussian variable ($\phi$) are equal to zero and ignoring all higher order terms ($>\phi^4$), we obtain

$$\begin{aligned}\langle\delta_{\rm m}^{(1)}(\boldsymbol{k})\delta_{\rm m}^{(1)}(\boldsymbol{q})\rangle &= \Bigg\langle F_k^{(1)}F_q^{(1)}\beta^2k^2q^2T_kT_q \\ &\quad\times\left\{\phi_k\phi_q+\frac{f_{\rm NL}^2}{c^4}\left[I^k_{ab}\phi_a\phi_b-\delta_{\rm D}(\boldsymbol{k})\langle\phi^2\rangle\right]\times\left[I^q_{gh}\phi_g\phi_h-\delta_{\rm D}(\boldsymbol{q})\langle\phi^2\rangle\right]+2\phi_k\frac{g_{\rm NL}}{c^4}\left[I^q_{ilm}\phi_i\phi_l\phi_m-\frac{3}{(2\pi)^3}\phi_q\langle\phi^2\rangle\right]\right\}\Bigg\rangle \\ &= F_k^{(1)}F_q^{(1)}\beta^2k^2q^2T_kT_q\times\left\{(2\pi)^3\delta_{\rm D}^{kq}P_k^\phi+\frac{f_{\rm NL}^2}{c^4}\left[I^k_{ab}I^q_{gh}\langle\phi_a\phi_b\phi_g\phi_h\rangle-2I^k_{ab}\langle\phi_a\phi_b\rangle\delta_{\rm D}^q\langle\phi^2\rangle+\delta_{\rm D}^k\delta_{\rm D}^q\langle\phi^2\rangle^2\right]\right. \\ &\quad\left.+\frac{g_{\rm NL}}{c^4}\left[I^q_{ilm}\langle\phi_i\phi_l\phi_m\phi_k\rangle-\frac{3}{(2\pi)^3}\langle\phi_k\phi_q\rangle\langle\phi^2\rangle\right]\right\} \\ &= F_k^{(1)}F_q^{(1)}\beta^2k^2q^2T_kT_q \\ &\quad\times\left\{(2\pi)^3\delta_{\rm D}^{kq}P_k^\phi+\frac{f_{\rm NL}^2}{c^4}\left[2I^k_{ab}(2\pi)^3\delta_{\rm D}^{abq}P_a^\phi P_b^\phi+\delta_{\rm D}^k\delta_{\rm D}^q\langle\phi^2\rangle^2-2\delta_{\rm D}^k\delta_{\rm D}^q\langle\phi^2\rangle^2+\delta_{\rm D}^k\delta_{\rm D}^q\langle\phi^2\rangle^2\right]\right. \\ &\quad\left.+\frac{g_{\rm NL}}{c^4}\left[I^q_{ilm}3\langle\phi_i\phi_l\rangle\langle\phi_m\phi_k\rangle-3\delta_{\rm D}^{kq}P_k^\phi\langle\phi^2\rangle\right]\right\} \\ &= F_k^{(1)}F_q^{(1)}\beta^2k^2q^2T_kT_q(2\pi)^3\delta_{\rm D}^{kq} \\ &\quad\times\left\{P_k^\phi+\frac{2f_{\rm NL}^2}{c^4}\int\frac{d\boldsymbol{p}_a^3}{(2\pi)^3}P_a^\phi P^\phi_{|\boldsymbol{k}-\boldsymbol{p}_a|}+\frac{g_{\rm NL}}{c^4(2\pi)^3}\left[3P_k^\phi\int d\boldsymbol{p}_l^3P_l^\phi-3P_k^\phi\langle\phi^2\rangle\right]\right\} \\ &= F_k^{(1)}F_q^{(1)}\beta^2k^2q^2T_kT_q(2\pi)^3\delta_{\rm D}^{kq}\left\{P_k^\phi+\frac{2f_{\rm NL}^2}{c^4}\int\frac{d\boldsymbol{p}_a^3}{(2\pi)^3}P_a^\phi P^\phi_{|\boldsymbol{k}-\boldsymbol{p}_a|}\right\}, \end{aligned} \tag{B10}$$

where $\delta_{\rm D}^{kq}=\delta_{\rm D}(\boldsymbol{k}+\boldsymbol{q})$.

*B1.2* $P_{12}$

$$\begin{aligned}2\langle\delta_{\rm m}^{(1)}(\boldsymbol{k})\delta_{\rm m}^{(2)}(\boldsymbol{q})\rangle &= \left\langle 2F_k^{(1)}I^q_{ab}F^{(2)}_{ab}\delta_k^\ell\delta_a^\ell\delta_b^\ell\right\rangle \\ &= \Bigg\langle 2F_k^{(1)}\beta^3k^2T_kI^q_{ab}q_a^2T_aq_b^2T_bF^{(2)}_{ab} \\ &\quad\times\left\{\phi_k+\frac{f_{\rm NL}}{c^2}\left[I^k_{cd}\phi_c\phi_d-\delta_{\rm D}(\boldsymbol{k})\langle\phi^2\rangle\right]+\frac{g_{\rm NL}}{c^4}\left[I^k_{def}\phi_e\phi_f\phi_g-\frac{3}{(2\pi)^3}\phi_k\langle\phi^2\rangle\right]\right\} \\ &\quad\times\left\{\phi_a+\frac{f_{\rm NL}}{c^2}\left[I^a_{hi}\phi_h\phi_i-\delta_{\rm D}(\boldsymbol{p}_a)\langle\phi^2\rangle\right]+\frac{g_{\rm NL}}{c^4}\left[I^a_{lmn}\phi_l\phi_m\phi_n-\frac{3}{(2\pi)^3}\phi_a\langle\phi^2\rangle\right]\right\} \\ &\quad\times\left\{\phi_b+\frac{f_{\rm NL}}{c^2}\left[I^b_{or}\phi_o\phi_r-\delta_{\rm D}(\boldsymbol{p}_b)\langle\phi^2\rangle\right]+\frac{g_{\rm NL}}{c^4}\left[I^b_{stv}\phi_s\phi_t\phi_v-\frac{3}{(2\pi)^3}\phi_b\langle\phi^2\rangle\right]\right\}\Bigg\rangle\,. \end{aligned} \tag{B11}$$

This at maximum order $\phi^4$ returns only one term proportional to $f_{\rm NL}$:

$$\begin{aligned}2\langle\delta_{\rm m}^{(1)}(\boldsymbol{k})\delta_{\rm m}^{(2)}(\boldsymbol{q})\rangle &= 2F_k^{(1)}\beta^3k^2T_kI^q_{ab}q_a^2T_aq_b^2T_bF^{(2)}_{ab}\frac{f_{\rm NL}}{c^2} \\ &\quad\times\left\{I^k_{cd}\langle\phi_c\phi_d\phi_a\phi_b\rangle-\langle\phi_c\phi_d\rangle\delta_{\rm D}(\boldsymbol{k})\langle\phi^2\rangle+2I^a_{hi}\langle\phi_k\phi_b\phi_h\phi_i\rangle-2\langle\phi_h\phi_i\rangle\delta_{\rm D}(\boldsymbol{p}_a)\langle\phi^2\rangle\right\} \\ &= (2\pi)^6\frac{4f_{\rm NL}}{c^2}F_k^{(1)}\beta^3k^2{\rm T}_kI^q_{ab}q_a^2T_aq_b^2T_bF^{(2)}_{ab}\times\left\{I^k_{cd}\delta_{\rm D}^{ac}\delta_{\rm D}^{bd}P_a^\phi P_b^\phi+2I^a_{hi}\delta_{\rm D}^{ki}\delta_{\rm D}^{bh}P_k^\phi P_b^\phi\right\} \\ &= (2\pi)^3\delta_{\rm D}^{kq}\frac{4f_{\rm NL}}{c^2}F_k^{(1)}\beta^3k^2T_k\int\frac{d\boldsymbol{p}_a^3}{(2\pi)^3}p_a^2T_a|-\boldsymbol{k}-\boldsymbol{p}_a|^2T_{|-\boldsymbol{k}-\boldsymbol{p}_a|}F^{(2)}_{a,-\boldsymbol{k}-\boldsymbol{p}_a}P^\phi_{|-\boldsymbol{k}-\boldsymbol{p}_a|}\left[P_a^\phi+2P_k^\phi\right]\,. \end{aligned} \tag{B12}$$

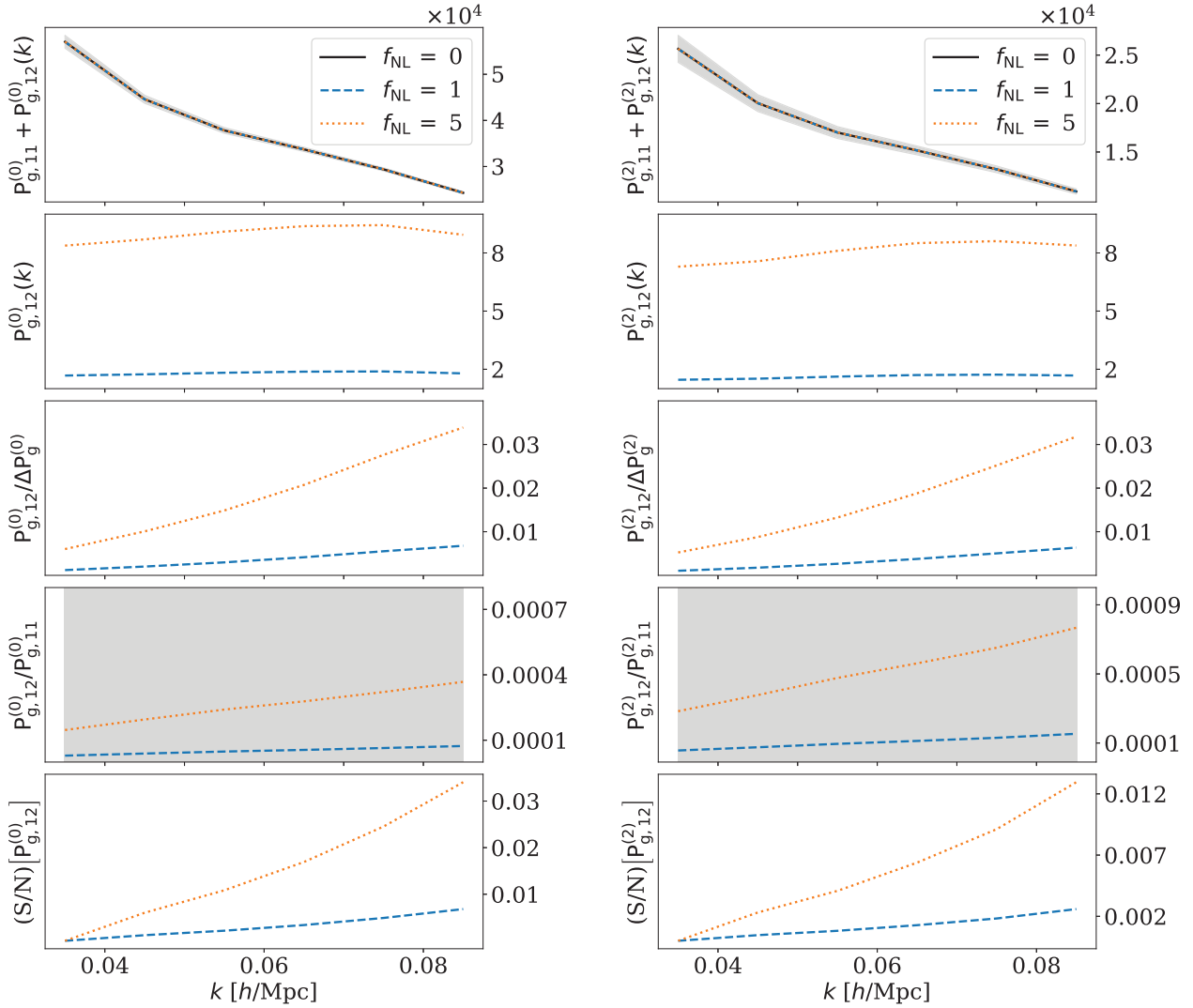



**Figure B1.** $P^{(0,2)}_{g}$: primordial non-Gaussianity contribution. In the first row, the power spectrum monopole and quadrupole model is shown for different values of local primordial non-Gaussianity, $f_{NL} = 0, 1, 5$. The shaded area corresponds to the error bars derived from the 1400 galaxy mocks measurements (Kitaura et al. 2016). The second row shows only the power spectrum terms linearly proportional to $f_{NL}$. The third row shows the ratio between the primordial term and the data-point error bar width, defined as $\Delta P^{(0,2)}_{g}(k_i) = \sqrt{\mathrm{Cov}_{ii}}$. The fourth row shows the ratio between the primordial term and the gravitational collapse one for the power spectrum. In the last row, we show the primordial part cumulative signal to noise as a function of the number of $k$-bins included in the analysis. This is defined as $(\mathrm{S/N})\left[P^{(0,2)}_{g,12}\right]_i = \sqrt{P^{(0,2),\intercal}_{g,12,i} \cdot \mathrm{Cov}^{-1} \cdot P^{(0,2)}_{g,12,i}}$, where $P^{(0,2)}_{g,12,i}$ is power spectrum monopole/quadrupole data vector up to the $k$-bin $k_i$. $\mathrm{Cov}^{-1}$ is the covariance matrix for the reduced data vector $P^{(0,2)}_{g,12,i}$.

The other terms in the power spectrum expansion, $P_{22}$ and $P_{13}$, are at first order already proportional to $\phi^4$ and, therefore, in our case, they just return the standard loop correction terms for the Gaussian initial conditions. In Fig. B1, the overall and relative effect of primordial non-Gaussianity can be observed for $f_{NL=0,1,5}$ in comparison to the current error bars.

### B2 Bispectrum

Also for the bispectrum, we limit the expansion to the terms proportional to $\phi^4$ so that we do not need to use fourth-order perturbation theory. However, for clarity, we will list all terms up to order $\phi^6$ even if we are not going to compute them explicitly.

$$\begin{aligned}\langle\delta_s\delta_s\delta_s\rangle &= \langle(\delta^{(1)}_m + \delta^{(2)}_m + \delta^{(3)}_m + O(\delta^{(4)}_m))(\delta^{(1)}_m + \delta^{(2)}_m + \delta^{(3)}_m + O(\delta^{(4)}_m))(\delta^{(1)}_m + \delta^{(2)}_m + \delta^{(3)}_m + O(\delta^{(4)}_m))\rangle \\ &= \langle\delta^{(1)}_m\delta^{(1)}_m\delta^{(1)}_m\rangle + 3\langle\delta^{(1)}_m\delta^{(1)}_m\delta^{(2)}_m\rangle \\ &\quad + 3\langle\delta^{(2)}_m\delta^{(2)}_m\delta^{(1)}_m\rangle + 3\langle\delta^{(1)}_m\delta^{(1)}_m\delta^{(3)}_m\rangle \\ &\quad + 3\langle\delta^{(1)}_m\delta^{(1)}_m\delta^{(4)}_m\rangle + 6\langle\delta^{(1)}_m\delta^{(2)}_m\delta^{(3)}_m\rangle + \langle\delta^{(2)}_m\delta^{(2)}_m\delta^{(2)}_m\rangle \\ &= B_{111} + B_{112} + B_{122} + B_{113} + B_{114} + B_{123} + B_{222}\,. \end{aligned} \tag{B13}$$

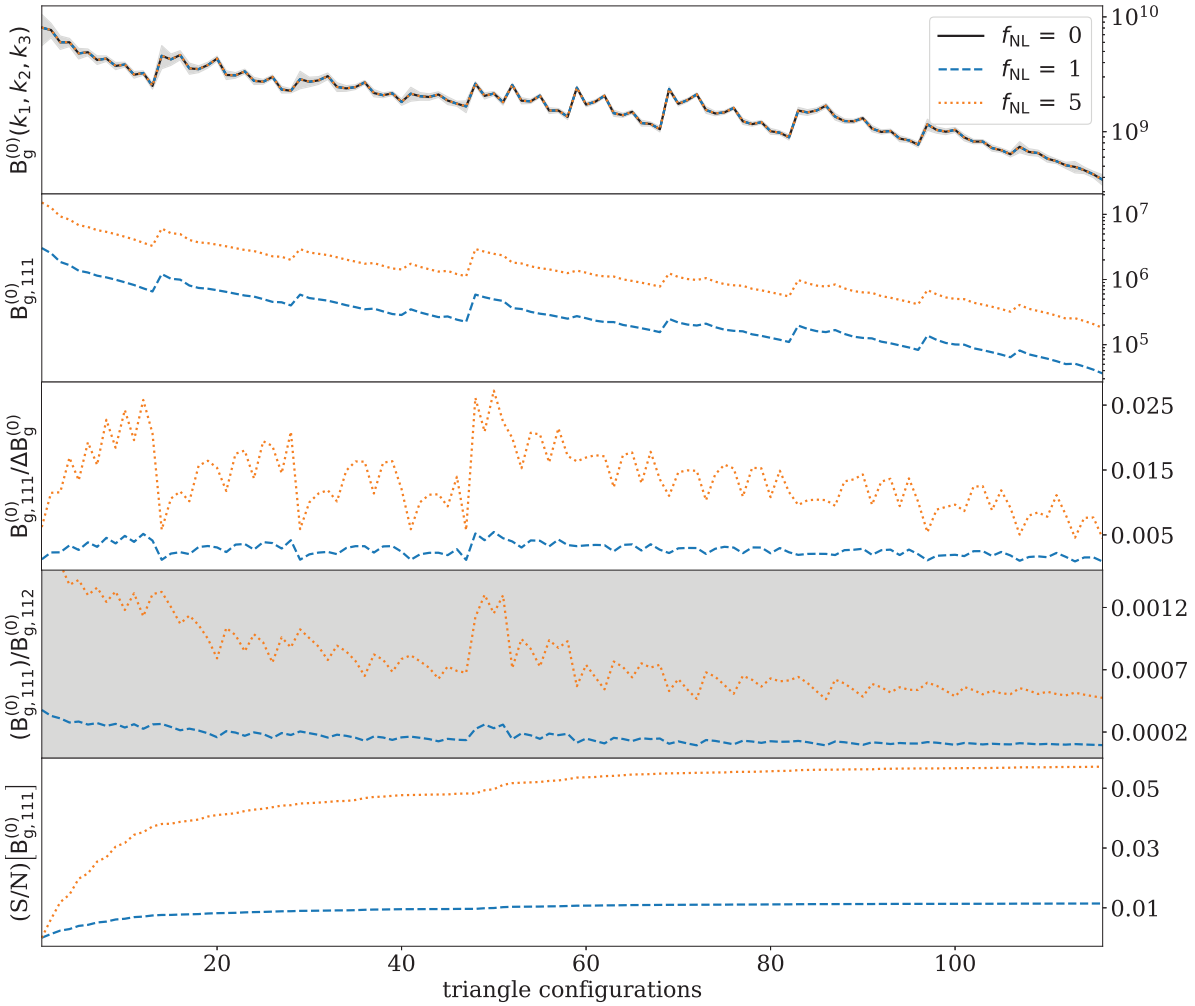



**Figure B2.** $B_g^{(0)}$: primordial non-Gaussianity contribution. Same as Fig. B1 for the galaxy bispectrum monopole.

From the expansion above, we can see that if we would only consider terms up to order $\phi^6$, then $B_{114}$, $B_{123}$, and $B_{222}$ would just result into loop corrections without primordial non-Gaussianity contributions. From $B_{122}$ and $B_{113}$, we would have only terms proportional to $f_{NL}$ since the ones given by Gaussian initial conditions are all equal to zero (odd moments).

For $B_{111}$, we would need to consider an additional parameter for the primordial non-Gaussianity contribution, in particular one proportional to $\phi^4$. The other terms up to order $\phi^6$ originating from $B_{111}$ would be either proportional to $f_{NL}^3$ or to the product $f_{NL}g_{NL}$. Limiting ourselves only at order $\phi^4$, from $B_{111}$, we will obtain only one term proportional to $f_{NL}$.

Finally, $B_{112}$ would return in the case of Gaussian initial conditions the standard tree level expression for the bispectrum. When PNG are considered up to order $\phi^6$, $B_{112}$ contains also terms proportional to both $f_{NL}^2$ and $g_{NL}$.

### *B2.1* $B_{111}$

We proceed then with the only term containing primordial non-Gaussianity contributions up to order $\phi^4$:

$$\begin{aligned}\langle \delta_m^{(1)} \delta_m^{(1)} \delta_m^{(1)} \rangle = \Bigg\langle & F_{k_1}^{(1)} F_{k_2}^{(1)} F_{k_3}^{(1)} \beta^3 k_1^2 k_2^2 k_3^2 T_{k_1} T_{k_2} T_{k_3} \\ & \times \left\{ \phi_{k_1} + \frac{f_{NL}}{c^2} \left[ I_{ab}^{k_1} \phi_a \phi_b - \delta_D(\boldsymbol{k}_1) \langle \phi^2 \rangle \right] + \frac{g_{NL}}{c^4} \left[ I_{cde}^{k_1} \phi_c \phi_d \phi_e - \frac{3}{(2\pi)^3} \phi_{k_1} \langle \phi^2 \rangle \right] \right\} \\ & \times \left\{ \phi_{k_2} + \frac{f_{NL}}{c^2} \left[ I_{fg}^{k_2} \phi_f \phi_g - \delta_D(\boldsymbol{k}_2) \langle \phi^2 \rangle \right] + \frac{g_{NL}}{c^4} \left[ I_{hil}^{k_2} \phi_h \phi_i \phi_l - \frac{3}{(2\pi)^3} \phi_{k_2} \langle \phi^2 \rangle \right] \right\} \\ & \times \left\{ \phi_{k_3} + \frac{f_{NL}}{c^2} \left[ I_{mn}^{k_3} \phi_m \phi_n - \delta_D(\boldsymbol{k}_3) \langle \phi^2 \rangle \right] + \frac{g_{NL}}{c^4} \left[ I_{oqr}^{k_3} \phi_o \phi_q \phi_r - \frac{3}{(2\pi)^3} \phi_{k_3} \langle \phi^2 \rangle \right] \right\} \Bigg\rangle . \end{aligned} \tag{B14}$$

The resulting three terms proportional to $f_{\rm NL}$ are equivalent, different only by the permutation between $k_1$, $k_2$, and $k_3$:

$$\begin{aligned}\langle\delta^{(1)}_{\rm m}\delta^{(1)}_{\rm m}\delta^{(1)}_{\rm m}\rangle &= F^{(1)}_{k_1}F^{(1)}_{k_2}F^{(1)}_{k_3}\beta^3k_1^2k_2^2k_3^2T_{k_1}T_{k_2}T_{k_3}\frac{f_{\rm NL}}{c^2}\\ &\quad\times\left\{I^{k_1}_{ab}\langle\phi_{k_2}\phi_{k_3}\phi_a\phi_b\rangle-\delta_{\rm D}(\boldsymbol{k}_1)\langle\phi_{k_2}\phi_{k_3}\rangle\langle\phi^2\rangle\right\}\;+\;\text{cyc.}\\ &= F^{(1)}_{k_1}F^{(1)}_{k_2}F^{(1)}_{k_3}\beta^3k_1^2k_2^2k_3^2T_{k_1}T_{k_2}T_{k_3}\frac{f_{\rm NL}}{c^2}I^{k_1}_{ab}2(2\pi)^6\delta_{\rm D}^{k_2a}\delta_{\rm D}^{k_3b}P^{\phi}_{k_2}P^{\phi}_{k_3}\;+\;\text{cyc.}\\ &= (2\pi)^3F^{(1)}_{k_1}F^{(1)}_{k_2}F^{(1)}_{k_3}\beta^3k_1^2k_2^2k_3^2T_{k_1}T_{k_2}T_{k_3}\frac{2f_{\rm NL}}{c^2}\delta_{\rm D}(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3)P^{\phi}_{k_2}P^{\phi}_{k_3}\;+\;\text{cyc.}\\ &= (2\pi)^3\delta_{\rm D}(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3)F^{(1)}_{k_1}F^{(1)}_{k_2}F^{(1)}_{k_3}\beta^{-1}k_1^2k_2^{-2}k_3^{-2}\frac{T_{k_1}}{T_{k_2}T_{k_3}}\frac{2f_{\rm NL}}{c^2}P^{\rm m}_{k_2}P^{\rm m}_{k_3}\;+\;\text{cyc.}\,,\end{aligned}\tag{B15}$$

where, in the last line, the primordial power spectrum was converted into the late-time matter power spectrum. The primordial non-Gaussianity contribution to the galaxy bispectrum monopole is shown in Fig. B2.

## APPENDIX C: OUTLIERS ANALYSIS

In the left-hand panels of Figs 1, 4 and 5, we can observe that for few galaxy catalogues, the relative parameter constraints improvements are negative. In the same figures, we also check that the overall average improvement on the parameters constraints is always positive.

We show in Fig. C1 the 1–2D marginalized posterior distribution for one of the galaxy catalogues showing this negative improvement for some of the constrained parameters. In particular, we show the posterior for the MCMC on the full data vector (including 116 triangle configurations) and two different runs of the GEOMAX algorithm applied to the larger set of triangles (2734). The first run is relative to the compression being computed using the same fiducial cosmology used throughout this paper. For the second run, we computed the compression using the best-fitting parameters set obtained from the first run.

Indeed, we wanted to check whether the observed negative improvement for certain parameters was due to the difference between the assumed fiducial cosmology and the best-fitting one. This is not the case. None the less, in this way, we verified on this single case that the GEOMAX compression performance is not sensitive to the fiducial cosmology used to derive the compression weights.

We then conclude that the negative improvement observed for certain parameters is a statistical effect counterbalancing the above average improvement for the rest of the parameters. This second hypothesis is supported by the average column in the left-hand panels of Figs 1, 4, and 5.

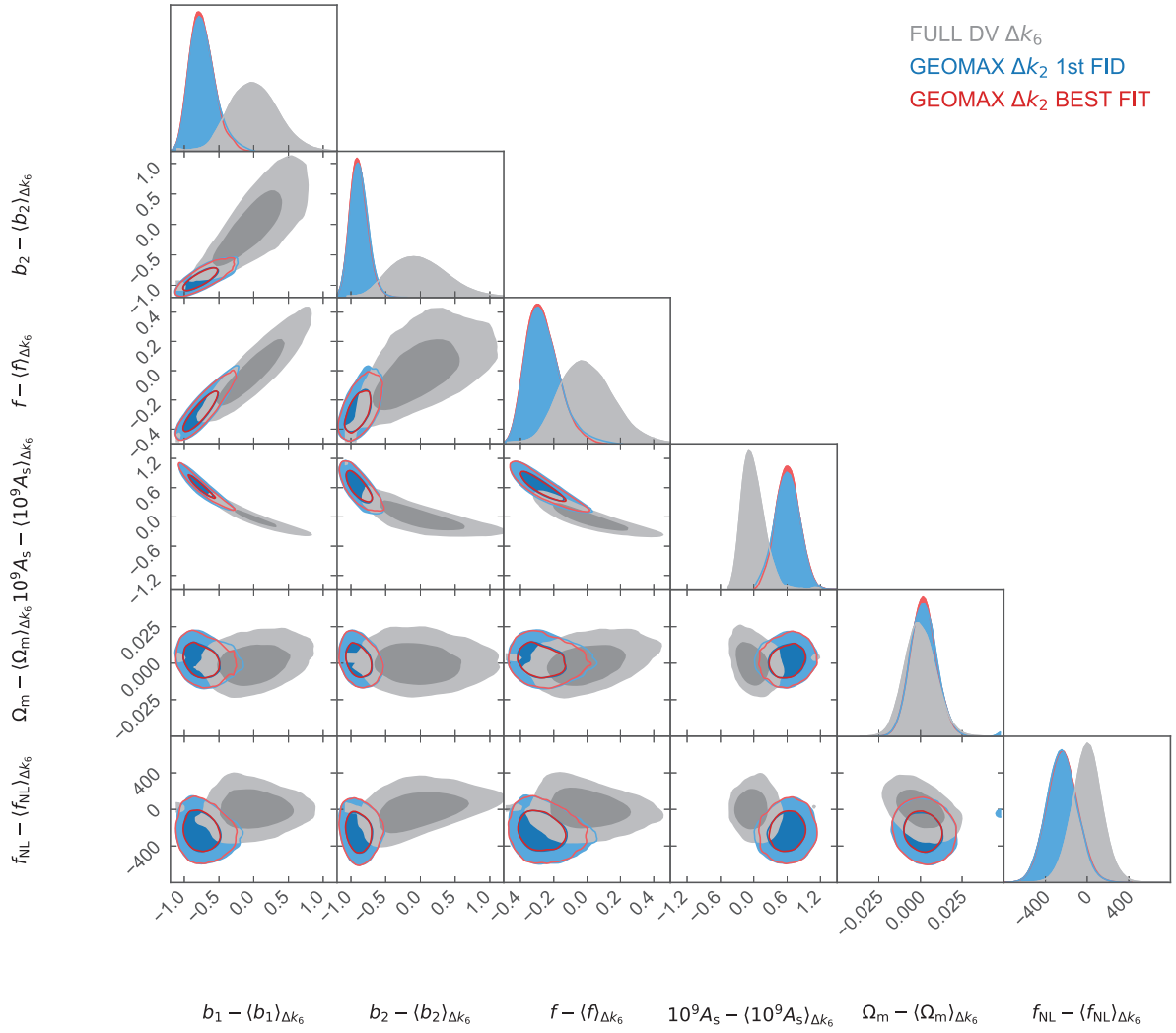



**Figure C1.** Posterior for the galaxy catalogue 37: Beside the contours relative to the MCMC on the full data vector containing 116 triangle configurations, we show the result of GEOMAX compression applied to 2734 triangles' bispectra. In the first case ('FID'), we used the fiducial cosmology to compute the derivatives needed for the compression. The second set of contours ('BEST FIT') where instead derived by using the best-fitting parameters obtained through the first run ('FID') to compute the derivatives needed in the geometrical step. Even if these two sets have differences larger than $1\sigma$ intervals for certain parameters, the 1–2D posterior contours given by GEOMAX do not significantly differ.

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