Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **484,** L29–L34 (2019)
Advance Access publication 2019 January 2

doi:10.1093/mnrasl/sly242

Geometrical compression: a new method to enhance the BOSS galaxy bispectrum monopole constraints

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

[1] *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[2] *Institut Lagrange de Paris (ILP), Sorbonne Université, 98 bis Boulevard Arago, 75014 Paris, France*
[3] *Laboratoire de Physique Nucléaire et de Hautes Energies, Université Pierre et Marie Curie, Paris, France*
[4] *Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra (Barcelona) Spain*
[5] *Centre for Mathematical Sciences, DAMTP, Cambridge University, Wilberforce Rd, Cambridge CB3 0WA, UK*

Accepted 2018 December 20. Received 2018 December 06; in original form 2018 August 5

ABSTRACT

We present a novel method to compress galaxy clustering three-point statistics and apply it to redshift space galaxy bispectrum monopole measurements from BOSS DR12 CMASS data considering a k-space range of $0.03 - 0.12\,h/\mathrm{Mpc}$. The method consists in binning together bispectra evaluated at sets of wavenumbers forming closed triangles with similar geometrical properties: the area, the cosine of the largest angle, and the ratio between the cosines of the remaining two angles. This enables us to increase the number of bispectrum measurements, for example by a factor of 23 over the standard binning (from 116 to 2734 triangles used), which is otherwise limited by the number of mock catalogues available to estimate the covariance matrix needed to derive parameter constraints. The 68 per cent credible intervals for the inferred parameters (b_1, b_2, f, σ_8) are thus reduced by (−39 per cent, −49 per cent, −29 per cent, −22 per cent), respectively. We find very good agreement with the posteriors recently obtained by alternative maximal compression methods. This new method does not require the a-priori computation of the data vector covariance matrix and has the potential to be directly applicable to other three-point statistics (e.g. galaxy clustering, weak gravitational lensing, 21-cm emission line) measured from future surveys such as DESI, Euclid, PFS, and SKA.

Key words: methods: analytical – methods: data analysis – methods: statistical – cosmology: cosmological parameters – cosmology: large-scale structure of Universe.

1 INTRODUCTION

Three-point (3pt) statistics will be indispensable to fully exploit the large data sets from current and forthcoming cosmological surveys. Their most recent applications to galaxy clustering data sets have been on BOSS for both the bispectrum (Gil-Marín et al. 2017) while (Slepian et al. 2017a) used the 3pt correlation function. Slepian et al. (2017b) also measured baryonic acoustic oscillations (BAO) using the 3pt correlation function and Pearson & Samushia (2018) detected them using the bispectrum. Moreover a new interferometric basis has also been developed by (Child et al. 2018) to highlight the BAO signal in the bispectrum. For the 21-cm emission line, 3pt statistics have been investigated by Hoffmann et al. (2018).

Weak-lensing 3pt statistics in Fourier and real space have also been explored (Takada & Jain 2004; Schneider, Kilbinger & Lombardi 2005; Joachimi, Shi & Schneider 2009; Kayo & Takada 2013; Kayo, Takada & Jain 2013; including early applications to data (Kilbinger & Schneider 2005; Fu et al. 2014).

As shown also recently by Yankelevich & Porciani (2018), considering the bispectrum together with the power spectrum significantly improves the constraints on cosmological parameters, even if using only the bispectrum monopole severely limits these improvements. In order to include higher multipoles, from the data analysis side, data-vector compression becomes essential.

In previous work, we introduced two compression methods for the redshift space galaxy bispectrum in Gualdi et al. (2018a), Paper I hereafter, and tested them on bispectrum monopole measurements from BOSS DR12 data in Gualdi et al. (2018b), Paper II hereafter. Both methods are variations of the method presented in Heavens, Jimenez & Lahav (2000) and named MOPED, which achieves maximal compression of the original data vector by extending to the multiple parameters case the Karhunen–Loève algorithm first intro-

⋆ E-mail: davide.gualdi.14@ucl.ac.uk

duced in Tegmark, Taylor & Heavens (1997). The two techniques require an approximate analytic expression for the data vector covariance matrix. These methods compress the original data vector to a new one with dimension corresponding to the number of model parameters constrained, hence the name 'maximal' compression. Since in Paper II we have shown that the two methods produce very similar results, for the sake of clarity in this work we compare the new method to just one of them. In particular, we use the method consisting in running a Markov Chain Monte Carlo sampling (MCMC) on the compressed data vector, labelling these results 'maximal compression'.

Here, we present a new compression method which consists in averaging bispectra triangle configurations of wavenumbers that have similar geometrical properties. In order to derive parameter constraints, we again use MCMC sampling on the compressed data vector. We label the method geometrical compression (MC-GC).

In particular, we define new bins in terms of the triangle configurations area and functions of the internal angles. The area parametrizes the physical scales information encoded in the two power spectra products present in the bispectrum analytic expression. At the same time, the angles are the variables on which depends the value of the second-order perturbation theory kernel. This can be seen in fig. 2 of (Gil-Marín et al. 2017) where the oscillating pattern in the bispectrum repeats itself because the angles are the same, even if the sizes of k1,k2,k3 increase. Therefore, using these parameters to compress the bispectrum proves to be much more optimal than simply using larger bins defined in terms of the triangle configuration sides.

In Section 2, the data set and the galaxy mocks used together with the settings of our analysis are described. In Section 3, we present the analytical model used for the considered data vector. Section 4 introduces the transformation through which we compress the original data vector. Section 5 describes how to optimally choose the number of bins for the new parameters characterizing the compressed data vector. In Section 6, we compare the MC-GC results with the ones from standard MCMC and one of the two alternative maximal compression techniques, described in Paper II. In Section 7, we conclude and discuss potential future extensions.

2 DATA, MOCKS, AND ANALYSIS

The power spectrum monopole, quadrupole, and bispectrum monopole have been measured from the DR12 CMASS sample ($0.43 \leq z \leq 0.70$) of the Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al. 2013) which is part of the Sloan Digital Sky Survey III (Eisenstein et al. 2011). For more details, see Gil-Marín et al. (2017) and Alam et al. (2017).

The covariance matrix used to estimate the cosmological parameters of interest via standard MCMC on the full data vector has been numerically estimated using 1400 of the 2048 galaxy catalogues of the MultiDark Patchy BOSS DR12 mocks by Kitaura et al. (2016). We only use 700 when the compressed data vector is used in order to consistently compare the new method presented here with the results obtained in Paper II. The underlying cosmology used to realize these mocks is: $\Omega_\Lambda(z=0) = 0.692885$, $\Omega_{\rm m}(z=0) = 0.307115$, $\Omega_{\rm b}(z=0) = 0.048$, $\sigma_8 = 0.8288$, $n_{\rm s} = 0.96$, $h_0 = 0.6777$.

We fix the bin size for the power spectrum monopole and quadrupole to $\Delta k = 0.01h/\mathrm{Mpc}$. We estimated the bispectrum monopole from both data and mocks using different multiples of the fundamental frequency defined as $k_f^3 = \frac{(2\pi)^3}{V_{\rm s}}$, where $V_{\rm s}$ is the survey volume. $V_{\rm s}$ has been set to the mocks case value, which was the one of a cubic box volume $V_{\rm s} = L_{\rm b}^3 = (3500\,\mathrm{Mpc}/h)^3$.

For the bispectrum, we considered the bin sizes $\Delta k_{6,5,2} = 6, 5, 2 \times k_f$ respectively, corresponding to 116 and 2734 triangles used between $0.02 < k_i\,[h/\mathrm{Mpc}] < 0.12$. The largest bin size Δk_6 corresponds to the one used in the standard BOSS analysis performed by Gil-Marín et al. (2017).

For the same reason, we use the same range of scales: $k_{\rm min} = 0.03\,h/\mathrm{Mpc}$ and $k_{\rm max} = 0.09\,h/\mathrm{Mpc}$ for both power spectrum monopole and quadrupole, $k_{\rm min} = 0.02\,h/\mathrm{Mpc}$ and $k_{\rm max} = 0.12\,h/\mathrm{Mpc}$ for the bispectrum monopole.

The fiducial cosmology chosen for the analysis corresponds to a flat ΛCDM model similar to the one reported in Planck Collaboration et al. (2016) and recently in Planck Collaboration et al. (2018). In particular, we set $\Omega_{\rm m}(z=0) = 0.31$, $\Omega_{\rm b}(z=0) = 0.049$, $A_{\rm s} = 2.21 \times 10^{-9}$, $n_{\rm s} = 0.9624$, $h_0 = 0.6711$. As in Paper II, in order to compute the numerical derivatives of the data vector with respect to the model parameters, we fixed the fiducial value of the bias model parameters, the growth rate, and the amplitude of dark matter oscillations to the ones obtained by running a preliminary low-resolution MCMC ($b_1 = 2.5478$, $b_2 = 1.2127$, $f = 0.7202$, $\sigma_8 = 0.4722$).

Since for the range of scales considered (quasi-linear regime), the Fingers-of-God parameters for both power spectrum and bispectrum were compatible with zero, $\sigma_{{\rm B}_k}^{\rm FoG}$ and $\sigma_{{\rm P}_k}^{\rm FoG}$ have been set to zero. In Paper II, we tested that the choice of fiducial parameters used to compute the analytical covariance matrix and the derivatives of the mean of the data vector does not significantly influence the results of the compression.

3 DATA VECTOR

We use the estimators described in Gil-Marín et al. (2015) and Gil-Marín et al. (2017) to measure the power spectrum monopole and quadrupole together with the bispectrum monopole from the data and the galaxy catalogues. In this work, we constrain the model parameters using the joint data vector obtained by combining the power spectrum monopole and quadrupole with the bispectrum monopole.

Almost all the two-point (2pt) statistics signal is contained in the first two multipoles of the redshift space galaxy power spectrum, the monopole and the quadrupole ($\ell = 0, 2$). These can be found by integrating the galaxy power spectrum:

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

where $L_\ell(\mu)$ is the ℓ-order Legendre polynomial and $\mathrm{P}_{\rm g}^{(s)}(k, \mu)$ is the redshift space galaxy power spectrum defined in Paper II and originally in the appendix of Gil-Marín et al. (2014).

We adopt the effective model presented in Gil-Marín et al. (2014) for the redshift space galaxy bispectrum. This consists in the modification of the redshift space distortion kernels derived from perturbations theory (see the appendix of the paper above for the full expressions).

The monopole of the bispectrum is obtained by averaging all the possible orientations of a triangle configuration with respect to the line of sight. It can therefore be computed through the integration of the two angular coordinates:

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

where μ_i is the angle between the $\boldsymbol{k}_i$ vector and the line of sight. The angle ϕ 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 $\boldsymbol{k}_2$. More details are given in the appendix of Paper II.

4 NEW TRIANGLE GEOMETRICAL PARAMETRIZATION

We want to regroup the bispectrum data vector elements in bins defined by different parameters describing the triangle configurations. The idea underlying this procedure is that similar triangular shapes will result in similar sensitivity to the cosmological parameters. This is because the perturbation kernels depend, in particular, on the cosine of the angles between the sides of the triangle.

Given the three triangle sides (k_1, k_2, k_3) normally characterizing an element of the redshift space galaxy bispectrum monopole data vector, we define three new variables. The first is the square root of the area of the triangle, which we label $\aleph$ (”aleph”). It can be computed using Heron’s formula:

$$A = \sqrt{s(s-k_1)(s-k_2)(s-k_3)} \quad \Longrightarrow \quad \aleph \equiv \sqrt{A}, \qquad (3)$$

where $s = \frac{1}{2}(k_1 + k_2 + k_3)$ is the semiperimeter of the triangle. The $\aleph$ parameter keeps track of the physical scales probed by the triangle configuration. Therefore, $\aleph$ is a variable that encodes the information the two linear power spectra present in the bispectrum tree-level expression (see Paper I or II appendixes for the explicit expression).

The second variable which we use to characterize a triangle is the cosine of the largest angle,[1] $\daleth = \cos\psi_{\max}$ (pronounced ‘daleth’). This choice allows one to describe whether the triangle is acute or obtuse. If $\cos(\pi/3) = 1/2 > \daleth > 0$, the triangle is acute. In this case, either the three sides are all approximately the same or two of them are larger than a third one. If $-1 < \daleth < 0$, the triangle is obtuse. The triangle could then have either a side much larger than the other two (the one opposite to $\psi_{\max}$) or two sides of similar length with a third smaller one. In order to distinguish between the pair of possibilities described above, as a third variable we consider the ratio between the cosines of the intermediate and smallest angles, $\gimel = \cos\psi_{\mathrm{int}}/\cos\psi_{\min}$ (pronounced ‘gimel’). All the cosines can be computed using the cosine rule for a triangle

$$k_l^2 = k_m^2 + k_n^2 - 2k_m k_n \cos\psi_{mn}, \qquad (4)$$

where $\cos\psi_{mn}$ is the angle between the triangle sides k_m and k_n. The variables $\daleth$, $\gimel$ encode the geometrical information strongly affecting the value of the second-order perturbation theory kernel present in the bispectrum expression. These variables allow to regroup together triangle configurations returning similar kernel values because of the similar geometrical properties. Therefore, each triangle configuration can be described as a function of the three variables ($\aleph$, $\daleth$, $\gimel$) and the same is true for each bispectrum monopole data vector element

$$\mathrm{B}_{\mathrm{g}}^{(0)}(k_1, k_2, k_3) \quad \Longrightarrow \quad \mathrm{B}_{\mathrm{g}}^{(0)}(\aleph, \daleth, \gimel)\,. \qquad (5)$$

The vice-versa relation is also valid, each set ($\aleph$, $\daleth$, $\gimel$) corresponds to a triangle configuration described by a choice of the three sides (k_1, k_2, k_3). The compression consists in using large-enough bins for the new variables ($\aleph$, $\daleth$, $\gimel$) so that the bispectra of triangles with similar geometrical properties contribute to the same new data vector element. Once the coordinate conversion has been done for all the triangle configurations, the binning for the new coordinates can be defined by finding the minimum and maximum values for the new parameters ($\aleph$, $\daleth$, $\gimel$). Given a choice for the number of bins for each new coordinate ($n_\aleph$, $n_\daleth$, $n_\gimel$), the potential dimension of the new data vector is $n_\aleph \times n_\daleth \times n_\gimel$. However, as is the case when using the three sides (k_1, k_2, k_3) to describe the triangle, several combinations of ($\aleph$, $\daleth$, $\gimel$) actually do not satisfy the triangle inequalities, and therefore no triplet (k_1, k_2, k_3) will contribute to that particular bin. Moreover even if a particular combination of $(\aleph, \daleth, \gimel)_k$ does represent a triangle configuration, it is not certain that the triangle bin defined by $(\aleph, \daleth, \gimel)_k$ will contain modes since the original number of triangles in (k_1, k_2, k_3) coordinates was finite. The new data-vector $\boldsymbol{g}$ is obtained by averaging over all the bispectra in the non-empty triangle sets defined by different combinations of the coordinates ($\aleph$, $\daleth$, $\gimel$):

$$g_k(\aleph, \daleth, \gimel)_k = \frac{1}{N_i^{\mathrm{tr.}}} \sum_{j\,:\,(k_1,k_2,k_3)_j \in (\aleph,\daleth,\gimel)_k}^{N_i^{\mathrm{tr.}}} \mathrm{B}_{\mathrm{g}}^{(0)}(k_1, k_2, k_3)_j\,, \qquad (6)$$

where each new data vector element has been normalized by dividing by the number of triangles belonging to the same set defined by a particular combination of $(\aleph, \daleth, \gimel)_k$, $N_k^{\mathrm{tr.}}$.

5 NUMBER OF BINS: OPTIMAL CHOICE

For the construction of the new data vector, it is necessary to define how many bins will be used to divide the range of each parameter. In order to optimize the choice of these three numbers, ($n_\aleph$, $n_\daleth$, $n_\gimel$), we suggest the following procedure. The idea is to ‘sample’ the sensitivity of the new data vector to the considered model parameters for the different choices of ($n_\aleph$, $n_\daleth$, $n_\gimel$). The most straightforward way to do so is to consider the derivatives of the data vector model with respect to the parameters. These can be computed assuming a fiducial cosmology which, in our case, was described in Section 2.

In order to transform the derivatives of the standard bispectrum monopole data vector into the derivatives of the new one, it is sufficient to apply the same algorithm used to convert the bispectrum into $\boldsymbol{g}$ given in equation (6), because the transformation is linear. At this point, we have a list of $\boldsymbol{g}_{,i} = \partial\boldsymbol{g}/\partial\theta_i$ for all the elements of the model parameter vector $\boldsymbol{\theta}$. The target is to combine these vectors into a unique number expressing the sensitivity of the new data vector $\boldsymbol{g}$ for a certain choice of ($n_\aleph$, $n_\daleth$, $n_\gimel$). We call $N_{\boldsymbol{g}}$ the dimension of the new data vector $\boldsymbol{g}$ and $N_{\boldsymbol{\theta}}$ the number of model parameters. $N_{\boldsymbol{g}}$ is of course a function of the number of bins of the new coordinates, $N_{\boldsymbol{g}}(n_\aleph, n_\daleth, n_\gimel)$. For each of the model parameters θ_i and for a particular choice of the number of bins $(n_\aleph, n_\daleth, n_\gimel)_j$, we derive a single number defined as

$$\mathrm{S}_{ij} \equiv \sum_{k=1}^{N_{\boldsymbol{g}}(n_\aleph,n_\daleth,n_\gimel)_j} \frac{1}{N_k^{\mathrm{tr.}}} \left| \frac{\partial g_k}{\partial\theta_i} \right|. \qquad (7)$$

S_{ij} is a proxy for the sensitivity of the new data vector $\boldsymbol{g}$ defined for a particular choice of number of bins $(n_\aleph, n_\daleth, n_\gimel)_j$ with respect to variations of the model parameter θ_i. Notice that each term of the sum, before being added, is normalized by the number of triangles regrouped in the new bin defined by a set of coordinates $(\aleph, \daleth, \gimel)_k$.

[1] In this case, we mean the ‘interior’ angle of the triangles, which differs from the angles between k-vectors used in the perturbation theory kernels by a factor of π, since the sum of k-vectors must be equal to zero.

The next step consists of combining these proxies for all the model parameters. This in order to obtain a single number describing the overall sensitivity of $\boldsymbol{g}$ for a determinate choice of $(n_{\aleph}, n_{\daleth}, n_{\gimel})_j$. We then normalize each ith S_{ij} dividing by the maximum value of S_{ij} for all the possible $(n_{\aleph}, n_{\daleth}, n_{\gimel})_j$ combinations

$$\mathrm{s}_{ij} \equiv \frac{\mathrm{S}_{ij}}{\max\left[\mathrm{S}_{ij}\right]_{\forall j}}, \tag{8}$$

so that for all θ_i then $0 < \mathrm{s}_{ij} \le 1$. Finally, all the $N_{\boldsymbol{\theta}}$ s_{ij} for each $(n_{\aleph}, n_{\daleth}, n_{\gimel})_j$ combination can be merged into a unique number by doing

$$\bar{\mathrm{s}}_j \equiv \sum_{i=1}^{N_{\boldsymbol{\theta}}} \mathrm{s}_{ij}. \tag{9}$$

We consider $\bar{\mathrm{s}}_j$ as the proxy encoding the overall sensitivity to the model parameters of the new data vector $\boldsymbol{g}$, defined by a particular choice of the triplet $(n_{\aleph}, n_{\daleth}, n_{\gimel})_j$. Since we may want to limit the dimension of $\boldsymbol{g}$ in the algorithm, we include a condition setting $\bar{\mathrm{s}}_j = 0$ when $N_{\boldsymbol{g}}(n_{\aleph}, n_{\daleth}, n_{\gimel})_j \ge N_{\boldsymbol{g}}^{\max}$. The standard BOSS analysis bispectrum data vector, limited to the range of scales that we consider, has 116 triangles (Δk_6 binning case defined in Section 2). We use the measurements done for the Δk_2 binning case corresponding to 2734 triangles for the bispectrum monopole.

We consider two cases, $N_{\boldsymbol{g}}^{\max} = 117$ and $N_{\boldsymbol{g}}^{\max} = 196$, compressing the original bispectrum monopole by a factor of ∼23 and ∼14, respectively. The $N_{\boldsymbol{g}}^{\max} = 196$ is used to study the difference between MC-GC and the standard MCMC on the full data vector given by the Δk_5 binning of the triangle sides.

For $N_{\boldsymbol{g}}^{\max} = 117$, $\bar{\mathrm{s}}_j$ has been computed for all the $(n_{\aleph}, n_{\daleth}, n_{\gimel})_j$ combinations with $1 \le n_{\aleph}, n_{\daleth}, n_{\gimel} \le 25$. With these settings, we obtained the highest value for $\bar{\mathrm{s}}_j$ in the case of $(n_{\aleph} = 10,\ n_{\daleth} = 9,\ n_{\gimel} = 19)$ corresponding to a dimension $N_{\boldsymbol{g}}(10, 9, 19) = 115$. For $N_{\boldsymbol{g}}^{\max} = 196$, $\bar{\mathrm{s}}_j$ has been computed for all the $(n_{\aleph}, n_{\daleth}, n_{\gimel})_j$ combinations with $5 \le n_{\aleph}, n_{\daleth}, n_{\gimel} \le 30$. With these settings, we obtained the highest value for $\bar{\mathrm{s}}_j$ in the case of $(n_{\aleph} = 22,\ n_{\daleth} = 10,\ n_{\gimel} = 16)$ corresponding to a dimension $N_{\boldsymbol{g}}(22, 10, 16) = 194$. Fig. 1 shows the variation of $\bar{\mathrm{s}}_j$ as function of each number of bins for the $N_{\boldsymbol{g}}^{\max} = 117$ case, keeping the others fixed to the optimal value. In the last two columns of Table 1 and Table 2, we show that the difference between the mean of the 1D posterior distributions obtained for the two cases $N_{\boldsymbol{g}}^{\max} = 117$ and $N_{\boldsymbol{g}}^{\max} = 196$ is small and that improvement on parameter constraints are similar.

6 COMPRESSION PERFORMANCE

We can compare the results obtained via MC-GC (Δk_2 case) in terms of 1D and 2D the posterior distributions obtained via the standard MCMC sampling (Δk_6 and Δk_5 cases) and maximal compression (Δk_2 case). The comparison is shown in Fig. 2. Even if it does not need to analytically model the covariance matrix in order to compress the data vector, MC-GC produces a posterior distribution very close to the one given by the maximal compression method. The agreement is remarkable, especially considering that these compression methods are fairly independent of each other (they have in common only the use of the data vector derivatives). The precise values of the 1D 68 per cent confidence intervals and of the means of the distribution are reported in Tables 1 and 2.

It is important to notice the difference between the MCMC with 116, 195 triangles and the MC-GC results using 115 and 194 combinations of the original 2734 triangles, respectively. It is clear from both Table 2 and Fig. 2 that when the same number of data vector elements are considered, MC-GC produces much tighter constraints since it is able to exploit the constraining power of the original 2734 triangles.

The observed shift between MCMC results using 116 triangles and MC-GC/maximal compression using 2734 triangles is due to the strong degeneracy between the model parameters which is partially lifted when more triangle configurations are used in the data-vector.

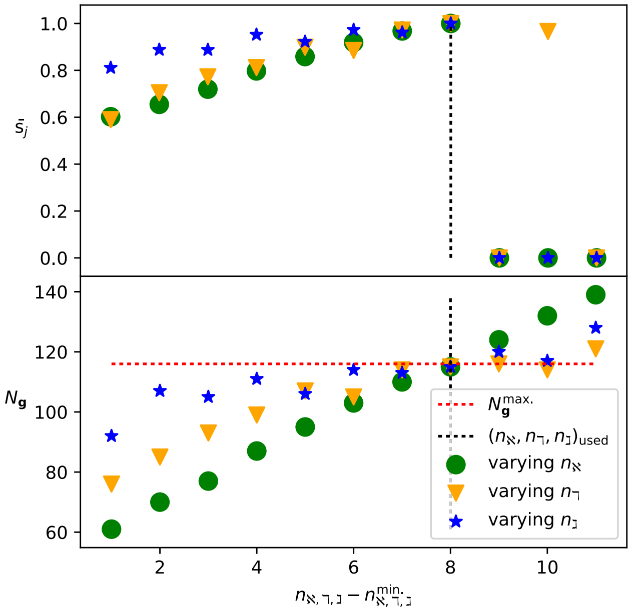

Figure 1. Variation of the parameter $\bar{\mathrm{s}}_j$ in the $N_{\boldsymbol{g}}^{\max} = 117$ case, used to choose the number of bins for the new parameters, and of the number of elements of the new data vector $N_{\boldsymbol{g}}$ as a function of $(n_{\aleph}, n_{\daleth}, n_{\gimel})$. $n^{\mathrm{min.}}_{\aleph,\daleth,\gimel}$ is a normalization on the x-axis used to show the same number of different configurations, obtained by varying one of the bins numbers $(n_{\aleph}, n_{\daleth}, n_{\gimel})$ while keeping the other fixed to the optimal value, on both left and right sides of the optimal set $(n_{\aleph}, n_{\daleth}, n_{\gimel})_{\mathrm{opt.}}$. In particular, for the case shown, we used $(n_{\aleph}^{\mathrm{min.}} = 3, n_{\daleth}^{\mathrm{min.}} = 2, n_{\gimel}^{\mathrm{min.}} = 12)$. The horizontal red line shows the imposed upper limit to the number of new data vector elements, $N_{\boldsymbol{g}}^{\mathrm{max.}}$. The vertical black line indicates the chosen set of $(n_{\aleph}, n_{\daleth}, n_{\gimel})$ for which s_j was the highest for $N_{\boldsymbol{g}} < N_{\boldsymbol{g}}^{\mathrm{max.}}$.

Table 1. Best-fit parameters. Mean values of the posterior distributions and 68% credible intervals for the MCMC sampling on the full data vector, the 'maximal' and the MC-GC compression methods. The largest k-binning Δk_6, the size used in the BOSS analysis, corresponds to the lowest number of triangles (116). For it, we show the best-fit parameters obtained via MCMC sampling using the full data vector. For the thinnest binning Δk_2, corresponding to the highest number of triangles (2734), we compare the three compression methods. The results shown for the MC-GC method are relative to the cases with $N_{\boldsymbol{g}}^{\mathrm{max.}} = 196$ (orange) and $N_{\boldsymbol{g}}^{\mathrm{max.}} = 117$ (yellow). The observed shift in the mean values as a function of the number of considered triangles is due to the strong degeneracy present between the model parameters which gets partially lifted when, due to the compression, more triangle configurations are considered.

	Δk_6		Δk_2	
	MCMC	Max. Comp.	MC-GC	
b_1	2.41 ± 0.22	2.33 ± 0.14	2.25 ± 0.15	2.22 ± 0.14
b_2	1.00 ± 0.40	0.72 ± 0.22	0.64 ± 0.25	0.68 ± 0.21
f	0.69 ± 0.08	0.63 ± 0.06	0.64 ± 0.06	0.65 ± 0.06
σ_8	0.50 ± 0.04	0.53 ± 0.03	0.52 ± 0.04	0.53 ± 0.03

Table 2. Improvement in parameter constraints showing the relative change of the 68% credible intervals for the Δk_2 k-binning case with respect to the Δk_6 and Δk_5 (green) cases. For the MC-GC method, we show two cases, for $N_{\boldsymbol{g}}^{\mathrm{max.}} = 196$ (orange) and $N_{\boldsymbol{g}}^{\mathrm{max.}} = 117$ (yellow). MC-GC obtains very similar improvements, in terms of tighter parameter constraints, to the ones obtained via maximal compression. Notice the substantial difference in parameter constraint improvements between the standard MCMC case using 195 triangles and the MC-GC case recombining 2734 triangles into 194 new data vector elements.

	$\Delta\theta^{\mathrm{mc}}_{\Delta k_6}$	$\frac{\Delta\theta^{\mathrm{comp.}} - \Delta\theta^{\mathrm{mc}}_{\Delta k_6}}{\Delta\theta^{\mathrm{mc}}_{\Delta k_6}}$ [%]			
	MCMC	MCMC	Max. Comp.	MC-GC	
	$N_{\mathrm{tr}} = 116$	$N_{\mathrm{tr}} = 195$	$N_{\mathrm{el.}} = 4$	$N_{\boldsymbol{g}} = 115$	$N_{\boldsymbol{g}} = 194$
Δb_1	0.22	−23.9	−37.1	−30.5	−38.8
Δb_2	0.40	−34.9	−46.1	−39.3	−48.6
Δf	0.08	−18.5	−27.8	−24.4	−29.1
$\Delta\sigma_8$	0.04	−12.9	−22.8	−16.6	−22.0
$\left\langle \frac{\Delta\theta - \Delta\theta^{\mathrm{mc}}_{\Delta k_6}}{\Delta\theta^{\mathrm{mc}}_{\Delta k_6}} [\%] \right\rangle$		-22.5	-33.5	-27.7	-34.1

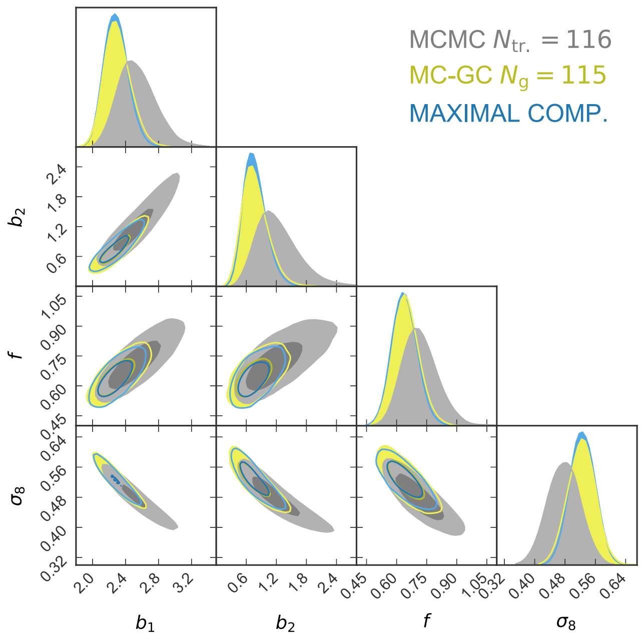

Figure 2. Compression performance: 2-D 68 per cent and 95 per cent credible regions are shown, respectively, for the Δk_6 standard MCMC sampling (MCMC in grey, 116 triangles), Δk_2 maximal compression (MCMC on the compressed data vector in blue, obtained using the maximal compression method presented in Gualdi et al. (2018a) on the original 2734 triangles) and Δk_2 geometrical compression (MC-GC in yellow, 2734 triangles) cases. The MC-GC case shown is obtained by imposing that the dimension of the compressed data vector satisfies $N_{\boldsymbol{g}}^{\mathrm{max.}} = 116$. The agreement between maximal compression and MC-GC posterior distributions is remarkable. Without the need of an analytical modelling of the covariance matrix, MC-GC recovers very close posterior distributions to the ones derived using the maximal compression method. The observed shift between MCMC results using 116 triangles and MC-GC/maximal compression using 2734 triangles is due to the strong degeneracy between the model parameters which is partially lifted when more triangle configurations are used. In particular, the shift happens along the degeneration direction of b_1, b_2 and f with σ_8 and as described in Gualdi et al. (2018b) and it may have a statistical origin.

In terms of time and computing resources, MC-GC is equivalent to standard MCMC sampling and maximal compression method (details given in Paper II).

7 CONCLUSIONS

The new compression method presented in this work consists in binning together bispectra evaluated at sets of wavenumbers forming closed triangles with similar geometrical properties: the area, the cosine of the largest angle, and the ratio between the cosines of the remaining two angles.

The advantage of the geometrical compression (MC-GC) technique, with respect to maximal compression methods, introduced in Gualdi et al. (2018a) and applied to BOSS data in Gualdi et al. (2018b), is that it does not require an analytical modelling of the covariance matrix. This is due to the fact that MC-GC is based on the similarities between the geometrical properties of different triangle configurations and not on their bispectrum values covariance. In terms of resources and computing time required, these are approximately the same as for the maximal compression method (see Paper II for details), i.e. the time taken by the geometrical compression step is negligible. The MC-GC compression is not ‘maximal’ as the ones presented in Gualdi et al. (2018a). We compressed using MC-GC the bispectrum of 2734 triangle configurations into data vectors up to ∼23 times shorter.

By compressing the data vector using the geometrical compression before running the MCMC sampling, we improved BOSS constraints, reducing the 68 per cent credible intervals for the inferred parameters (b_1, b_2, f, σ_8) by (−39 per cent, −49 per cent, −29 per cent, −22 per cent), respectively.

Future work will include the development of extensions of the MC-GC method to higher order statistics, like the trispectrum and tetraspectrum, always using geometrical properties of the k-vectors’ configurations. Moreover, we are interested in applying MC-GC to weak-lensing and 21-cm emission line 3pt statistics. Given its immediate and straightforward applicability, we hope that MC-GC will become a standard procedure for future data sets to study the bispectra and 3pt functions of the cosmological fields of interest. Another interesting point would be to study whether it is possible to efficiently compress 3pt statistics using different geometrical properties of the triangle configurations than the ones used here.

ACKNOWLEDGEMENTS

DG is supported by the Perren and the IMPACT studentships. HGM is supported by Labex ILP (reference ANR-10-LABX-63) part of the Idex SUPER, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme Investissements d’avenir under the reference ANR-11-IDEX-0004-02.e. MM. acknowledges funding from STFC Consolidated Grants RG84196 and RG70655 LEAG/506 and has received funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 6655919. O.L. acknowledges support from a European Research Council Advanced Grant FP7/291329. C (Kernighan 1988) and PYTHON 2.7 (Rossum 1995) have been used together with many packages like IPYTHON (Perez & Granger 2007), Numpy (van der Walt, Colbert & Varoquaux 2011), Scipy (Jones et al. 2001) and Matplotlib (Hunter 2007). The corner plots have been realized using PYGTC developed by Bocquet & Carter (2016)

REFERENCES

Alam S. et al., 2017, MNRAS, 470, 2617
Bocquet S., Carter F. W., 2016, J. Open Source Softw., 1
Child H. L., Takada M., Nishimichi T., Sunayama T., Slepian Z., Habib S., Heitmann K., 2018, Phys. Rev. D, 98, 123521
Dawson K. S. et al., 2013, AJ, 145, 10
Eisenstein D. J. et al., 2011, AJ, 142, 72
Fu L. et al., 2014, MNRAS, 441, 2725
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
Gil-Marín H., Wagner C., Noreña J., Verde L., Percival W., 2014, J. Cosmol. Astropart. Phys., 12, 029
Gualdi D., Gil-Marín H., Schuhmann R. L., Manera M., Joachimi B., Lahav O., 2018b, MNRAS
Gualdi D., Manera M., Joachimi B., Lahav O., 2018a, MNRAS, 476, 4045
Heavens A. F., Jimenez R., Lahav O., 2000, MNRAS, 317, 965
Hoffmann K., Mao Y., Mo H., Wandelt B. D., 2018, preprint (arXiv:1802.02578)
Hunter J. D., 2007, Comput. Sci. Engg., 9, 90
Joachimi B., Shi X., Schneider P., 2009, A&A, 508, 1193
Jones E., Oliphant T., Peterson P. et al., 2001, SciPy: Open source scientific tools for, http://www.scipy.org/
Kayo I., Takada M., 2013, preprint (arXiv:1306.4684)
Kayo I., Takada M., Jain B., 2013, MNRAS, 429, 344
Kernighan B. W., 1988, The C Programming Language. 2nd edn. Prentice Hall Professional Technical Reference
Kilbinger M., Schneider P., 2005, A&A, 442, 69
Kitaura F.-S. et al., 2016, MNRAS, 456, 4156
Pearson D. W., Samushia L., 2018, MNRAS, 478, 4500
Perez F., Granger B. E., 2007, Comput. Sci. Eng., 9, 21
Planck Collaboration et al., 2016, A&A, 594, A13
Planck Collaboration et al., 2018, preprint (arXiv:1807.06209)
Rossum G., 1995, Technical report, Python Reference Manual. Amsterdam, The Netherlands
Schneider P., Kilbinger M., Lombardi M., 2005, A&A, 431, 9
Slepian Z. et al., 2017a, MNRAS, 468, 1070
Slepian Z. et al., 2017b, MNRAS, 469, 1738
Takada M., Jain B., 2004, MNRAS, 348, 897
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
van der Walt S., Colbert S. C., Varoquaux G., 2011, CoRR, 13, 22
Yankelevich V., Porciani C., 2018, preprint (arXiv:1807.07076)

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