Fitting covariance matrix models to simulations

Alessandra Fumagalli, a,b,c,d,* Matteo Biagetti, c,e,b,d Alex Saro, c,e,b,d Emiliano Sefusatti, b,c,d Anže Slosar, f Pierluigi Monaco a,b,c,d and Alfonso Veropalumbo g,h

a Astronomy Unit, Dipartimento di Fisica, Università di Trieste, via Tiepolo 11, 34143 Trieste, Italy
b Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Trieste, via Tiepolo 11, 34143 Trieste, Italy
c Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy
d Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, via Valerio 2, 34127 Trieste, Italy
e SISSA - International School for Advanced Studies, via Bonomea 265, 34136 Trieste, Italy
f Physics Department, Brookhaven National Laboratory, Upton NY 11973, U.S.A.
g Dipartimento di Fisica, Università degli Studi Roma Tre, via della Vasca Navale 84, 00146 Roma, Italy
h Istituto Nazionale di Fisica Nucleare, Sezione di Roma Tre, via della Vasca Navale 84, 00146 Roma, Italy

E-mail: alessandra.fumagalli@inaf.it, matteo.biagetti@sissa.it, alexandro.saro@inaf.it, emiliano.sefusatti@inaf.it, anze@bnl.gov, pierluigi.monaco@inaf.it, alfonso.veropalumbo@uniroma3.it

Received June 24, 2022
Revised October 15, 2022
Accepted November 18, 2022
Published December 15, 2022

* Corresponding author.
Abstract. Data analysis in cosmology requires reliable covariance matrices. Covariance matrices derived from numerical simulations often require a very large number of realizations to be accurate. When a theoretical model for the covariance matrix exists, the parameters of the model can often be fit with many fewer simulations. We write a likelihood-based method for performing such a fit. We demonstrate how a model covariance matrix can be tested by examining the appropriate $\chi^2$ distributions from simulations. We show that if model covariance has amplitude freedom, the expectation value of second moment of $\chi^2$ distribution with a wrong covariance matrix will always be larger than one using the true covariance matrix. By combining these steps together, we provide a way of producing reliable covariances without ever requiring running a large number of simulations. We demonstrate our method on two examples. First, we measure the two-point correlation function of halos from a large set of 10000 mock halo catalogs. We build a model covariance with 2 free parameters, which we fit using our procedure. The resulting best-fit model covariance obtained from just 100 simulation realizations proves to be as reliable as the numerical covariance matrix built from the full 10000 set. We also test our method on a setup where the covariance matrix is large by measuring the halo bispectrum for thousands of triangles for the same set of mocks. We build a block diagonal model covariance with 2 free parameters as an improvement over the diagonal Gaussian covariance. Our model covariance passes the $\chi^2$ test only partially in this case, signaling that the model is insufficient even using free parameters, but significantly improves over the Gaussian one.

Keywords: Bayesian reasoning, cosmological parameters from LSS, cosmological simulations

ArXiv ePrint: 2206.05191
1 Introduction

Modern studies of the large-scale structure rely on mapping the Universe using a variety of tracers and using the statistical properties of their distribution to put constraints on cosmological models. In most analyses, data are compressed into summary statistics that capture certain properties of cosmological perturbations, with examples including traditional measurements like power spectra and higher-order correlations [1–6]. Theoretical and modeling progress have finally made it possible to include statistics beyond the power spectrum, like the bispectrum, in cosmological analyses of galaxy surveys [4, 7, 8]. Large collaborations, such as Euclid, are already preparing to exploit these statistics. More complex observables have also been explored, such as various topological measures [9–12], synthetic tracers such as voids [13] or galaxies marked by local density, morphology or stellar mass [14–16], etc. This implies that data vectors will soon have hundreds of components. In all cases, however, it is necessary to obtain a robust estimate of the uncertainty affecting the measurement in order to derive solid constrains on the cosmological model.

Relying on a numerical estimate based on simulations and mocks has been so far the most common approach, as it allows including more easily subtle observational effects (see, e.g. [17–19]). For experiments with complex selection functions, one might be able to generate
$O(10)$ or $O(100)$ realistic realizations, but it is unlikely that 10000 would be feasible. For this purpose, several approximate, but efficient, methods to produce galaxy mock catalogs have been proposed over the past decade (see [20, 21] for a review and a recent proposal or [22–24] for comparisons in terms of the predicted covariance). On the other hand, upcoming galaxy surveys will pose a challenge to this method as mock galaxy catalogs will require high-resolution simulations over very large volumes and covering a very large parameter space, making their production extremely expensive when feasible at all. In this respect, over the last years, several works addressed the problem of a poor numerical estimate of the precision matrix [25–29], while various strategies to reduce the number of required realizations have been explored [30–36].

An alternative approach based on analytical models for the covariance matrix is also gaining attention [37–42]. Analytical models can be affected by systematic theoretical errors and they still need to be validated against simulations. In particular, often such models depend on nuisance parameters to be determined by proper fits to numerical results.

In this paper, we discuss a hybrid approach, which builds up on recent techniques, with the goal of obtaining covariance matrices that are simultaneously reliable and based on few simulations. There are two insights our result is based on. The first one is that the covariance matrix estimation is, fundamentally, an estimate of two-point correlations of realizations of the data vector. This is a problem that has been solved many times in various fields, and for which an exact likelihood can be written and evaluated. It can be shown that the naively calculated numerical matrix contains all the information needed to evaluate the likelihood of a theoretical model of the same covariance matrix. Previous work has exploited this insight to fit a model covariance to mock simulations for the Baryon Acoustic Oscillations (BAO) analysis of SDSS DR7 [43] and the two-point correlation function [44], the bispectrum [45] and the power spectrum [46] of BOSS data. In particular, [44] and [46] have used these fits to reduce the number of mocks needed to build a reliable covariance.

The second insight is that, ultimately, any covariance matrix is used to compare some data with a model. The fundamental quantity of interest is therefore the $\chi^2$ distribution. A good covariance matrix produces $\chi^2$–distributed values when presented with realizations of data vectors. A biased or inappropriate covariance matrix produces $\chi^2$ values that are not drawn from the correct distribution. This then tells us how to test a given covariance matrix. Applications of this idea have been investigated recently [47, 48]. Essentially, we propose to combine these two insights into a single algorithm to produce reliable and cheap covariance matrices. Assuming we have 50 realizations of a data vector of size 100, therefore corresponding to a covariance matrix of size $100 \times 100$, we are unable to estimate the covariance with precision, but we can both i) test candidate covariance matrices (i.e. 50 $\chi^2$ values are sufficient to show compatibility with the correct distribution) and ii) fit a few parameter model covariance matrices.

This paper is structured as follows. In section 2 we present our method. We develop two different tests to illustrate our method and show the results of the tests in sections 4 and 5. The last section presents conclusions and outlines future work.

2 Covariance matrix estimation is covariance estimation

Let us assume we are interested in the covariance matrix for a quantity that can be represented by an $N$ dimensional vector. An example might be a measurement of the power spectrum of galaxies in $N$ bins, or the measurement of a bispectrum, or both.
Ultimately, we want to run a Bayesian analysis, comparing a measurement vector \( \mathbf{m} \) (of size \( N \)) with a theory prediction \( \mathbf{t}(\theta_m) \), where \( \theta_m \) are the parameters of the model for the expectation value of \( \mathbf{m} \). Assuming a Gaussian likelihood, we have

\[
P(\theta_m|\mathbf{m}) \propto P(\mathbf{m}|\theta_m) = (2\pi)^{-N/2} |C|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{m} - \mathbf{t}(\theta_m) \right\}^T C^{-1} \left[ \mathbf{m} - \mathbf{t}(\theta_m) \right]. \tag{2.1}
\]

In principle, the matrix \( C \) depends as well on the parameters \( \theta_m \). If such dependence is significant, our method can be straightforwardly extended to accommodate it as well. For now, we assume that this is not the case.

Suppose we have \( N_{\text{sims}} \) simulations providing as many realizations of the measurements \( \mathbf{m}_i \), where \( i = 1 \ldots N_{\text{sims}} \). We assume that these simulations are drawn from the same underlying theory and that the output varies only because of a different realization of the cosmic structure and noise. In practical terms, they are identical runs except for the random seed. By construction, the ensemble mean value of those realizations is \( \langle m_i \rangle \equiv \mathbf{m}(\theta_{m,\text{sim}}) \) where \( \theta_{m,\text{sim}} \) are the fiducial values of the parameters adopted to create the simulations. These are assumed to be the same for all realizations. Defining \( \mathbf{d}_i = \mathbf{m}_i - \mathbf{m}(\theta_{m,\text{sim}}) \), we see that the \( \mathbf{d}_i \) are normally distributed around zero with covariance \( C \).

We can write the naive matrix estimator

\[
C_n = \frac{1}{N_{\text{sims}}} \sum_{i=1}^{N_{\text{sims}}} \mathbf{d}_i \mathbf{d}_i^T. \tag{2.2}
\]

We refer to this covariance matrix as the “numerical” covariance matrix.

Now, let us assume that we have a theoretical model for \( C \) depending on some model parameters \( \theta \), i.e. \( C = C(\theta) \). That is, for a given set of \( \theta \), we can predict all values of \( C \). Note that parameters \( \theta \) cannot contain \( \theta_m \), since we have just assumed that these do not affect \( C \).

The model for \( C \) can be either physical, based on theoretical expectation about \( C \), but it can also be purely phenomenological, e.g. assuming \( C \) has a Toeplitz form, or that the off-diagonal terms beyond the second diagonal vanish. In any case, we can now use the Bayesian theorem to write a likelihood for the covariance parameters \( \theta \) as:

\[
P(\theta|\mathbf{d}) \propto P(\mathbf{d}|\theta)\Pi(\theta), \tag{2.3}
\]

where we can put any prior information in \( \Pi(\theta) \) and which we assume to be unity and where

\[
L = P(\mathbf{d}|\theta) \propto \prod_{i=1}^{N_{\text{sims}}} |C(\theta)|^{-1/2} \exp \left[ -\frac{1}{2} \mathbf{d}_i^T C^{-1}(\theta) \mathbf{d}_i \right]. \tag{2.4}
\]

For the classically educated cosmologist, this equation looks quite familiar. It is the equation representing the likelihood of a Gaussian field. The log-likelihood function \( L \equiv \log L \) equals up to a constant to

\[
L(\mathbf{d}|\theta) = -\frac{N_{\text{sims}}}{2} \log |C(\theta)| - \frac{1}{2} \sum_i \mathbf{d}_i^T C^{-1}(\theta) \mathbf{d}_i = -\frac{N_{\text{sims}}}{2} \left[ \log |C(\theta)| + \text{Tr}(C^{-1}(\theta)C_n) \right]. \tag{2.5}
\]

\(^1\)Note that in this paper, as is often the practice, we subtract the mean of simulation results rather than the theoretical true model. For normally distributed residuals and an accurate theoretical model this is equivalent given a sufficient number of simulations. If the difference is significant, one has a bigger fish to fry first before worrying about the covariance matrix.
This is the form of the equation that we use in what follows. This equation appears in this form already in [44, 45], but in the context of very concrete models for the covariance matrix. In this paper, we stress that this form is general and that it can be applied to any model for covariance matrix and for observables beyond 2-point correlation function, as we will do in section 5. The most interesting aspect is that the likelihood can be rewritten in a form that depends only on the numerical covariance matrix $C_n$. In other words, we can compress the results of $N_{\text{sims}}$ simulations into a single $N \times N$ matrix. If $N_{\text{sims}} > N$ this form offers useful information compression. It is also true if $N_{\text{sims}} < N$, i.e. $C_n$ can in principle be even non-invertible and still contain all the available information at the 2-point level from the simulation suite. Given a sufficient number of simulations that the uncertainty on $\theta$ is negligible compared to the measurement noise on $m$, it suffices to find the maximum likelihood point in the $\theta$ space and use the resulting matrix.

In what follows, we demonstrate this technique in practice with two examples.

### 3 Goodness of fit for covariance matrix model

After deriving the best-fit parameters for a given model of the covariance matrix, we would like to determine if the result is indeed a useful covariance matrix. Note that this is not a model comparison exercise, but a problem of goodness-of-fit: is the resulting covariance matrix actually fit for purpose?

A common use of the covariance matrix is the evaluation of the $\chi^2$ for a cosmological model likelihood. Under the assumption of Gaussianity, a good covariance matrix is the one providing correctly distributed $\chi^2$ values. A simple test on the inverse can be done by verifying that the residuals $d_i = m_i - m(\theta_{m, \text{sim}})$ are $\chi^2$–distributed with the right number of degrees of freedom,

$$
\chi^2_{th,i} = (m_i - m(\theta_{m, \text{sim}})) C^{-1} [m_i - m(\theta_{m, \text{sim}})]^T.
$$

While one would ideally have a separate set of simulations, this test can also be performed on the same simulations that were used to infer the parameters. This is equivalent to fitting a theory to data and then using the same data to check the resulting $\chi^2$ without subtracting the model degrees of freedom. This is a valid procedure where the number of simulations is much larger than the number of free parameters in the theory. When this condition is violated, the only safe thing to do is to split simulations into a “training” and “testing” subsets.

We will now consider how do the moments of $\chi^2$ distribution respond to being tested with a wrong covariance matrix. First consider a simple model where we simply fit for the covariance matrix amplitude, i.e.

$$
C(\theta_A) = \theta_A C_0.
$$

At this point, we make no claims of whether $C_0$ is a good or poor approximation of the true covariance matrix, it is simply a matrix. We have $|C(\theta_A)| = \theta_A^N |C_0|$ and $C^{-1}(\theta_A) = \theta_A^{-1} C_0^{-1}$.

Plugging these expressions into equation (2.5) we find that the maximum likelihood point is given by

$$
\theta_A = \frac{\text{Tr}(C_0^{-1} C_n)}{N}.
$$
For the simulation realization \( i \), the \( \chi^2 \) is given by \( \chi^2_i = d^T_i C^{-1} d \) (for some covariance matrix \( C \)) and so the mean over the set of simulations is given by

\[
\text{Mean } \chi^2 = \frac{1}{N_{\text{sims}}} \sum_{i=1}^{N_{\text{sims}}} \chi^2_i = \text{Tr}(C^{-1}) = N, \tag{3.4}
\]

\[
\text{Var } \chi^2 = \frac{1}{N_{\text{sims}}} \sum_{i=1}^{N_{\text{sims}}} (\chi^2_i)^2 - \left( \text{Mean } \chi^2 \right)^2 = \frac{1}{N_{\text{sims}}} \sum_i (d^T_i C^{-1} d)^2 - N^2, \tag{3.5}
\]

The last equality of equation (3.4) comes from using solution of equation (3.3) and doing some straightforward manipulation. Even if \( C_0 \) is a completely wrong, the terms containing trace of \( C_0^{-1} C_n \) cancel exactly and so one always has \( \text{Mean } \chi^2 = N \). In other words, if our model for covariance matrix has a freedom to adjust the amplitude, then the maximum likelihood is such that the first moment of \( \chi^2 \) matches the theoretical expectation. Note that in general, our model for \( C \) will have many more parameters, but as long as there is a subspace of the model which corresponds to a simple amplitude rescaling, this statement will be true.

There is no such simplification for variance. In this case we can calculate the expectation value and find after some simplifications that:

\[
\langle \text{Var } \chi^2 \rangle = 2 \text{Tr}(C^{-1} C_{\text{true}} C^{-1}(\theta) C_{\text{true}}), \tag{3.6}
\]

where \( C_{\text{true}} \) is the true covariance matrix, i.e. the one from which vectors \( d \) are drawn. If \( C(\theta) = C_{\text{true}} \), i.e. if our model covariance matrix is indeed true, we find the standard moment of \( \chi^2 \) distribution, i.e. \( \langle \text{Var } \chi^2 \rangle = 2 \text{Tr}(I) = 2N \).

Next let us assume \( C(\theta) \) is different from \( C_{\text{true}} \), so that its inverse is given by \( C^{-1}(\theta) = C_{\text{true}}^{-1} + X \). Additionally, let us assume that the first moment is correctly recovered as will always happen when the model has the freedom to rescale the matrix. In this case we find that

\[
\text{Tr}(C_{\text{true}} X) = \text{Tr}(C_{\text{true}} C^{-1}(\theta)) - \text{Tr}(C_{\text{true}} C_{\text{true}}^{-1}) = \langle \chi^2 \rangle - N = 0. \tag{3.7}
\]

Therefore

\[
\langle \text{Var } \chi^2 \rangle = 2N + 2 \text{Tr}(C_{\text{true}} X C_{\text{true}} X) > 2N \tag{3.8}
\]

To show the second line inequality, we note that \( C_{\text{true}} \) is a positive definite matrix and we can always rotate it into the frame \( C_{\text{true}}' \) where it is diagonal with its eigen values \( \lambda_i \) on the diagonal. In this frame \( X \) becomes \( X' \) and the second term becomes \( 2 \text{Tr}(C_{\text{true}} X' C_{\text{true}} X') = \sum_{ij} \lambda_i \lambda_j X_{ij}^2 > 0 \).

Finally, let’s consider the error on the \( \chi^2 \) on a given data vector, which appears as a result of using the wrong covariance matrix. For a given \( d \)

\[
\Delta \chi^2(d) = d^T C(\theta)^{-1} d - d^T C^{-1} d = d^T X d \tag{3.9}
\]

Using the same assumption as above it is easy to show that the first and second moment of this quantity are

\[
\langle \Delta \chi^2 \rangle = 0 \tag{3.10}
\]

\[
\langle \text{Var } \Delta \chi^2 \rangle = 2 \text{Tr}(C X C) \tag{3.11}
\]
In other words, for data vectors drawn from the true covariance matrix, using the wrong covariance matrix produces additional scatter around true $\chi^2$ values. This additional scatter has zero mean and variance given by eq. (3.11). Since variance add, this results in a distribution of $\chi^2$ values that is broader by the same amount.

To recap, we have shown three simple but powerful results. If the model covariance matrix has freedom to vary in amplitude, the maximum likelihood will adjust its value so that the expectation value of $\chi^2$ is the correct value. In that limit, the values of individual $\chi^2$ computed with a wrong covariance matrix scatter around their true values: some of them are lower and some of them are higher than what they should be. Because variances add, this results in a second moment of $\chi^2$ distribution that is larger that $2N$, with equality holding when $C(\theta)$ is the correct covariance matrix.

This gives us a direct handle on accuracy of covariance matrix. If we are unwilling to tolerate more than $\Delta \chi^2 = 1$ in $\chi^2$ error, then we should really find a model in which both the first and the second moment are reproduced to this accuracy. In practice, we might find that a significantly more relaxed $\chi^2$ can still produce essentially unchanged constraints.

4 Test 1: two-point correlation function

For the first test of our method, we consider the two-point correlation function as observable, measured from a very large number $N_{\text{sims}} = 10000$ of mock halo catalogs. Let us assume as original problem a test of the two-point correlation function model based on the correct recovery of the cosmological parameters of the simulations. For the sake of simplicity, we only consider $\Omega_m$ and $\sigma_8$. Such test requires the knowledge of the covariance matrix. Given the large number of mocks, we can build a reliable numerical covariance matrix, $C_n$. Using the method presented in section 2, we show how we can obtain an equally reliable covariance matrix using only a fraction of the $N_{\text{sims}}$ simulations available.

4.1 The two-point correlation function and its covariance

The basic ingredients of this test involve a model for the two-point correlation function of halos corresponding to $t(\theta_m)$ in the previous section and the model for covariance matrix corresponding to $C(\theta)$.

For the purpose of testing our covariance fitting technique, we take a simple linear model for the halo power spectrum including shot noise,

$$P_h(k) = b^2 P_m(k) + \frac{1 + \alpha}{\bar{n}},$$

where $b$ is the linear bias, $P_m$ is the linear matter power spectrum, $\bar{n}$ is the halo number density and $\alpha$ parametrizes deviations from Poissonian shot noise, $1/\bar{n}$. The corresponding two-point correlation function is given by,

$$\xi_h(r) = b^2 \xi_m(r),$$

where $\xi_m$ is the Fourier transform of $P_m$. When comparing to simulations, we fit the linear bias to the measurements of the two-point correlation function. In order to compute the covariance for eq. (4.2), we Fourier transform the leading\(^2\) term of the power spectrum.

\(^2\)The full covariance would include a trispectrum term [49], which we assume to be negligible [50].
covariance

\[ C(b, \alpha) \simeq \frac{2}{V} \int \frac{\mathrm{d}k}{2\pi^2} \left( b^2 P_m(k) + \frac{1+\alpha}{n} \right)^2 W_i(k)W_j(k), \]

where \( V \) is the box volume and \( W_i(k) \) is the Fourier transform of the i-th radial shell

\[ W_i(k) = \int \frac{\mathrm{d}^3r}{V_i} j_0(kr) = \frac{r^3_i}{r^3_i+} W_{\text{th}}(kr_{i,+}) - \frac{r^3_i}{r^3_i-} W_{\text{th}}(kr_{i,-}) \]

where \( V_i = \frac{4\pi}{3} \left( r^3_{i,+} - r^3_{i,-} \right) \) is the shell volume, \( j_0(kr) \) is the zero-th order spherical Bessel function, and \( W_{\text{th}}(kr) \) is the Fourier-transform of the top-hat window function. Our model covariance therefore depends on two parameters \( \theta = \{b, \alpha\} \). A simple choice for these parameters would be to use the bias we fit to the two-point correlation function measurements \( b \equiv \hat{b} \) and \( \alpha = 0 \).

### 4.2 Description of the data

We take advantage of a very large set of \( N_{\text{sims}} = 10000 \) mock halo catalogs obtained with the PINOCCHIO code [51–53]. The code provides an approximate dark matter halo distribution based on particle displacements computed from 3rd order Lagrangian Perturbation Theory and a halo identification from ellipsoidal collapse. In comparison with full N-body numerical simulations the latest version of the code is able to recover the halo mass function and linear halo bias of Friends-of-Friend halos with an accuracy respectively within 5% and 10% of the full N-body.

The mock catalogs are built from \( 1000^3 \) dark matter particles in a cubic box of side \( L = 1500 h^{-1} \) Mpc, reproducing the set-up and adopting the cosmology of the Minerva N-body simulations [50]. The mass threshold is defined requiring that the large-scale amplitude of the PINOCCHIO catalogs power spectrum (including shot-noise) matches the amplitude of the same power spectrum measured in the Minerva simulations catalogs, the latter defined by a minimal mass of \( M \simeq 1.12 \times 10^{13} h^{-1} \text{M}_\odot \). We refer the reader to [54] for a more detailed description of the mock halo catalog construction, that allows to reproduce the power spectrum and bispectrum variance within 10% [54, 55]. The 2-point correlation function is measured using the standard Landy-Szalay estimator [56] as implemented in the CosmoBolognaLib code [57], adapted, however, to account for the box periodic boundary conditions in the pair counting.

The numerical covariance is estimated directly from the PINOCCHIO mocks as

\[ C_n(r_i, r_j) = \langle (\hat{\xi}_h(r_i) - \bar{\xi}_h(r_i))(\hat{\xi}_h(r_j) - \bar{\xi}_h(r_j)) \rangle, \]

where \( \hat{\xi}_h(r_i) \) is the measurement from a single mock and \( \bar{\xi}_h \) is the mean value, and the brackets indicate the average over the \( N_{\text{sims}} = 10000 \) mocks. The numerical covariance obtained from eq. (4.5) has been compared with full simulations results, showing good agreement (but with a different definition of the mass threshold) in [22].

We should notice that the covariance model above does not account for the specific geometry of our distribution, that is the effect of the exact number of pairs in the case of a box with periodic boundary conditions [58, 59]. We therefore expect that the recovered value of the covariance model parameters accounts in part also for this neglected effect.
4.3 Results

In this section we outline the results of this test. The goal is to show that our model covariance is as reliable as the numerical one obtained from a very large set of simulations. The final test is therefore to demonstrate that cosmological parameters, in our case $\Omega_m$ and $\sigma_8$, are recovered with the correct value and uncertainty using the model covariance. The procedure can be summarized as follows:

1. Build a set of numerical covariance matrices using eq. (4.5) for varying number of simulations $N_{\text{sims}} = 10000, 1000, 100, 50, 30$.

2. Build a model covariance with free parameters, following eq. (4.3). The free parameters are the linear bias $b$ and the deviation from Poisson shot noise, $\alpha$.

3. Maximize the likelihood of eq. (2.5) varying the free parameters, using the sets of $C_n$ from point 1) to get a best-fit model covariance $C(b_{\text{fit}}, \alpha_{\text{fit}})$.

4. Verify the reliability of the model covariance matrix, and of all the numerical covariance matrices, for varying number of simulations.

We proceed then with details on the maximization procedure and the check of reliability of the model covariance. We also show results obtained by fixing the covariance parameters to the fit to the two-point correlation function, $C_\xi$.

4.3.1 Maximizing the likelihood

We fit our model covariance $C(b, \alpha)$ using the technique described in section 2. We perform the inference for $b$ and $\alpha$ using the likelihood of eq. (2.5) with the Nested Sampling Monte Carlo library PYMULTINEST [60]. We assume flat uninformative priors for the two parameters, $b = [0, 5]$ and $\alpha = [-1, 1]$. We repeat the fit using several subsets of simulations $N_{\text{sims}} = 10000, 1000, 100, 50, 30$ to build the numerical covariance $C_n$. In figure 1, we show a comparison between the numerical and the model covariance matrices for three different subsets of simulations, $N_{\text{sims}} = 10000$, $N_{\text{sims}} = 1000$ and $N_{\text{sims}} = 100$. The errorbars are given by the standard deviation computed on a set of $n = N_{\text{sims,tot}}/N_{\text{sims}}$ independent subsets of simulations, where $N_{\text{sims,tot}} = 10000$ is the total number of available simulations. In this case, we fix the model covariance to be $C_\xi$, i.e. fixing $b$ to the value fit to the 2-point function measurements and $\alpha = 0$. As we decrease the number of simulations, the model deviates from the numerical covariance.

It is useful to show the posteriors resulting from the maximization of the likelihood, in the left panel of figure 2. As expected, as we use an increasing number of simulations, the contours shrink, but they are consistent with each other. Even with only 30 simulations we can broadly constrain the parameters well within the prior. Interestingly, the data prefer a non-zero value of the shot noise parameter, and a value of the bias that disagrees with the one we have fit to the two-point correlation function (dashed line). This effect is distinguishable even with a very small number of simulations and it becomes very significant at $N_{\text{sims}} = 10000$ with an over 30 sigma tension between the standard Poisson shot noise and our measurement of it. As mentioned, this is not necessarily reflecting a proper departure from the Poisson limit of the halo distribution shot-noise, but it can include additional systematics, e.g. geometry.
Figure 1. Comparison between the model $C_\xi$ (dashed lines) and numerical (solid lines) covariance, computed with different number $N_{\text{sims}}$ of simulations. Different colors represent different terms of the matrices: diagonal in blue, first off-diagonal in red, second off-diagonal in green and fourth off-diagonal in orange. In the bottom panels, percent residuals between numerical and model covariances. Errorbars are heavily correlated, which makes the scatter between neighboring points less than that implied by errorbars. Effects, and the incompleteness of the model.\textsuperscript{3} Another interesting comment is that, although the terms containing $\alpha$ are always subdominant, they contribute to the fitting process in a non-negligible way and with approximately the same weight of the main term. Thus, even if the difference between $C_\xi$ and $C_n$ is only about 5\%, it is not unreasonable for the fit to prefer a value of $\alpha$ with about a 20\% boost over the fiducial value ($\alpha \sim 0.2$ instead of $\alpha = 0$).

4.3.2 A $\chi^2$ test for the inverse of the covariance

In figure 3 we plot the histogram of $\chi^2$ values for the model covariances built from the subsets of simulations as described in section 3. As a reference, we also calculate the histogram for the numerical covariance matrix using $N_{\text{sims}} = 10000$ simulations and the model covariance where we fix the bias from the correlation function fit and $\alpha = 0$, $C_\xi$. All these candidate models for the covariance are compared to the theoretical $\chi^2$ distribution for this setup, which has mean $\mu = 11$ and variance $\sigma^2 = 22$. We calculate the mean and variance of each histogram; their errors are estimated by bootstrap, i.e. by resampling the data and evaluating the confidence region from the bootstrap distribution of the resampled statistics. From this test it is clear that our model covariance agrees reasonably well with the theoretical curve already with only $N_{\text{sims}} = 30$ simulations, even though with large error bars. On the other hand, the model covariance with fixed parameters is significantly away from the predicted distribution.

\textsuperscript{3} As a further check, we have fit $b$ and $\alpha$ directly on power spectrum measurements performed on the same set of simulations. The fitted values are also in disagreement with our best-fit values found using our model covariance. This confirms that the discrepancy is explained by the incompleteness of the model covariance, rather than a true deviation from Poisson shot-noise.
Figure 2. Left panel: contour plots at 68 and 95 per cent of confidence level for the covariance parameters, fitted over different numbers of simulations. The gray dashed lines represent the reference values, given by $\alpha = 0$ (Poissonian shot-noise) and bias fitted from the 2-point correlation function. Right panel: model covariances for the cases $N_{\text{sims}} = 10000$, 100, 30 and $\xi(r)$ fit (respectively solid dark, dashed light, dotted gray and black lines). In the bottom panel, percent residuals with respect to the full numerical matrix. Color code as in figure 1.

Figure 3. $\chi^2$ distribution corresponding to the fitted parameters for different numbers of simulations (colored histograms), compared to the predicted distribution (black). In the bottom panels, comparison with the matrix with parameters fitted from $\xi(r)$ and the numerical matrix. The errors on the mean and the variance are computed with bootstrap method.
4.3.3 Effects on cosmological parameters

The final test is to show that from the clustering analysis of the two-point correlation function we are able to correctly recover cosmological parameters from the simulations by using our model covariance. We consider a Gaussian likelihood for the observable, the two-point correlation function of halos. The theoretical model is the one given in eq. (4.2), where we vary the linear bias and two cosmological parameters, $\Omega_m$ and $\sigma_8$, which enter through the linear matter power spectrum. We choose the following flat priors for these parameters: $\{b, \Omega_m, \sigma_8\} = \{[0, 5], [0.20, 0.35], [0.7, 0.9]\}$. As for the covariance, we use our model covariance set to the best-fit values of figure 2. Note that we allow the bias of the covariance model to be different to the bias in the correlation function model. This is because we expect the former to also play a role in absorbing the missing terms in the covariance model, and thus lose its physical meaning of bias. For comparison, we also consider the case in which covariance and correlation function biases are the same, i.e. $C_\zeta$. All the cases are compared to the results with the full numerical covariance $C_n$. The results are shown in the left panel of figure 4. As expected, our model covariance with only $N_{\text{sims}} = 30$ simulations does not bias the contours of the parameters with respect to the numerical one built from $N_{\text{sims}} = 10000$ simulations. By eye, the fixed model covariance seems to be as good as the others, deviating by a small amount from the numerical one. To better evaluate the differences between the results, we estimate the accuracy of the parameter posteriors by computing the figure of merit [61] in the $\Omega_m$-$\sigma_8$ plane

$$\text{FoM}(\Omega_m, \sigma_8) = \frac{1}{\sqrt{\text{det}[\text{Cov}(\Omega_m, \sigma_8)]}},$$

(4.6)

where $C(\Omega_m, \sigma_8)$ is the parameter covariance computed from the sampled points. To take into account the statistical uncertainty of the likelihood maximization process, we compute the figure of merit as the average over $n = 10$ realizations, with errors given by the standard error $\varepsilon = \sigma/\sqrt{n}$. Clearly, the figure of merit here is to be intended as an overall estimate of the parameters errors and it is relevant only in its departure from the fiducial value. A good covariance matrix would give the same FoM as numerical fit on 10,000 realizations: a higher FoM indicates errors that are too small, while a low FoM indicates errors that are too large.

The result is shown in the right panel of figure 4. It confirms that our model covariance works well already with $N_{\text{sims}} = 30$, and provides a perfect match for $N_{\text{sims}} = 100$ simulations. We also find a curious effect: using too few simulations provides a covariance matrix that not only varies from one set of 30 simulations to another, but one that is also systematically biased towards high FoM (underestimation of the errorbars). With too few simulations, the numerical matrix turns out to be inaccurate, providing a biased fit of the model. As shown in left panel of figure 2, the covariance fitted from 30 simulations underestimates the true covariance; different results can be obtained from different subsets of simulations, as the data on which performing the fit can be biased in the opposite direction. The inaccuracy of such fit is confirmed by the FoM of figure 4.

Finally, we see that using a covariance where we fix the bias to the fitted value to the two-point correlation function measurements and $\alpha = 0$, $C_\zeta$, provides a $\sim 20\%$ underestimation of the error with respect to the numerical one.

5 Test 2: bispectrum

As a second test, we study the bispectrum, i.e. the three-point correlation function in Fourier space, on the same halo catalogs used for the first test. The motivation for choosing this
Figure 4. Left panel: contour plots at 68 and 95 per cent of confidence level for the cosmological parameters $\Omega_m$ and $\sigma_8$, computed with different covariance matrices: full numerical (gray), no-fit model (pink), $N_{\text{sims}} = 30$ fit (blue), $N_{\text{sims}} = 100$ fit (yellow), $N_{\text{sims}} = 10000$ fit (red). Right panel: figure of merit in the $\Omega_m - \sigma_8$ plane for different covariance cases: the numerical covariance matrix drawn from the full set of $N_{\text{sims}} = 10000$ simulation, our model covariance for varying $N_{\text{sims}}$ where we use the best-fit values for $b$ and $\alpha$ as found from the likelihood minimization, and the model covariance where we use $b$ fit from the two-point correlation function and $\alpha = 0$. The error bars are computed from an average of $n = 10$ realizations, as $\epsilon = \sigma/\sqrt{n}$.

observable as a test is twofold: first, it is an example where the covariance matrix can have very large dimensions. As compared to the power spectrum, where the covariance matrix is built from a data vector with typically $N \sim 20–50$ components, for the bispectrum the data vector may contain hundreds to thousands of triangles. Such a large data vector, and consequently covariance matrix, makes it very tricky to use the numerical covariance matrix as defined in eq. (2.2). This is because we usually need $N_{\text{sims}} \gg N$ to beat numerical noise, $N_{\text{sims}}$ being the number of simulations and $N$ the dimension of the data vector. Moreover, whenever the number of simulations is lower than the dimension of the data vector, the numerical covariance matrix cannot be inverted. Secondly, it is a way of testing a more realistic scenario, as the modeling of the bispectrum is significantly more involved than the one for a two-point function. In fact, as we will see, the model covariance we are going to use is knowingly incomplete, and our method will not be able to perform as well as a numerical covariance matrix drawn from $N_{\text{sims}} = 10000$ simulations. It will show, however, that a fit with parameters can help in building a reliable covariance matrix, even if the model is incomplete, as compared to considering a diagonal Gaussian covariance matrix.

5.1 The bispectrum and its covariance

Let us first set the notation on the bispectrum. Differently than the previous example, we work in Fourier space, and the bispectrum is defined as

$$\langle \delta(k_1)\delta(k_2)\delta(k_3) \rangle = \frac{\delta_K(k_1 + k_2 + k_3)}{k_3^3} B(k_1, k_2, k_3),$$

(5.1)
where $\delta(k)$ is the discrete Fourier transform of the density contrast $\delta(x)$, $\delta_k$ is the Kronecker symbol (equal to unity when the argument vanishes, zero otherwise) and $k_f = 2\pi/L$ is the fundamental frequency of a cubic box of volume $L^3$. We measure the bispectrum for all the 10000 Pinocchio halo catalogs. We use unbiased estimators for the measurement of the bispectrum following the definition of [62, 63]. We implement a fourth-order density interpolation and the interlacing scheme described in [64]. We divide $k$-modes into bins of width $\Delta k = k_f$ and up to a $k_{\text{max}} = 0.12 h$/Mpc, for a total of 29 $k$-bins and 2766 triangles.

We consider the approximation for the bispectrum covariance given by\footnote{The complete formula for the bispectrum covariance would read}

$$C_{ij}^{B} = C_{ij}^{B,(PPP)} + C_{ij}^{B,(BB)} + C_{ij}^{B,(PT)} + C_{ij}^{B,(Pn)},$$

\begin{equation}
(5.3)
\end{equation}

where $\alpha$ and $\beta$ are free parameters and $\delta_k^{K_{k_1k_2}}$ is a Kronecker symbol which is non-vanishing when the two triangles $i$ and $j$ have the smallest side in common, i.e. $k_1^i$ is equal to $k_3^j$. The first term, $C_{ij}^{B,(PPP)}$, is the Gaussian contribution, which we write down in the thin-shell approximation as

$$C_{ij}^{B,(PPP)} \approx \delta_k^{SB} \frac{k_3^{SB}}{N_{tr}^{i}} P(k_1^{i}) P(k_2^{j}) P(k_3^{j}),$$

\begin{equation}
(5.4)
\end{equation}

where $s_B = 6, 2, 1$ is the symmetry factor accounting for the shape of the triangles (equilateral, isosceles and scalene, respectively) and $N_{tr}^{i}$ is the number of fundamental triangles in the triangle bin $\{k_1^i, k_2^j, k_3^j\}$. The second term is defined as

$$C_{ij}^{B,(BB)} \approx B_i B_j \left( \Sigma_{ij}^{bb} + 8 \text{perm.} \right),$$

\begin{equation}
(5.5)
\end{equation}

where $i$ and $j$ indicate the triangle bins $\{k_1^i, k_2^i, k_3^i\}$ and $\{k_1^j, k_2^j, k_3^j\}$, respectively, $B_i$ is the bispectrum for the triangle bin $i$ and $\Sigma_{ij}^{bb}$ is a mode-counting factor that depends again on the shape of the triangle. Both terms are computed using measurements of the power spectrum and bispectrum directly, without using any perturbative calculation. The resulting covariance has a block-diagonal structure, which we show in the left panel of figure 5 by plotting the cross-correlation matrix

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}},$$

\begin{equation}
(5.6)
\end{equation}

which helps in visualizing the importance of off-diagonal elements with respect to diagonal ones.

For comparison, we also plot the numerical covariance for $N_{\text{sims}} = 10000$, in the right panel of the same figure. The numerical covariance shows a similar block diagonal structure as the modeled one. Indeed, these particular non-diagonal entries of the bispectrum covariance are the largest terms in the non-Gaussian covariance [42]. Based on theoretical considerations [42, 67], we expect that $\alpha \simeq 1$ and $\beta \simeq 2$. In particular, $\beta = 2$ assumes that the contribution due the product of the power spectrum and trispectrum of the field can be proportional to the BB term and neglect $P_n$. See [65] for more details.
approximated by the term in eq. (5.5), a good approximation for squeezed triangles, but not for generic shapes. In the numerical covariance, we can clearly see more structure outside the blocks, which eq. (5.3) does not model.

5.2 Results

Having defined the model covariance for the bispectrum test, we can proceed with the same steps as done for the two-point correlation function in the section 4.

5.2.1 Maximizing the likelihood

We maximize the likelihood of eq. (2.5) varying $\alpha$ and $\beta$ using sets of $C_n$ for $N_{\text{sims}} = 30$, 50, 100, 1000 and 10000 as in the previous test. We plot the contour plots for $\alpha$ and $\beta$ in figure 6. Interestingly, the theoretical values do not fall within the contours for $N_{\text{sims}} = 1000$ and $N_{\text{sims}} = 10000$. In fact, this is not surprising, since the model covariance is an incomplete model of the full covariance, which has non-zero elements also outside the block-diagonal structure modeled by eq. (5.3). These non-zero elements might be sourced by a connected 6-point function, or by correlated noise. Even though these terms are small compared to the ones we model, when considering a large number of simulations they become significant with respect to sample noise and the fit tries to adapt to them shifting the central values of $\alpha$ and $\beta$.

5.2.2 A $\chi^2$ test for the inverse of the covariance

Using the best-fit values for the $\alpha$ and $\beta$ parameters obtained from maximizing the likelihood, we perform the $\chi^2$ test introduced in the previous section in order to verify the performance of our model covariance. We show results in figure 7. Differently than for the two-point correlation function, our model for the bispectrum covariance does not perform as well as the numerical covariance obtained from $N_{\text{sims}} = 10000$ simulations. Indeed, even though the mean $\mu$ of the $\chi^2$ distribution of our model covariance is in good agreement with the theoretical value of $\mu = 2766$ already at $N_{\text{sims}} = 100$ simulations, the variance $\sigma^2$ is significantly

Figure 5. Cross-correlation matrix $r_{ij}$ of the halo bispectrum covariance for the model covariance (left) and the numerical covariance computed on $N_{\text{sims}} = 10000$ simulations (right).
Figure 6. Same of left panel of figure 2 for the bispectrum covariance. The parameters $\alpha$ and $\beta$ represent the amplitude of the Gaussian and (part of the) non-Gaussian covariance, respectively, as defined in eq. (5.3).

Figure 7. Same of figure 3, for the bispectrum covariance.

off.\footnote{Again, we can expect this given that our model covariance is incomplete. Nevertheless, it is useful to point out that, in the absence of a theoretical prior on the free parameters, and a numerical covariance, this method would prove useful in building a good approximation of the true covariance. Indeed, given that the mean of the $\chi^2$ is within the theoretical

\footnotetext{It is interesting to point out that even the numerical covariance does not fit perfectly well the reference $\chi^2$. We attribute this fact to the actual likelihood of the bispectrum not being Gaussian, at this volume.}
value, we expect that our model covariance does not bias strongly parameter estimation. For comparison, we compute the histogram also for the pure Gaussian covariance, i.e. a model covariance built fixing $\alpha = 1$ and $\beta = 0$, see eq. (5.4). In this case, both the mean and the variance of the histogram are significantly off from the theoretical curve.

In order to further confirm that the failure of the $\chi^2$ test is linked to the incompleteness of the model, we perform the following test: we calculate the numerical covariance matrix where we only keep the block diagonals that are modeled by our model covariance, putting all other off-diagonal terms to zero. We then redo the $\chi^2$ test using this numerical covariance and compare to the result that we got using the model covariance with $N_{\text{sims}} = 10000$ simulations. We show the result of this test in figure 8. Indeed, if we only consider the blocks that we model in the numerical covariance, the $\chi^2$ test gives a very similar result as for the model covariance.

6 Conclusions

In this paper, we have explored a two-steps method to build a reliable and cheap covariance matrix. It is based on using two basic ingredients: an even imperfect model covariance with free parameters, and a $\chi^2$-test. Upon successfully finding the best-fit values for the free-parameters for which the $\chi^2$-test is passed, the methods provides a reliable covariance matrix having to run a small number of simulations, typically smaller than the data-vector itself. We have applied the method to two contexts, using the two-point correlation function and the bispectrum of mock halo catalogs as observables. We employed knowingly incomplete models of the covariance with two free parameters in both cases. We have shown how to verify their reliability without relying on knowing the true covariance via a $\chi^2$ test. In the case of the two-point correlation function, using only 100 simulations, we are able to recover unbiased estimates of the cosmological parameters of the simulation. We found that our covariance matrix fit prefers a non-Poisson shot noise in our simple model. Since the effective shot noise
is different when derived from covariance matrix compared to direct power spectrum fit, the most likely explanation is that we are detecting higher-order corrections. In the second example, using a non-Gaussian model for the bispectrum covariance we improve significantly with respect to a Gaussian covariance, but we are not able to match the numerical covariance drawn from $N_{\text{sims}} = 10000$.

Consistent with our analytic results in section 3 we find in both cases that the first moment of $\chi^2$ is consistent with theoretical expectation. The second moment is correct for the case of correlation function, but too high in the case of bispectrum. There the wrong covariance matrix causes an extra excess on the $\chi^2$ values with rms of $\sim 60$, similar to intrinsic scatter from $\chi^2$ distribution.

The main strength of the method is that it can be applied to test a proposed covariance matrix even in the case where the number of simulated universes is not sufficient to generate even a positive-definite matrix. It relies on examining the consistency of $\chi^2$ values derived from a model covariance matrix with theoretical expectation. A simple way of doing this is to measure mean and variance of realization $\chi^2$ values, comparing these with expectation values for a $\chi^2$ distribution. Even when the test is performed on the same simulated realization used for fitting the covariance matrix, the result might be inconsistent if the covariance model is insufficiently flexible to describe the actual covariance. The results also quantify the badness of the covariance matrix by giving typical expected biases in $\chi^2$ values.

We also found that the adopted model can have intrinsic biases when fitting an insufficient number of simulations. For example, when fitting with only 30 simulations, the derived best-fit covariance matrix not only varies with sample variance but is systematically underestimated (see figure 2) for this subset of simulations. Investigating this effect further and developing covariance models that are unbiased (in the sense that they in average produce an unbiased covariance matrix) is left to future work.

In practice we have found that even covariance matrices that have a demonstrably biased $\chi^2$ distributions often perform well enough in practical situations, giving cosmological parameters constraints that are acceptably biased with respect to the ideal case (see figure 4). This indicates that while our method allows for a proper propagation of covariance matrix uncertainties coming from sample variance on the number of realizations used into cosmological inferences, this might very rarely be used in practice. At the same time, the distribution of simulated $\chi^2$ values might allow us to “recalibrate” the goodness of fit measures. In our bispectrum example, the fitted covariance matrix produces a distribution of $\chi^2$ values which is too broad. This would allow us to better quantify the goodness of fit on the real data, which might be formally bad, but consistent with the distribution obtained with simulation with an imperfect covariance matrix.

While our test shows in principle the potential of our method, there are several more realistic setups where this method could prove to be crucial. For instance, it will be important to further test to what degree the fitting procedure can compensate for an incomplete modeling of the covariance and if approximate phenomenological terms can be added into covariance matrix model that absorb terms missing from the theory. A good testing ground for these tests is the galaxy/halo bispectrum framework we introduced in section 5, since the covariance in this case has large off-diagonal terms for which we have only an incomplete model. We leave these tests to future work.
Acknowledgments

M.B. acknowledges support from the Netherlands Organization for Scientific Research (NWO), which is funded by the Dutch Ministry of Education, Culture and Science (OCW) under VENI grant 016.Veni.192.210. E.S. and P.M. are partially supported by the INFN INDARK PD51 grant. A.S. acknowledges hospitality of Institute for Fundamental Physics of the Universe in Trieste where this work was initiated. A.F. and A.S. are supported by the ERC-StG ‘ClustersXCosmo’ grant agreement 716762, by the FARE-MIUR grant ‘ClustersXEuclid’ R165SBKTMA, and by INFN InDark Grant.

References

[1] C. Blake et al., The WiggleZ Dark Energy Survey: the growth rate of cosmic structure since redshift z=0.9, Mon. Not. Roy. Astron. Soc. 415 (2011) 2876 [arXiv:1104.2948] [ INSPIRE].

[2] S. de la Torre et al., The VIMOS Public Extragalactic Redshift Survey (VIPERS). Galaxy clustering and redshift-space distortions at z=0.8 in the first data release, Astron. Astrophys. 557 (2013) A54 [arXiv:1303.2622] [ INSPIRE].

[3] BOSS collaboration, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample, Mon. Not. Roy. Astron. Soc. 470 (2017) 2617 [arXiv:1607.03155] [ INSPIRE].

[4] H. Gil-Marín et al., The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies, Mon. Not. Roy. Astron. Soc. 465 (2017) 1757 [arXiv:1606.00439] [ INSPIRE].

[5] KiDS collaboration, KiDS-1000 Cosmology: cosmic shear constraints and comparison between two point statistics, Astron. Astrophys. 645 (2021) A104 [arXiv:2007.15633] [ INSPIRE].

[6] DES collaboration, Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing, Phys. Rev. D 105 (2022) 023520 [arXiv:2105.13549] [ INSPIRE].

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

[8] G. D’Amico et al., The Cosmological Analysis of the SDSS/BOSS data from the effective field theory of large-scale structure, JCAP 05 (2020) 005 [arXiv:1909.05271] [ INSPIRE].

[9] M. Biagetti, A. Cole and G. Shiu, The persistence of large scale structures I: primordial non-gaussianity, JCAP 04 (2021) 061 [arXiv:2009.04819] [ INSPIRE].

[10] S. Heydenreich, B. Brück and J. Harnois-Déraps, Persistent homology in cosmic shear: constraining parameters with topological data analysis, Astron. Astrophys. 648 (2021) A74 [arXiv:2007.13724] [ INSPIRE].

[11] M. Biagetti, J. Calles, L. Castiblanco, A. Cole and J. Noreña, Fisher forecasts for primordial non-Gaussianity from persistent homology, JCAP 10 (2022) 002 [arXiv:2203.08262] [ INSPIRE].

[12] S. Heydenreich et al., Persistent homology in cosmic shear II: A tomographic analysis of DES-Y1, Astron. Astrophys. 667 (2022) A125 [arXiv:2204.11831] [ INSPIRE].

[13] G. Lavaux and B.D. Wandelt, Precision cosmology with voids: definition, methods, dynamics, Mon. Not. Roy. Astron. Soc. 403 (2010) 1392 [arXiv:0906.4101] [ INSPIRE].

[14] R. Skibba, R.K. Sheth, A.J. Connolly and R. Scranton, The luminosity-weighted or ‘marked’ correlation function, Mon. Not. Roy. Astron. Soc. 369 (2006) 68 [astro-ph/0512463] [ INSPIRE].
[15] M. White, *A marked correlation function for constraining modified gravity models*, *JCAP* **11** (2016) 057 [arXiv:1609.08632] [inSPIRE].

[16] J. Armijo, Y.-C. Cai, N. Padilla, B. Li and J.A. Peacock, *Testing modified gravity using a marked correlation function*, *Mon. Not. Roy. Astron. Soc.* **478** (2018) 3627 [arXiv:1801.08632] [inSPIRE].

[17] M. Manera et al., *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: a large sample of mock galaxy catalogues*, *Mon. Not. Roy. Astron. Soc.* **428** (2012) 1036 [arXiv:1203.6609] [inSPIRE].

[18] J. Armijo, Y.-C. Cai, N. Padilla, B. Li and J.A. Peacock, *Testing modified gravity using a marked correlation function*, *Mon. Not. Roy. Astron. Soc.* **478** (2018) 3627 [arXiv:1801.08975] [inSPIRE].

[19] DES collaboration, *Dark Energy Survey Year 1 Results: galaxy mock catalogues for BAO*, *Mon. Not. Roy. Astron. Soc.* **479** (2018) 94 [arXiv:1712.06232] [inSPIRE].

[20] P. Monaco, *Approximate methods for the generation of dark matter halo catalogs in the age of precision cosmology*, *Galaxies* **4** (2016) 53 [arXiv:1605.07752] [inSPIRE].

[21] M. Colavincenzo et al., *Comparing approximate methods for mock catalogues and covariance matrices — I. Correlation function*, *Mon. Not. Roy. Astron. Soc.* **482** (2019) 1786 [arXiv:1806.09477] [inSPIRE].

[22] M. Lippich et al., *Comparing approximate methods for mock catalogues and covariance matrices — I. Correlation function*, *Mon. Not. Roy. Astron. Soc.* **482** (2019) 1786 [arXiv:1806.09477] [inSPIRE].

[23] L. Blot et al., *Comparing approximate methods for mock catalogues and covariance matrices II: power spectrum multipoles*, *Mon. Not. Roy. Astron. Soc.* **485** (2019) 2806 [arXiv:1806.09497] [inSPIRE].

[24] M. Colavincenzo et al., *Comparing approximate methods for mock catalogues and covariance matrices — III: bispectrum*, *Mon. Not. Roy. Astron. Soc.* **482** (2019) 4883 [arXiv:1806.09499] [inSPIRE].

[25] Y. Feng, M.-Y. Chu, U. Seljak and P. McDonald, *FastPM: a new scheme for fast simulations of dark matter and haloes*, *Mon. Not. Roy. Astron. Soc.* **463** (2016) 2273 [arXiv:1603.00476] [inSPIRE].

[26] M. Hartlap, T. Schrabback, P. Simon and P. Schneider, *The non-Gaussianity of the cosmic shear likelihood or how odd is the Chandra Deep Field South?, Astron. Astrophys.* **504** (2009) 689 [arXiv:0901.3269] [inSPIRE].

[27] A. Taylor, B. Joachimi and T. Kitching, *Putting the precision in precision cosmology: how accurate should your data covariance matrix be?, Mon. Not. Roy. Astron. Soc.* **432** (2013) 1928 [arXiv:1212.4359] [inSPIRE].

[28] W.J. Percival et al., *The Clustering of Galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: including covariance matrix errors*, *Mon. Not. Roy. Astron. Soc.* **439** (2014) 2531 [arXiv:1312.4841] [inSPIRE].

[29] E. Sellentin and A.F. Heavens, *Parameter inference with estimated covariance matrices*, *Mon. Not. Roy. Astron. Soc.* **456** (2016) L132 [arXiv:1511.05969] [inSPIRE].

[30] R. Scoccimarro, *The bispectrum: from theory to observations*, *Astrophys. J.* **544** (2000) 597 [astro-ph/0004086] [inSPIRE].

[31] A.J.S. Hamilton, C.D. Rimes and R. Scoccimarro, *On measuring the covariance matrix of the nonlinear power spectrum from simulations*, *Mon. Not. Roy. Astron. Soc.* **371** (2006) 1188 [astro-ph/0511416] [inSPIRE].
[32] A.C. Pope and I. Szapudi, *Shrinkage estimation of the power spectrum covariance matrix*, *Mon. Not. Roy. Astron. Soc.* **389** (2008) 766 [arXiv:0711.2509] [SPIRE].

[33] B. Joachimi, *Non-linear shrinkage estimation of large-scale structure covariance*, *Mon. Not. Roy. Astron. Soc.* **466** (2017) L83 [arXiv:1612.00752] [SPIRE].

[34] D.J. Paz and A.G. Sanchez, *Improving the precision matrix for precision cosmology*, *Mon. Not. Roy. Astron. Soc.* **454** (2015) 4326 [arXiv:1508.03162] [SPIRE].

[35] N. Chartier and B.D. Wandelt, *CARPool covariance: fast, unbiased covariance estimation for large-scale structure observables*, *Mon. Not. Roy. Astron. Soc.* **509** (2021) 2220 [arXiv:2106.11718] [SPIRE].

[36] N.S.M. de Santi and L.R. Abramo, *Improving cosmological covariance matrices with machine learning*, *JCAP* **09** (2022) 013 [arXiv:2205.10881] [SPIRE].

[37] F. Lacasa, *The impact of braiding covariance and in-survey covariance on next-generation galaxy surveys*, *Astron. Astrophys.* **634** (2020) A74 [arXiv:1909.00791] [SPIRE].

[38] X. Fang, T. Eifler and E. Krause, *2D-FFTLog: efficient computation of real space covariance matrices for galaxy clustering and weak lensing*, *Mon. Not. Roy. Astron. Soc.* **497** (2020) 2699 [arXiv:2004.04833] [SPIRE].

[39] N.S. Sugiyama, S. Saito, F. Beutler and H.-J. Seo, *Perturbation theory approach to predict the covariance matrices of the galaxy power spectrum and bispectrum in redshift space*, *Mon. Not. Roy. Astron. Soc.* **497** (2020) 1684 [arXiv:1908.06234] [SPIRE].

[40] D. Wadekar and R. Scoccimarro, *Galaxy power spectrum multipoles covariance in perturbation theory*, *Phys. Rev. D* **102** (2020) 123517 [arXiv:1910.02914] [SPIRE].

[41] D. Wadekar, M.M. Ivanov and R. Scoccimarro, *Cosmological constraints from BOSS with analytic covariance matrices*, *Phys. Rev. D* **102** (2020) 123521 [arXiv:2009.00622] [SPIRE].

[42] M. Biagetti, L. Castiblanco, J. Noreña and E. Sefusatti, *The covariance of squeezed bispectrum configurations*, *JCAP* **09** (2022) 009 [arXiv:2111.05887] [SPIRE].

[43] X. Xu, N. Padmanabhan, D.J. Eisenstein, K.T. Mehta and A.J. Cuesta, *A 2% distance to z=0.35 by reconstructing baryon acoustic oscillations - II: fitting techniques*, *Mon. Not. Roy. Astron. Soc.* **427** (2012) 2146 [arXiv:1202.0091] [SPIRE].

[44] R. O’Connell, D. Eisenstein, M. Vargas, S. Ho and N. Padmanabhan, *Large covariance matrices: smooth models from the two-point correlation function*, *Mon. Not. Roy. Astron. Soc.* **462** (2016) 2681 [arXiv:1510.01740] [SPIRE].

[45] Z. Slepian et al., *The large-scale three-point correlation function of the SDSS BOSS DR12 CMASS galaxies*, *Mon. Not. Roy. Astron. Soc.* **468** (2017) 1070 [arXiv:1512.02231] [SPIRE].

[46] D.W. Pearson and L. Samushia, *Estimating the power spectrum covariance matrix with fewer mock samples*, *Mon. Not. Roy. Astron. Soc.* **457** (2016) 993 [arXiv:1509.00064] [SPIRE].

[47] A. Hall and A. Taylor, *A Bayesian method for combining theoretical and simulated covariance matrices for large-scale structure surveys*, *Mon. Not. Roy. Astron. Soc.* **483** (2019) 189 [arXiv:1807.06875] [SPIRE].

[48] DES collaboration, *Dark Energy Survey year 3 results: covariance modelling and its impact on parameter estimation and quality of fit*, *Mon. Not. Roy. Astron. Soc.* **508** (2021) 3125 [arXiv:2012.08568] [SPIRE].

[49] R. Scoccimarro, M. Zaldarriaga and L. Hui, *Power spectrum correlations induced by nonlinear clustering*, *Astrophys. J.* **527** (1999) 1 [astro-ph/9901099] [SPIRE].

[50] J.N. Grieb, A.G. Sánchez, S. Salazar-Albornoz and C. Dalla Vecchia, *Gaussian covariance matrices for anisotropic galaxy clustering measurements*, *Mon. Not. Roy. Astron. Soc.* **457** (2016) 1577 [arXiv:1509.04293] [SPIRE].
[51] P. Monaco, T. Theuns and G. Taffoni, *Pinocchio: pinpointing orbit-crossing collapsed hierarchical objects in a linear density field*, Mon. Not. Roy. Astron. Soc. **331** (2002) 587 [astro-ph/0109323] [inSPIRE].

[52] P. Monaco et al., *An accurate tool for the fast generation of dark matter halo catalogs*, Mon. Not. Roy. Astron. Soc. **433** (2013) 2389 [arXiv:1305.1505] [inSPIRE].

[53] E. Munari et al., *Improving fast generation of halo catalogues with higher order Lagrangian perturbation theory*, Mon. Not. Roy. Astron. Soc. **465** (2017) 4658 [arXiv:1605.04788] [inSPIRE].

[54] A. Oddo, E. Sefusatti, C. Porciani, P. Monaco and A.G. Sánchez, *Toward a robust inference method for the galaxy bispectrum: likelihood function and model selection*, JCAP **03** (2020) 056 [arXiv:1908.01774] [inSPIRE].

[55] A. Oddo, F. Rizzo, E. Sefusatti, C. Porciani and P. Monaco, *Cosmological parameters from the likelihood analysis of the galaxy power spectrum and bispectrum in real space*, JCAP **11** (2021) 038 [arXiv:2108.03204] [inSPIRE].

[56] S.D. Landy and A.S. Szalay, *Bias and variance of angular correlation functions*, Astrophys. J. **412** (1993) 64 [inSPIRE].

[57] F. Marulli, A. Veropalumbo and M. Moresco, *CosmoBolognaLib: C++ libraries for cosmological calculations*, Astron. Comput. **14** (2016) 35 [arXiv:1511.00012] [inSPIRE].

[58] O.H.E. Philcox and D.J. Eisenstein, *Estimating covariance matrices for two- and three-point correlation function moments in arbitrary survey geometries*, Mon. Not. Roy. Astron. Soc. **490** (2019) 5931 [arXiv:1910.04764] [inSPIRE].

[59] Y. Li, S. Singh, B. Yu, Y. Feng and U. Seljak, *Disconnected covariance of 2-point functions in large-scale structure*, JCAP **01** (2019) 016 [arXiv:1811.05714] [inSPIRE].

[60] J. Buchner et al., *X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue*, Astron. Astrophys. **564** (2014) A125 [arXiv:1402.0004] [inSPIRE].

[61] A. Albrecht et al., *Report of the dark energy task force*, astro-ph/0609591 [inSPIRE].

[62] R. Scoccimarro, S. Colombi, J.N. Fry, J.A. Frieman, E. Hivon and A. Melott, *Nonlinear evolution of the bispectrum of cosmological perturbations*, Astrophys. J. **496** (1998) 586 [astro-ph/9704075] [inSPIRE].

[63] R. Scoccimarro, *Fast estimators for redshift-space clustering*, Phys. Rev. D **92** (2015) 083532 [arXiv:1506.02729] [inSPIRE].

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

[65] M. Biagetti, L. Castiblanco, J. Noreña and E. Sefusatti, *The covariance of squeezed bispectrum configurations*, JCAP **09** (2022) 009 [arXiv:2111.05887] [inSPIRE].

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

[67] A. Barreira, *The squeezed matter bispectrum covariance with responses*, JCAP **03** (2019) 008 [arXiv:1901.01243] [inSPIRE].