CARPool covariance: fast, unbiased covariance estimation for large-scale structure observables

Nicolas Chartier, Benjamin D. Wandelt

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **509,** 2220–2233 (2022) https://doi.org/10.1093/mnras/stab3097
Advance Access publication 2021 October 25

CARPool covariance: fast, unbiased covariance estimation for large-scale structure observables

Nicolas Chartier[1,2]★ and Benjamin D. Wandelt[2,3]

[1]*Laboratoire de Physique de l'École Normale Supérieure, ENS, Universite PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France*
[2]*Institut d'Astrophysique de Paris, Sorbonne Université, CNRS, UMR 7095, 98 bis bd Arago, F-75014 Paris, France*
[3]*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*

Accepted 2021 October 18. Received 2021 October 18; in original form 2021 July 27

ABSTRACT

The covariance matrix $\mathbf{\Sigma}$ of non-linear clustering statistics that are measured in current and upcoming surveys is of fundamental interest for comparing cosmological theory and data and a crucial ingredient for the likelihood approximations underlying widely used parameter inference and forecasting methods. The extreme number of simulations needed to estimate $\mathbf{\Sigma}$ to sufficient accuracy poses a severe challenge. Approximating $\mathbf{\Sigma}$ using inexpensive but biased *surrogates* introduces model error with respect to full simulations, especially in the non-linear regime of structure growth. To address this problem, we develop a matrix generalization of Convergence Acceleration by Regression and Pooling (CARPool) to combine a small number of simulations with fast surrogates and obtain low-noise estimates of $\mathbf{\Sigma}$ that are unbiased by construction. Our numerical examples use CARPool to combine GADGET-III *N*-body simulations with fast surrogates computed using COmoving Lagrangian Acceleration (COLA). Even at the challenging redshift $z = 0.5$, we find variance reductions of at least $\mathcal{O}(10^1)$ and up to $\mathcal{O}(10^4)$ for the elements of the matter power spectrum covariance matrix on scales $8.9 \times 10^{-3} < k_{\rm max} < 1.0\,{\rm h\,Mpc^{-1}}$. We demonstrate comparable performance for the covariance of the matter bispectrum, the matter correlation function, and probability density function of the matter density field. We compare eigenvalues, likelihoods, and Fisher matrices computed using the CARPool covariance estimate with the standard sample covariance and generally find considerable improvement except in cases where $\mathbf{\Sigma}$ is severely ill-conditioned.

Key words: methods: statistical – large-scale structure of Universe.

1 INTRODUCTION

In the era of precision cosmology, modelling the statistical properties of observables is crucial to derive cosmological parameters constraints from large-scale structure surveys. Particularly, the covariance matrix $\mathbf{\Sigma}$ of clustering statistics, such as the matter power spectrum and the matter bispectrum, as well as its inverse – the precision matrix – are key elements when building likelihood approximations, efficient estimators, or developing optimal summaries of observations for cosmological inference (Heavens, Jimenez & Lahav 2000; Eifler, Schneider & Hartlap 2009; Takahashi et al. 2009; Harnois-Déraps, Vafaei & Van Waerbeke 2012; Dodelson & Schneider 2013; Harnois-Déraps & Pen 2013; Blot et al. 2014; Percival et al. 2014; Taylor & Joachimi 2014; Alsing & Wandelt 2018; Harnois-Déraps, Giblin & Joachimi 2019; Hikage, Takahashi & Koyama 2020; Wadekar, Ivanov & Scoccimarro 2020).

Unfortunately, estimating the covariance matrix of large-scale structure observables is extremely challenging owing to both the large number of samples required and the computational cost per sample. A brute force solution to estimate covariance matrices would be to generate mock samples of survey statistics with computationally intensive *N*-body simulations reproducing the conditions of observation (volume, redshifts, sky area...), and then to compute the sample covariance matrix, which is an unbiased and positive (semi-) definite estimator of the true covariance. But high-quality estimates of the covariance are necessary because we actually need its inverse, the precision matrix. This will be dominated by the smallest eigenvalues of the covariance. It is just these small eigenvalues that need the largest number of samples to converge. For example, Blot et al. (2016) found in numerical experiments for a Euclid-like survey that at least 5000 independent *N*-body simulations are needed to estimate the power spectrum covariance in order to obtain cosmological parameter forecasts at an adequate level of accuracy given the precision of upcoming surveys. In spite of recent progress in optimization of various *N*-body codes, with GPU-acceleration and distributed-memory solutions (Springel 2005; Ishiyama, Fukushige & Makino 2009; Warren 2013; Habib et al. 2016; Potter, Stadel & Teyssier 2017; Garrison 2019), limited CPU time and memory resources and the large number of samples required mean that it will remain impractical to rely solely on full *N*-body simulations for covariance matrix estimation. For this reason, cosmologists have been investigating less costly alternatives to tackle next-generation data sets.

For certain clustering statistics that are amenable to perturbative treatment, analytical predictions allow computing covariance estimates rapidly, at the cost of making assumptions on the survey data. Such predictions typically use the Gaussian limit for the covariance. For example, Philcox et al. (2020) developed the RASCALC code that estimates the covariance of the two-point galaxy correlation function

★ E-mail: nicolas.chartier412@gmail.com

(2PCF) and only needs one data set as input: a shot noise rescaling, constrained by jackknife covariance matrices, describes deviations from Gaussianity, and as a result, the large-scale model covariance matrix is fully consistent with mocks. Philcox & Eisenstein (2019) had applied a similar approach to the auto- and cross-covariances of 2PCFs and three-point correlation functions (3PCFs) for general real-space survey statistics. Li et al. (2019) proposed an analytical computation of the 'disconnected' (Gaussian) part of the covariance, that is dominant on large scales, for both correlation functions and their Fourier space counterpart (power spectra): they found valuable accordance with mock estimates. Mohammed, Seljak & Vlah (2017) also experimented with an analytical decomposition of the covariance matrix motivated by perturbation theory and managed to get a 10 per cent agreement with simulations up to $k \approx 1\,\mathrm{h\,Mpc^{-1}}$. Mohammed & Seljak (2014) tested a simple model for the matter power spectrum motivated by the Zel'dovich approximation and stressed the influence of the simulation box volume on the convergence of the covariance matrix. Useful reviews of methods using theoretical predictions include Bernardeau et al. (2002) and Desjacques, Jeong & Schmidt (2018).

Alternatively, computational cosmologists have proposed various approximate solvers designed to be much faster than *N*-body simulations. One class of such methods exploits the availability of low order Lagrangian Perturbation Theory (LPT): Scoccimarro & Sheth (2002; PTHalos), Tassev & Zaldarriaga (2012), and Monaco et al. (2013) inspired by Taffoni, Monaco & Theuns (2002; PINOCCHIO) or Chuang et al. (2015; EZmocks). Futhermore, a significant number of PM (PARTICLE-MESH) codes, which treat the force as a field on a mesh, use the large-scale approximation provided by LPT: Tassev, Zaldarriaga & Eisenstein (2013; COLA) and Tassev et al. (2015; sCOLA) implemented by Leclercq et al. (2020) and Feng et al. (2016; FastPM) available in a distributed version by Modi, Lanusse & Seljak (2020), White, Tinker & McBride (2014; QPM), and Kitaura, Yepes & Prada (2014; PATCHY), to name a few.

Another family of approaches comprises mathematical models with free-parameters – *emulators* – that are trained on simulation suites covering a given range of cosmologies to then directly predict clustering parameters from upcoming data. Recent studies include DeRose et al. (2019), McClintock et al. (2019a,b), Zhai et al. (2019), Kasim et al. (2020), or Angulo et al. (2020). Although they can provide lightning-fast estimates, emulators are limited by the parameter range of the training set and do not guarantee unbiased results with respect to full solvers, of which they still need a large number of realizations to train. Some emulators based on deep learning architectures have been shown to reproduce particle positions or matter density fields from input initial conditions. In other words, they can produce snapshots of a low-resolution cosmological *N*-body code from which any clustering statistics can be extracted: He et al. (2019), Dai & Seljak (2020), and Kodi Ramanah et al. (2020). Fast approximate solvers – which we will refer to as *surrogates* and which comprise all the previous families of methods we mentioned – unfortunately do not match the accuracy of full *N*-body mocks, especially in the deeply non-linear regime. Blot et al. (2019), Colavincenzo et al. (2019), and Lippich et al. (2019), find statistical biases in parameters estimation with covariance matrices from surrogates up to 20 per cent higher than with covariances from full *N*-body codes.

Some works in cosmology are specifically dedicated to improving the estimation of covariance matrices. In the particular case of Gaussian-distributed weak lensing power spectra, Taylor, Joachimi & Kitching (2013) assessed the limits on parameter estimation imposed by the accuracy of the precision matrix and discussed solutions to relax them. Paz & Sánchez (2015) implemented a technique called *tapering* to estimate covariance and precision matrices and proved to be successful in reducing the confidence intervals of parameters without introducing bias. Pearson & Samushia (2016) fitted a theoretically motivated model with a mock catalogue in order to estimate the covariance matrix with fewer samples. Favole et al. (2020) provided more insight on the impact of jackknife resampling on covariance matrix estimates, which had also been experimented with by Escoffier et al. (2016) for the two-point galaxy clustering correlation function. In Hall & Taylor (2019), using a likelihood conditioned on both theoretical and simulated covariance matrices of summary statistics reduced the required number of simulations for covariance estimation. Pope & Szapudi (2008) applied the concept of linear shrinkage to the matter power spectrum covariance matrix: by optimally combining an empirical estimate with a specified simple target (for instance a diagonal covariance), they significantly improved the estimated covariance when few simulations are available. Regarding non-linear shrinkage, see for instance Joachimi (2017). In addition, the precision matrix being essential to derive parameters confidence bounds, the fact that the inverse of the unbiased covariance estimator is not an unbiased estimator of the precision is now well-known notably thanks to Hartlap, Simon & Schneider (2007). Numerous cosmology studies focus on precision matrix estimation and on the effects – parameters shifts for instance – of precision matrix biases (Friedrich & Eifler 2018; Sellentin & Heavens 2018; Friedrich et al. 2021; Percival et al. 2021; Philcox et al. 2021).

Chartier et al. (2020; CWAV20 from now on) developed the Convergence Acceleration by Regression and Pooling (CARPool) method, a general approach to reducing the number of simulations needed for low variance and explicitly unbiased estimates of clustering statistics. Equivalently, CARPool can be viewed as a way to obtain unbiased results from fast surrogates by running a small number of simulations. CWAV20 demonstrated a dramatic reduction of the number of simulations required to estimate the mean of a given statistic by exploiting the variance reduction principle known as *control variates* and combining a smaller number of costly simulations with a larger number of *surrogates*. CARPool exploits the correlation between full *N*-body runs and fast surrogates run on the same initial conditions, and proved to be very efficient for the estimation of the mean of the matter power spectrum, the bispectrum or the one-point probability density function (PDF).

It is therefore natural to study whether CARPool can improve the estimation of covariance matrices while reducing the number of simulations.[1] Showing this to be the case is the main contribution of this paper. We will first recall some theoretical results and generalize the CARPool approach to covariance matrices in Section 2. Then, we will show in Section 3, using likelihoods, eigenvalues, and Fisher information matrices computed from the estimated covariance matrices, that CARPool can reduce the number of mocks needed to estimate covariance matrices of clustering statistics by at least one order of magnitude and, depending on scale and observable, in some cases by several orders of magnitude. We will discuss our results and conclude in Section 4.

[1] Pontzen et al. (2016), Angulo & Pontzen (2016), and Villaescusa-Navarro et al. (2018) discuss variance reduction by designing special initial conditions that explicitly bias certain higher order n-point functions low. These would therefore not seem promising for improving estimates of the covariance matrix.

2 STATISTICAL METHODS

We adopt the same notation system as in CWAV20 up to a few necessary adaptations. Namely, let $\boldsymbol{y}$ be a costly simulation observable, constituted by scalar measurements y_i, $1 \leq i \leq p$, such that $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\mu} \in \mathbb{R}^p$, and $\boldsymbol{c}$ an approximate random observable with $\mathbb{E}[\boldsymbol{c}] = \boldsymbol{\mu_c} \in \mathbb{R}^q$. In computational cosmology, $\boldsymbol{y}$ and $\boldsymbol{c}$ can be, for instance, cold dark matter (CDM) power spectra with p and q power band bins, respectively.

2.1 Variance reduction with control variates

In this section, we recall some results dealing with the variance reduction technique known as *control variates*. For more details, see CWAV20.

2.1.1 General multivariate case

In order to compute an unbiased estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$, a straightforward solution is to use the sample mean from a set of *independent and identically distributed* realizations $\boldsymbol{y}_n$, $n = 1, \dots N$,

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}} \equiv \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{y}_n \,. \tag{1}$$

The standard deviation σ_i of each $\bar{y}_i$, $1 \leq i \leq p$, decreases slowly as $\mathcal{O}(N^{-\frac{1}{2}})$ when the number N of available samples increases. We are interested in computing a more precise and unbiased estimator of $\boldsymbol{\mu}$ so that we need less simulations $\boldsymbol{y}_n$ and thus less computational resources.

To this end, we can use fast surrogates $\boldsymbol{c}_n$ that are correlated with the costly simulations $\boldsymbol{y}_n$, $n = 1, \dots N$ by constructing the random vectors

$$\boldsymbol{x}_n(\boldsymbol{\beta}) = \boldsymbol{y}_n - \boldsymbol{\beta}\,(\boldsymbol{c}_n - \boldsymbol{\mu_c})\,, \tag{2}$$

with *control matrix* $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$.

The *control variates* estimator is then the sample mean of N samples from equation (2)

$$\hat{\boldsymbol{\mu}}(\boldsymbol{\beta}) = \bar{\boldsymbol{x}}(\boldsymbol{\beta}) = \bar{\boldsymbol{y}} - \boldsymbol{\beta}\,(\bar{\boldsymbol{c}} - \boldsymbol{\mu_c})\,. \tag{3}$$

This estimator is unbiased by construction, $\mathbb{E}[\hat{\boldsymbol{\mu}}(\boldsymbol{\beta})] = \boldsymbol{\mu}$ regardless of any bias in the surrogates. The unique choice

$$\boldsymbol{\beta}^\star = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}} \det\left(\boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta})\right) = \boldsymbol{\Sigma}_{\boldsymbol{yc}}\boldsymbol{\Sigma}_{\boldsymbol{cc}}^{-1}\,. \tag{4}$$

gives the minimum variance estimator (Rubinstein & Marcus 1985; see CWAV20 for a Bayesian derivation).

As shown in CWAV20, for many practical applications where $p >> 1$, imposing structure on $\boldsymbol{\beta}$ can lead to large reductions in computational cost when $\boldsymbol{\beta}$ must be estimated from simulation/surrogate pairs. In the following, we will first consider diagonal $\boldsymbol{\beta}$ and then the more general case of sparse $\boldsymbol{\beta}$.

2.1.2 Diagonal case

When $\boldsymbol{\beta}$ is diagonal, the problem reduces to estimating p independent quantities. It is easy to prove (CWAV20) that there exists a single control coefficient $\beta^\star$ that minimizes the variance of the resulting random variable

$$x(\beta) = y - \beta\,(c - \mu_c)$$
$$\beta^\star = \underset{\beta\in\mathbb{R}}{\operatorname{argmin}}\, \sigma^2_{x(\beta)} = \frac{\operatorname{cov}(y,\,c)}{\sigma_c^2}\,, \tag{5}$$

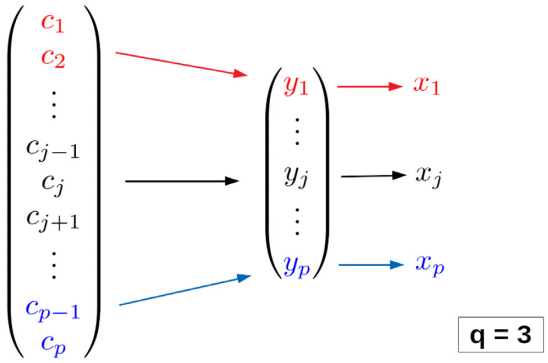

Figure 1. Illustration of sparse control matrix $\boldsymbol{\beta}$ using $q = 3$ elements of the surrogate $\boldsymbol{c}$ to reduce variance on an element of the simulation $\boldsymbol{y}$. In this example, we use adjacent elements of the surrogate; the boundary cases have $q = 2$.

and with subsequent variance reduction

$$\frac{\sigma^2_{x(\beta^\star)}}{\sigma_y^2} = 1 - \rho^2_{y,c}\,. \tag{6}$$

$\rho_{y,\,c}$ is the Pearson correlation coefficient between y and c. This case is equivalent to estimating the multivariate quantity $\boldsymbol{\mu}$, equation (3), while replacing $\boldsymbol{\beta}^\star$ with

$$\boldsymbol{\beta}^{\boldsymbol{diag}} = \operatorname{diag}\left(\frac{cov(y_1, c_1)}{\sigma^2_{c_1}}, \frac{cov(y_2, c_2)}{\sigma^2_{c_2}}, \dots, \frac{cov(y_p, c_p)}{\sigma^2_{c_p}}\right)\,. \tag{7}$$

CWAV20 applied this diagonal case to the estimation of the mean of the matter power spectrum, the matter bispectrum, and the one-dimensional matter PDF.

2.1.3 Sparse case

In some applications, multiple elements in $\boldsymbol{c}$ can help reduce variance of any element of $\boldsymbol{y}$. In this case, the problem reduces to p separate estimates with q control variates each

$$x(\beta) = y - \boldsymbol{\beta}^{\boldsymbol{T}}\,(\boldsymbol{c} - \boldsymbol{\mu_c})\,, \tag{8}$$

with $\boldsymbol{\beta} \in \mathbb{R}^q$. The optimal choice is

$$\boldsymbol{\beta}^\star_{\boldsymbol{m}} = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{cc}}\boldsymbol{\sigma}_{\boldsymbol{yc}} \tag{9}$$

with $\boldsymbol{\sigma}_{\boldsymbol{yc}}$, the q-dimensional column vector of covariances defined by $\boldsymbol{\sigma}_{\boldsymbol{yc}}[i] = cov(y, c_i)$, $1 \leq i \leq q$ (Lavenberg & Welch 1981; Rubinstein & Marcus 1985; Glynn & Szechtman 2002).

The attainable variance reduction with $\boldsymbol{\beta}^\star_{\boldsymbol{m}}$ is

$$\frac{\sigma^2_{x(\boldsymbol{\beta}^\star_{\boldsymbol{m}})}}{\sigma_y^2} = 1 - \frac{\boldsymbol{\sigma}^{\boldsymbol{T}}_{\boldsymbol{yc}}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{cc}}\boldsymbol{\sigma}_{\boldsymbol{yc}}}{\sigma_y^2}\,. \tag{10}$$

The Bayesian derivation in CWAV20 explains the form of the right-hand side of equation (10) as the ratio of the conditional and marginal covariances of y. The conclusion remains the same as before: the more the correlation, the smaller the variance. The idea is to use more of the elements of $\boldsymbol{c}$ to improve the variance reduction on y at the cost of estimating the vector $\boldsymbol{\beta}^\star_{\boldsymbol{m}}$ from equation (9) instead of the scalar β from equation (5). Fig. 1 illustrates the principle for $q = 3$ neighbouring surrogate variables per simulation variable. From now on, we will consider $\dim(\boldsymbol{y}) = \dim(\boldsymbol{c}) = p$ and have q designate the number of surrogates elements taken for each y in the sparse case.

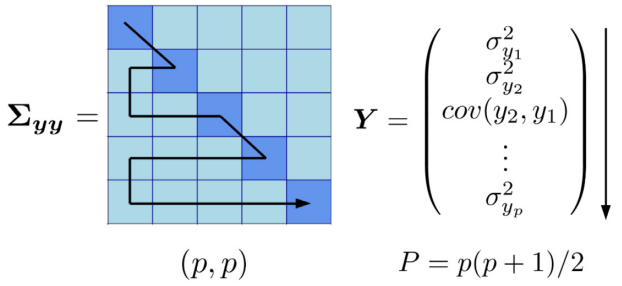

Figure 2. We generalize the CARPool method to covariance matrix estimation using the symmetric matrix vectorization above. This turns a $p \times p$ symmetric matrix into a vector of its $p(p+1)/2$ unique elements.

2.2 Application to covariance

In this section, we rewrite the covariance estimation problem as an instance of the CARPool method described above. We can replace the vectors $\boldsymbol{y}$ and $\boldsymbol{c}$ of size p by $\boldsymbol{Y}$ and $\boldsymbol{C}$ of size $P = p(p+1)/2$, representing the covariance matrix elements. If $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ are realizations of the random vector $\boldsymbol{y}$, then the samples $\{\boldsymbol{y}_1 - \bar{\boldsymbol{y}}, \ldots, \boldsymbol{y}_N - \bar{\boldsymbol{y}}\} \equiv \{\widetilde{\boldsymbol{y}}_1, \ldots, \widetilde{\boldsymbol{y}}_N\}$ are also realizations of a multivariate random variable; and so are the outer products $\{\widetilde{\boldsymbol{y}}_1 \otimes \widetilde{\boldsymbol{y}}_1, \ldots, \widetilde{\boldsymbol{y}}_N \otimes \widetilde{\boldsymbol{y}}_N\} \equiv \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ with P unique elements. We rewrite the sample covariance matrix in terms of these outer products

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{yy} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \bar{\boldsymbol{y}}\right)\left(\boldsymbol{y}_i - \bar{\boldsymbol{y}}\right)^{\boldsymbol{T}} \\ &\equiv \frac{N}{N-1} \times \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Y}_i \\ &\equiv \gamma\overline{\boldsymbol{Y}}\,, \end{aligned} \tag{11}$$

with Bessel's correction factor $\gamma = N/(N-1)$. This is an unbiased estimator of the true covariance $\boldsymbol{\Sigma}_{yy}$. Similarly, the surrogate samples $\boldsymbol{C}_i$ have sample mean

$$\overline{\boldsymbol{C}} \equiv \frac{1}{\gamma}\widehat{\boldsymbol{\Sigma}}_{cc} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{C}_i\,. \tag{12}$$

Note the additional constraint of $\boldsymbol{y}$ and $\boldsymbol{c}$ having finite fourth-order moments (i.e variance of the covariance).

Fig. 2 explains how we build such vectors from symmetric matrices by ensuring matrix elements remain neighbours in the resulting vector.[2]

For simplicity, in equations (11), (12), and (13), we have identified the capital bold vectors of size P with their reconstruction into a symmetric (p, p) matrix to emphasize that the CARPool covariance matrix can be framed as a standard CARPool estimate. From this point, we also drop the 'hat' of the estimated quantities for notational simplicity.

The CARPool covariance estimate is then simply an instance of the CARPool method, equation (3), applied to the vectorized sample covariances

$$\boldsymbol{\Sigma}_{yy}(\boldsymbol{\beta}) = \gamma\overline{\boldsymbol{X}}(\boldsymbol{\beta}) = \gamma\overline{\boldsymbol{Y}} - \gamma\boldsymbol{\beta}\left(\overline{\boldsymbol{C}} - \boldsymbol{\mathcal{M}}_C\right)\,, \tag{13}$$

where $\boldsymbol{\mathcal{M}}_C$ plays the role of $\boldsymbol{\mu}_c$ in the case of covariance estimation, that is to say it is the surrogate covariance matrix computed from a separate set of surrogate realizations. The estimator is unbiased by construction, $\mathbb{E}\left[\boldsymbol{\Sigma}_{yy}(\boldsymbol{\beta})\right] = \boldsymbol{\Sigma}_{yy}$, and corrected by the factor γ.

Note that the sample covariance and the unbiased estimator of the covariance from equation (13) do not necessarily yield, when inverted, an unbiased estimate of the precision. Hartlap et al. (2007) point out a correction factor (known as the 'Hartlap factor' in cosmology) for data sampled from a multivariate Gaussian distribution that has been widely used. In recent works, impacts of biases in the estimated precision matrix up until parameter constraints and shifts, as well as methods to improve the estimation, have been emphasized, e.g. in Friedrich & Eifler (2018), Percival et al. (2021), and Philcox et al. (2021). In the following numerical experiments, we will assess the performance of the inverse of our newly proposed covariance estimate through multiple tests: by visual comparison, through studying the eigenvalues of the covariances (which are simply the reciprocals of the eigenvalues of the precision) and by computing the Fisher matrices for parameter constraints, which are computed through the precision. A detailed description of tests will be given in Section 3.2.

Also, unlike the sample covariance matrix, the estimate $\boldsymbol{\Sigma}_{yy}(\boldsymbol{\beta})$ is not guaranteed to be positive semi-definite by construction for a finite number of samples, even though its expectation is. We will comment further on positive-definiteness in the numerical results of Section 3.3.

3 NUMERICAL EXPERIMENTS

Our numerical analysis assume a λcold dark matter (ΛCDM) cosmology congruent with the *Planck* constraints from Planck Collaboration VI (2020): $\Omega_{\rm m} = 0.3175$, $\Omega_{\rm b} = 0.049$, $h = 0.6711$, $n_s = 0.9624$, $\sigma_8 = 0.834$, $w = -1.0$, and $M_\nu = 0.0$ eV. For a reminder about how to apply CARPool, see Figure 1 in CWAV20: the principle stays the same, except the vectorized outer products of centered data, as explained above, play the role of the data samples to estimate the covariance matrix. The numerical analysis presented below compares the following unbiased covariance matrix estimators:

(i) GADGET, where we compute the sample covariance $\boldsymbol{\Sigma}_{yy}$ with equation (11) from N-body simulations only.

(ii) Diagonal CARPool applied individually to each unique covariance matrix element estimator, with equation (5) applied to the vectorized covariance $\boldsymbol{X}$. This framework can simply be referred to as '$q = 1$'.

(iii) Sparse CARPool, where we estimate each simulation covariance matrix element with $q > 1$ surrogate matrix elements according to equation (8).

We stress that for the CARPool estimate, we compute the control matrix $\boldsymbol{\beta}$ from the same N simulations entering $\bar{\boldsymbol{Y}}$ in equation (13).

3.1 Simulation and surrogate data

We briefly describe here the chosen simulation and surrogate solvers we use; for details, please refer to the numerical experiments of CWAV20 where the same simulations and surrogates are used. The solvers evolve $\mathcal{N}_{\rm p} = 512^3$ CDM particles in a box volume of $(1000\,{\rm h}^{-1}\,{\rm Mpc})^3$. The simulation-surrogate sample pairs take the same 2LPT initial conditions at starting redshift $z_s = 127.0$.

[2]Alternatively, NUMPY provides the function `tril_indices` (`triu_indices`) to simply extract the lower (upper) triangular part of a 2D-array in a row-major order.

3.1.1 N-*body solver*

The simulation outputs are part of the publicly available *Quijote* simulation suite[3] (Villaescusa-Navarro et al. 2020). The full *N*-body simulations were run with the TreePM code GADGET-III, stemming from GADGET-II by Springel (2005). In the following, we will use all 15 000 available realizations of the fiducial cosmology to evaluate the quality of the various estimates. We will take as the simulation 'ground truth' the sample mean and the sample covariance based on 12 000 of these simulations while retaining 3000 simulations to test likelihoods built using the various estimators, see Section 3.2.2. The force mesh grid size of all the simulations is $\mathcal{N}_{\rm m} = 1024$.

3.1.2 Surrogate solver

We generate the fast surrogate samples with The COmoving Lagrangian Acceleration (COLA) method from Tassev et al. (2013), which allows generating approximate gravitational *N*-body outputs using a smaller number of timesteps than our simulation code. The principle of COLA is to add residual displacements computed with a PARTICLE-MESH (PM) *N*-body solver to the trajectory given by analytical LPT approximations. See Izard, Crocce & Fosalba (2016) for comparisons of the capabilities and computational cost of COLA against *N*-body simulations in different configurations. Like in CWAV20, we used `L-PICOLA`, a publicly available and parallel (MPI) code implementation of COLA developed by Howlett, Manera & Percival (2015), with a force mesh grid size $N_m^{\rm cola} = 512$. We computed the matter power spectra, correlation functions and PDFs using `Pylians3` [4] and the matter bispectra with `pySpectrum`.[5]

3.2 Description of tests

The goal of this section is to briefly explain the different tests we have implemented to assess the reliability of the CARPool covariance estimates, and to compare with the standard (bias-corrected) sample covariance matrix from simulations only.

3.2.1 Variance reduction on matrix elements

This test, similarly to what was assessed concerning the mean of clustering statistics in CWAV20, consists in taking the empirical variance ratio, between a set of $\boldsymbol{Y}_i$ samples (vectorized outer products of centered data) and the corresponding $\boldsymbol{X}_i(\boldsymbol{\beta})$ samples, with $\boldsymbol{\beta}$ being estimated from the main set of paired simulation/surrogate samples. More precisely, in the experiments, the same set of seeds $\{s_i, i \in [\![1, 500]\!]\}$ serves to compute both $\boldsymbol{\beta}$ and the actual covariance estimate from equation (13), while paired simulation/surrogate samples from seeds $\{s_i, i \in [\![501, 2000]\!]\}$ are used to estimate $Var(\boldsymbol{Y})$ and $Var(\boldsymbol{X})$.

3.2.2 Negative Gaussian log-likelihood on test data

Under the assumption of Gaussianity of the matter statistics, the negative log-likelihood gives a loss-function of the test data $\{y_i\}_{\rm test}$ from simulations. In equation (14), the input is the covariance matrix (sample covariance or CARPool covariance) from 'training' data, i.e. the seeds we reserve for covariance estimation including the 'truth', while the $\boldsymbol{y}_i$ are the clustering statistics from test data simulations and $\boldsymbol{\mu}$ is replaced by $\bar{\boldsymbol{y}}$. More precisely, the sample covariance from the first 12 000 fiducial seeds of the *Quijote simulations* computes the 'true' covariance and gives the reference negative log-likelihood while the remaining seeds $\{s_i, i \in [\![12001, 15000]\!]\}$ constitute the unseen test data

$$-\ln\left[\mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{yy}\right)\right] = \frac{pN}{2}\ln\left[2\pi\right] + \frac{N}{2}\ln\left[\det\left(\boldsymbol{\Sigma}_{yy}\right)\right] + \frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)^{\boldsymbol{T}}\boldsymbol{\Sigma}_{yy}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}\right). \quad (14)$$

In Section 3.3, we show the convergence of the log-likelihood with increasing number of simulations. The loss function is computed only if the corresponding covariance estimate is positive-definite; in the numerical tests, we will see that for different statistics, the CARPool covariance estimates become positive-definite for a different minimum number of samples.

3.2.3 Eigenspectrum

We compare the eigenvalues of the sample covariance from simulations with these of the CARPool covariance from simulation/surrogate pairs. It is well-known that the sample covariance matrix from equation (11) of a vector statistics of size p is at most of rank $N - 1$ with N samples. Moreover, as demonstrated by Bai & Yin (1993), among others, the sample covariance matrix for $N \sim p$ is ill-conditioned even when full-rank: the smallest eigenvalues in particular disperse from the true values and are biased low even for unbiased estimates of the covariance matrix elements. Therefore, computing the ratio of the ordered eigenvalues of an unbiased covariance estimate and of the 'ground truth' covariance is a relevant indicator of the quality of the estimation. We stress that this test does not constitute a complete comparison of these matrices since the eigenbasis could still differ. Still, comparing the eigenspectra of two covariance estimates of the same random vector is common practise (e.g. Joachimi 2017; Pope & Szapudi 2008) and can be considered in the context of the other, complementary tests we show.

3.2.4 Fisher analysis and parameter uncertainties

The covariance matrix plays a central role in parameter inference and when forecasting parameter constraints for future data sets. In this context, the relevant performance metric is not the convergence of the covariance matrix in isolation but the way parameter-driven variations in the measured quantities are constrained. This is precisely what the Fisher information matrix measures, which is why it is an essential component in parameter estimation and forecasts.

For this test, we will model the likelihood of the simulated observable as a multivariate Gaussian with $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta})\right)$. In this approximation,[6] the Fisher matrix is the symmetric matrix of size (d, d)

$$\mathcal{F}_{ij} = \left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_i}\right)^{\boldsymbol{T}}\boldsymbol{\Sigma}_{yy}^{-1}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_j}\right). \quad (15)$$

[3] https://github.com/franciscovillaescusa/Quijote-simulations
[4] https://github.com/franciscovillaescusa/Pylians3
[5] Available at https://github.com/changhoonhahn/pySpectrum

[6] For Gaussian data whose covariance depends on the parameter values, equation (15) would include a second term. This term vanishes for the Gaussian approximation we consider here, where the statistics are approximated as Gaussian with constant covariance around a parameter-dependent mean (see Alsing & Wandelt 2018 for a succinct explanation and Carron 2013 and Kodwani, Alonso & Ferreira 2019 for further discussion).

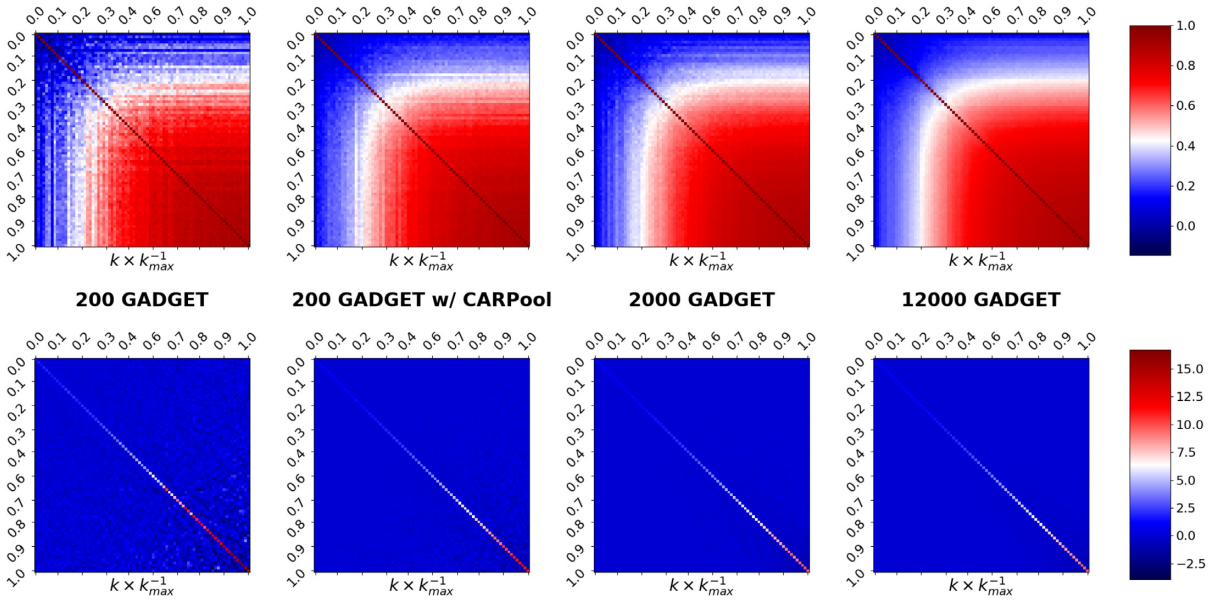

Figure 3. We display multiple power spectrum covariance estimates (top panels) and their inverse (bottom panels) in order to illustrate the improvement on the standard sample covariance matrix. For reference, the power spectrum is shown on the same axes in Fig. 4. Covariance matrices are shown as their correlation counterpart $\boldsymbol{D}^{-1}\widehat{\boldsymbol{\Sigma}}\boldsymbol{D}^{-1}$ with the diagonal $\boldsymbol{D} = \sqrt{\text{diag}\left(\widehat{\boldsymbol{\Sigma}}\right)}$. The precision matrices at the bottom are the inverse of the corresponding correlation matrix at the top.

We define the vector of parameters as $\boldsymbol{\theta} = (\Omega_m, \Omega_b, h, n_s, \sigma_8, M_\nu)^{\boldsymbol{T}}$, and $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{y}] = \boldsymbol{\mu}(\boldsymbol{\theta})$ is the expectation of $\boldsymbol{y}$ for fixed parameters $\boldsymbol{\theta}$.

The Cramér–Rao inequality then gives the lower bound of the variance of an unbiased estimator for parameter θ_i, marginalized over the other parameters

$$\sigma^2_{\theta_i} \geq \left[\mathcal{F}^{-1}\right]_{ii}. \tag{16}$$

The partial derivatives of the statistics are estimated numerically using finite differences from 500 *Quijote* simulations for each varying parameter exactly as in Villaescusa-Navarro et al. (2020) (see Table 1 of this work for the parameter values). When finite difference simulations are not already available, one can easily apply the CARPool method to the estimation of the mean of the derivatives: in CWAV20, especially for the matter power spectrum and bispectrum, the precision of the CARPool mean with 5 N-body simulations was comparable to that of the mean of 500 simulations. We do not further explore this application of CARPool, since the focus of this paper is covariance estimation.

3.3 Results on clustering statistics covariance at $z = 0.5$

Prior to extracting clustering statistics on each snapshot, we compute the matter overdensity $\rho(r = |\boldsymbol{r}|)$ on a grid with $\boldsymbol{r}$ the comoving-coordinates in h^{-1}Mpc. The density contrast field is then $\delta(\boldsymbol{r}) \equiv \rho(\boldsymbol{r})/\bar{\rho} - 1$. We present results at redshift $z = 0.5$, which is approximately the lowest redshift that is relevant for upcoming galaxy surveys of the large-scale structure. We found higher correlation for some statistics (power spectrum) at $z = 0.0$ than $z = 0.5$ and interpret that as the erasure by the non-linearities of discrepancies in the intermediate structure growth. Thus, the $z = 0.5$ case may be close to the worst case and we expect CARPool to be even more efficient both for higher and for lower redshifts, either for the mean estimation like in CWAV20 or for the covariance matrix in this study. The tests described in Section 3.2, for each clustering statistics, allow examining both the covariance estimator from equation (13) as well as its inverse, as an estimator of the precision matrix. We will assess the eigenspectrum, the negative log-likelihood loss function from equation (2), and the Fisher matrix as a proxy for the adequacy of the covariance matrix estimate for deriving parameter constraints. We also show the element-by-element variance reduction on the new estimate (performance of CARPool) and plot the covariance and precision matrices to provide a visual cue of the reduction in noise with respect to the sample covariance. As discussed in Section 2.2, the precision estimate $\boldsymbol{\Sigma}^{-1}_{yy}$ will not include the Hartlap factor, whether for the sample covariance or the CARPool covariance.

3.3.1 *Matter power spectrum*

The density contrast $\delta(\boldsymbol{x})$ is computed on a square grid of size $N_{\text{grid}} = 1024$ for each snapshot, then, in 3D Fourier space, the average of $|\delta(k)|^2$, $k \in [k - \Delta k, k + \Delta k]$ gives the power spectrum $P(k)$ for wave vector modulus k. The *Quijote* power spectra range from $k_{\text{min}} = 8.900 \times 10^{-3}\,\text{h}\,\text{Mpc}^{-1}$ to $k_{\text{max}} = 5.569\,\text{h}\,\text{Mpc}^{-1}$. The following analysis is restricted between $k_{\text{min}} = 8.900 \times 10^{-3}\,\text{h}\,\text{Mpc}^{-1}$ and $k_{\text{max}} \approx 1.0\,\text{h}\,\text{Mpc}^{-1}$, which results in $p = 79$ linearly spaced bins. Therefore, we have $P = 3160$ unique covariance matrix elements to estimate.

Fig. 3 shows the CARPool estimate of the covariance and the precision matrix using 200 simulations. For comparison, we show the sample covariance estimates for 200 and 2000 simulations as well as the 'ground truth' covariance measured from 12 000 GADGET simulations. The empirical variance reduction on each estimated covariance matrix element appears in Fig. 4: as expected, the variance reduction on the $\boldsymbol{X}(\boldsymbol{\beta})$ samples at large scales is much higher ($\sim 10^4$) than for variances and cross-covariances at small scales ($\sim$10-fold reduction). In Fig. 5, we see that the $q = 1$ case is positive definite from $N = 80$ simulations onward while with $q > 1$ (we stopped at $q = 3$ for the power spectrum) attains positive definiteness at $\sim$120

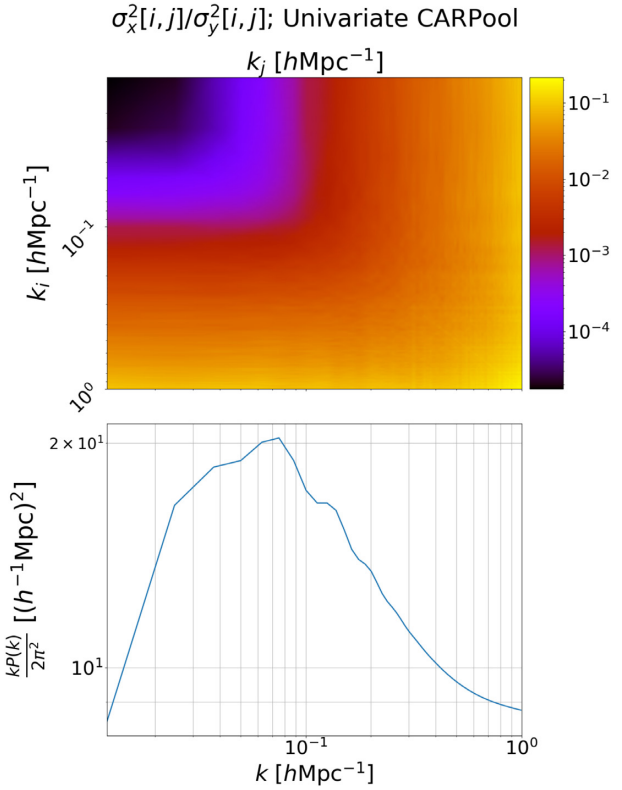

Figure 4. Top panel: We show the variance reduction with CARPool of the estimated power spectrum second-order moments up to $k_{\max} \approx 1.0\,\mathrm{h\,Mpc^{-1}}$. The control matrix that generates 3200 $\boldsymbol{X}_n$ CARPool samples is estimated with 200 paired realizations. The variance of the covariance elements of $\boldsymbol{\Sigma}_{yy}$ are estimated using 3200 available power spectra from the *Quijote* simulations. Bottom panel: for reference, reduced power spectrum from the mean of the *Quijote* simulations at $z = 0.5$.

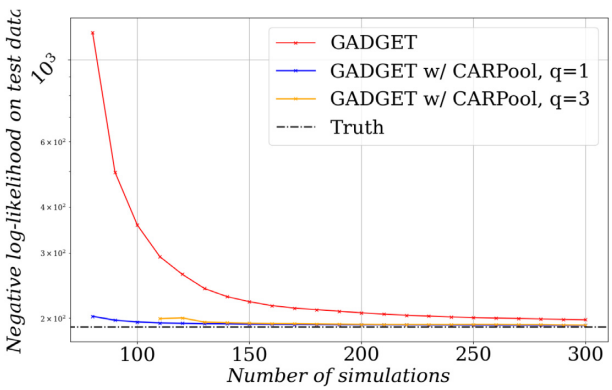

Figure 5. Negative log-likelihood on test data – acting as a cost function – evaluated for an increasing number of available N-body simulations used to estimate the covariance matrix of the matter power spectrum. We observe that the CARPool estimates converge much faster towards the true value of the cost function.

simulations and offers no improvement on the loss function on test data, at least for a small to moderate number of simulations and for our choice of neighbourhood induced by the vectorization of the covariance matrix, cf. Fig. 2. The log-likelihood of the CARPool covariance converges much faster to the log-likelihood of the 'ground truth' covariance than that of the sample covariance based on the same number of simulations. Additionally, Fig. 6 demonstrates

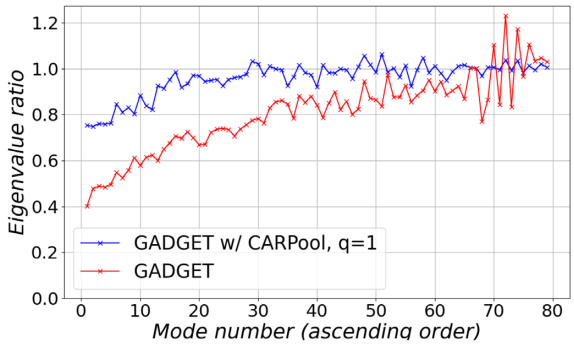

Figure 6. We show the improvement on the conditioning of the covariance estimate with CARPool by showing the ratio of the eigenvalues in ascending order between the estimated covariance using 200 simulations and the 'true' covariance matrix estimated with 12 000 simulations: $\lambda_i^{\mathrm{test}}/\lambda_i^{\mathrm{true}}$ for each index i. A constant line at 1 would indicate identical eigenspectra but would not imply that the eigenbases are the same, as discussed in Section 3.2.

that for the same number of simulations ($N = 200$ in this plot), the eigenvalue ratio greatly favours the CARPool estimate: small eigenvalues, which are the last to converge when using the sample covariance, are lifted up and the largest modes are more stable.

How does this improvement in the covariance matrix translate to the Fisher matrix? We can see in Fig. 7 that, with respect to the Fisher matrix computed using the 'ground truth' covariance, the sample covariance of size (79,79) using 200 simulations leads to a significant underestimate and, in some cases, rotation of the confidence regions of the parameters. The CARPool covariance using the same number of simulations (plus the paired and additional surrogate samples) gives a considerably more accurate Fisher matrix.

3.3.2 *Matter bispectrum*

In this subsection, we turn to estimating the covariance of the matter bispectrum. Like in CWAV20, we will consider two separate subsets of the matter bispectrum: the matter bispectrum monopole $B(k_1, k_2, k_3)$ of squeezed isosceles triangles on the one hand, and the reduced bispectrum monopole $Q(k_1, k_2, k_3)$ for equilateral configurations on the other hand. We will apply the same tests we used for the power spectrum covariance matrix, looking at the variance reduction, the negative log-likelihood, the eigenspectrum, and the Fisher matrix computed from the estimated bispectrum covariance matrix.

3.3.3 *Squeezed isosceles triangles*

We build the first group of samples by grouping triangle configurations for which $k_1 = k_2$ and by ordering the bispectrum monopoles in ascending order of the k_3/k_1 ratio. We keep squeezed triangles: $(k_3/k_1)_{\max} = 0.20$ ($p = 98$ and $P = 3851$). Since $q = 3$ gives a slight improvement over $q = 1$, the figures will show the results for $q = 3$.

Fig. 8 compares the CARPool covariance estimate with the sample covariance estimator applied 200, and 2000 GADGET simulations with the 'ground truth' computed from 12 000 simulations. The differences are visually most apparent in the precision matrix, where the CARPool estimate from 200 simulations looks visually similar to the standard estimate from 200 simulations. Fig. 9 gives a quantitative view of the CARPool variance reduction of the covariance matrix elements. As for the power spectrum covariance, CARPool also improves the eigenvalues of the covariance matrix (Fig. 11).

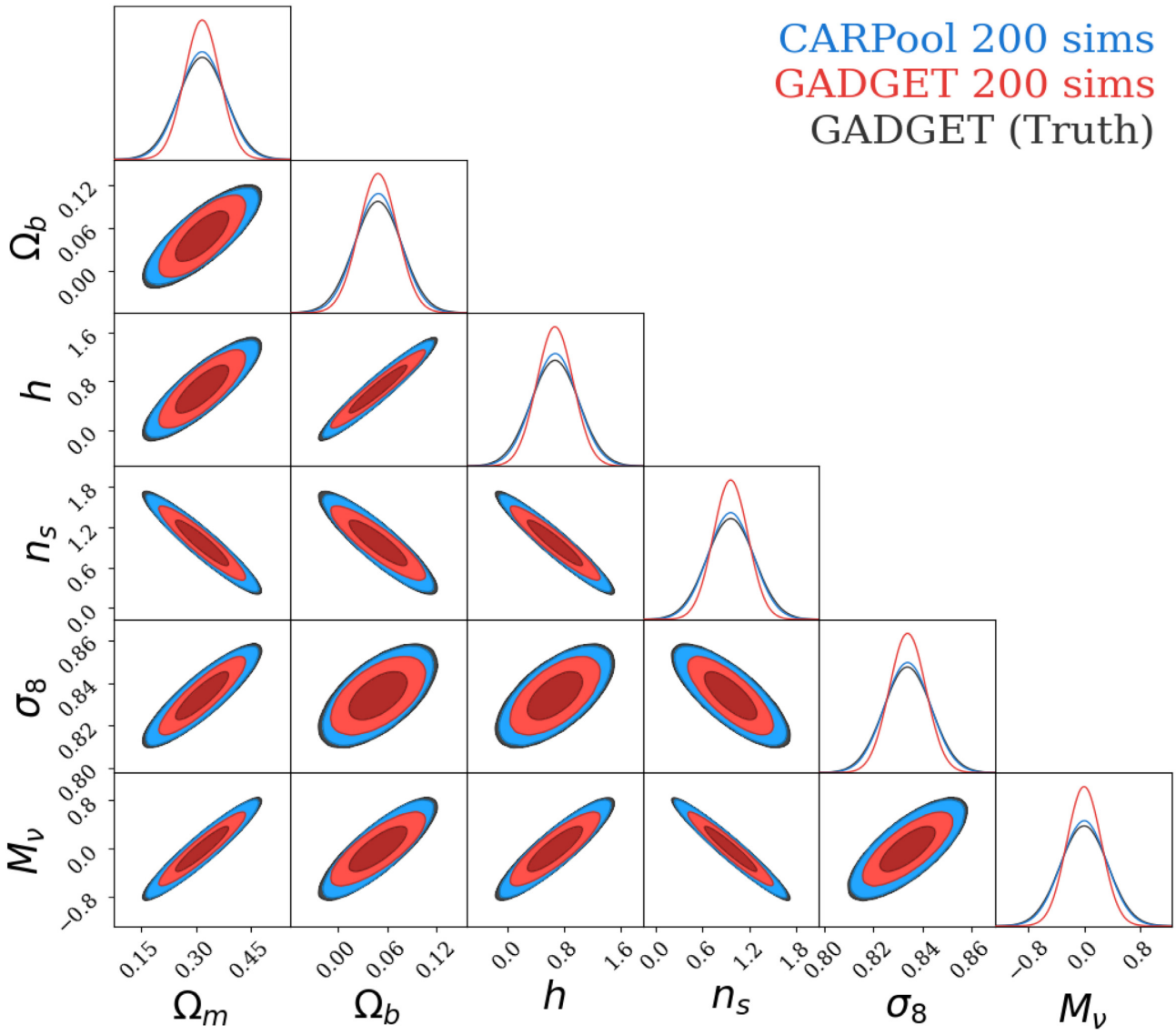

Figure 7. Confidence contours of the cosmological parameters computed using the Fisher matrix based on the estimated matter power spectrum covariance matrix. The 'truth' designates the confidence region using the sample covariance matrix of 12 000 N-body simulations, and the mean parameters are known from the ΛCDM models used in the simulations. We, thus, demonstrate the better conditioning of the CARPool covariance estimate with respect to the sample covariance for the same number of simulations.

For this case, the convergence of $q = 1$ and 3 CARPool estimates in term of negative log-likelihood of test data appears in Fig. 10. We also tested $q = 5$, but in that case the added noise in the estimate of $\boldsymbol{\beta}$, now a (5,5) matrix, for each empirical counterpart of equation (9) worsens the performance if we limit ourselves to a moderate number of simulations.

Fig. 12 shows that the Fisher matrix computed using the CARPool covariance based on 200 simulations is much more accurate even than the sample covariance using 300 simulations, where confidence contours are underestimated compared to the 'ground truth' reference computed based on 12 000 simulations.[7]

[7]We did not include M_ν here; the fiducial realizations of the bispectrum initialized with the Zel'dovich approximation that are needed to compute the neutrino mass derivative as in Villaescusa-Navarro et al. (2020) were not available.

3.3.4 Equilateral triangles

The second set of bispectrum statistics is comprised of equilateral triangles with $k_1 = k_2 = k_3$ varying up to $k_{\rm max} = 0.75\,{\rm h\,Mpc^{-1}}$ ($p = 40$ and $P = 820$). CARPool gives a particularly strong variance reduction for this case with a smaller p than before, so we focus on a case with only 100 simulations. Fig. 13 visually compares the covariance matrix estimators. Fig. 14 shows the strong variance reduction on the covariance matrix elements from $\sim \mathcal{O}(10^4)$ at large scales down to $\sim \mathcal{O}(10)$ at small scales. Fig. 15 emphasizes the improvements of the log-likelihood test, with the simplest, diagonal ($q = 1$) CARPool estimator being favoured (10 more simulations than the sample covariance are required for that in the $q = 3$ case). The eigenvalue ratios for the CARPool estimate, in Fig. 16, approach the 'ground truth' even with only 100 N-body simulations, except for seven smallest eigenvalues. The Fisher analysis presented in Fig. 17 exhibits the same behaviour as for the previous statistics: the sample covariance with few simulations underestimates parameter confi-

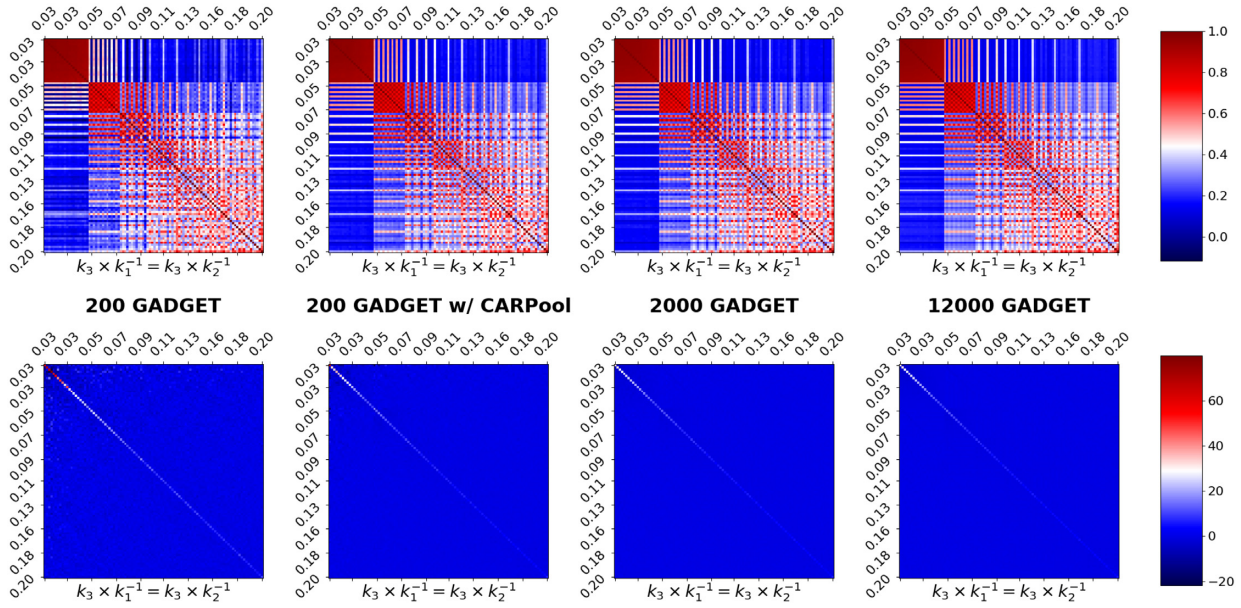

Figure 8. We plot different matter bispectrum covariance estimates (top panels) and their inverse (bottom panels), for squeezed isosceles triangles, similarly to Fig. 3.

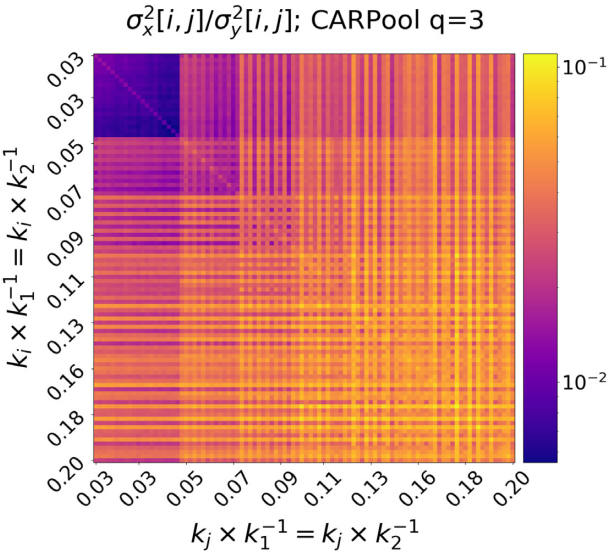

Figure 9. We demonstrate there is a significant variance reduction of the estimated matter bispectrum covariance matrix elements, for the set of squeezed isosceles triangles up to $(k_3/k_1)_{\max} = 0.2$. The control matrix that generates 1800 new $\boldsymbol{X}_n$ CARPool samples, to compute the variance of each vector element, is estimated with 200 paired realizations. To estimate the variance of each element of $\boldsymbol{\Sigma}_{yy}$, we use 1800 samples from *Quijote* simulations.

dence intervals, while the CARPool covariance with few simulation is much more representative of the knowledge about parameters given by the clustering statistic of interest. We note however, that the set of 40 equilateral triangles we treated is much less informative about the parameters than the other statistics we consider in this paper.

3.3.5 *Matter correlation function*

We also tested real-space clustering statistics, the first of which being the two-point matter correlation function $\xi(\boldsymbol{r})$ for $\boldsymbol{r} \in$

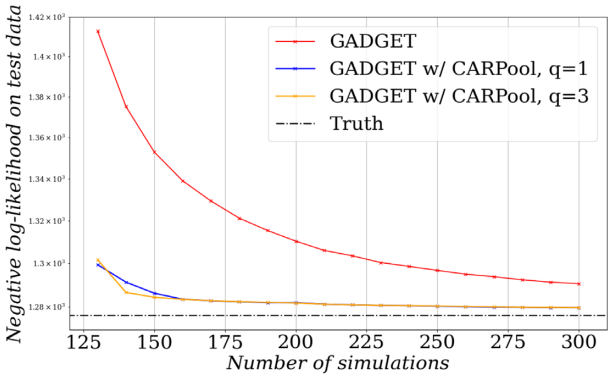

Figure 10. Negative log-likelihood on test data evaluated for an increasing number of simulations used to compute a covariance estimate, like in Fig. 5.

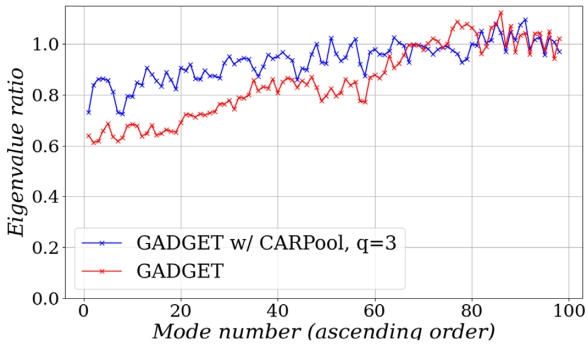

Figure 11. The computation method for the eigenvalue ratio of the matter bispectrum covariance is identical to Fig. 6.

$[5.0, 160.0]\,h^{-1}$Mpc ($p = 159$). As we did not experiment with the correlation function in CWAV20, we show the reduction of variance for the estimation of the mean in Fig. 18. With five *N*-body simulations and CARPool, we get an unbiased estimate of the

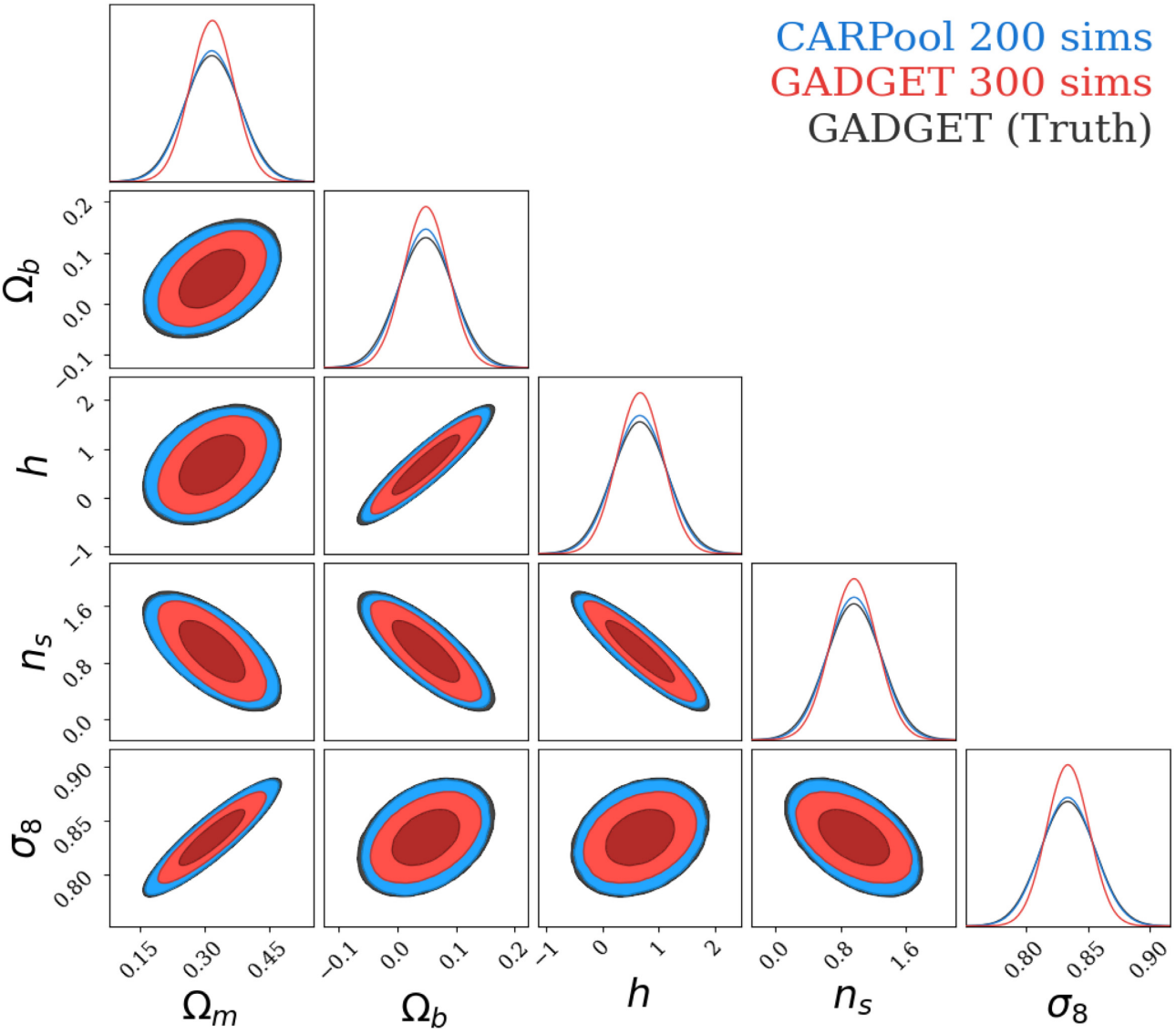

Figure 12. Confidence contours of the cosmological parameter computed using the Fisher matrix, based on the estimated squeezed matter bispectrum covariance matrix. Again, the CARPool covariance gives more realistic confidence regions than the sample covariance for an equivalent number of simulations. The 'truth' designates the confidence region using the sample covariance matrix of 12 000 *N*-body simulations; and the mean parameters are known from the ΛCDM models used in the simulations.

mean correlation function with an equivalent precision – in terms of 95 per cent confidence intervals – as with the sample mean of 500 simulations.

Then, regarding the covariance matrix, we show in Fig. 19 that we get consequential variance reduction on all the estimated second-order moments of the correlation function vector. We note then the reduction is 'homogeneous' in the matrix, since the highly correlated large-scale modes in Fourier space intervene at all scales in real-space by summation.

We did not, however, get significant improvement on the conditioning of the covariance matrix with CARPool comparatively to the sample covariance. In other words, CARPool does its job for the matter correlation function – that is to say reducing variance on covariance elements – but this improvement did not translate into a better eigenspectrum or strong improvements of the Fisher matrix contours.

3.3.6 Matter PDF

As in CWAV20, the matter PDF is computed on a grid with $N_{\rm grid} = 512$ and smoothed by a top-hat filter of radius $R = 5\,{\rm h}^{-1}\,{\rm Mpc}$. We have the raw 100 histogram bins in the range $\rho/\bar{\rho} \in \left[10^{-2}, 10^{2}\right]$. Given that the covariance for this case is formally degenerate since the histogram bins are linearly dependent – each bin can be written as 1 minus the sum of the others – we have taken all the bins that are non-zero across all samples up until the tails and down-sampled the PDF by a factor 2, which gives $p = 33$ bins. Even after this modification, the covariance is still nearly degenerate. There are also strong bin-to-bin correlations that suggest going to even coarser binning would improve the condition of the matrix; we proceeded without processing to test the CARPool covariance estimate in this regime.

Fig. 20 shows that the variance of the CARPool covariance estimate is only mildly reduced. A similar effect was seen in the

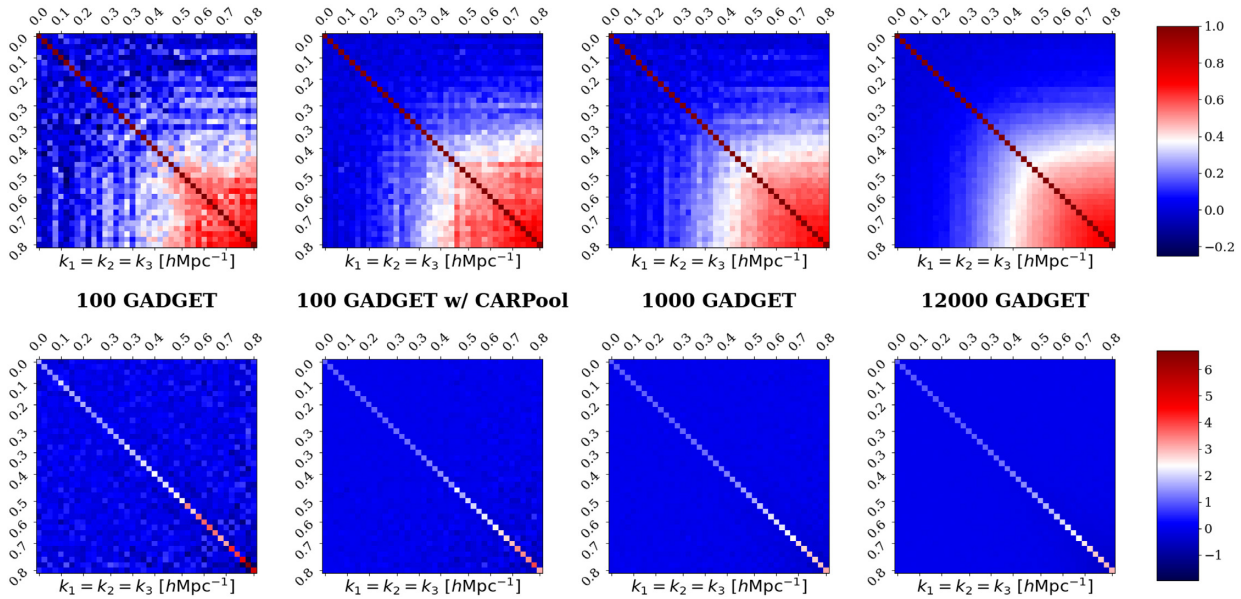

Figure 13. We plot covariance and precision matrices estimates for the reduced bispectrum of equilateral triangles, similarly to Fig. 3.

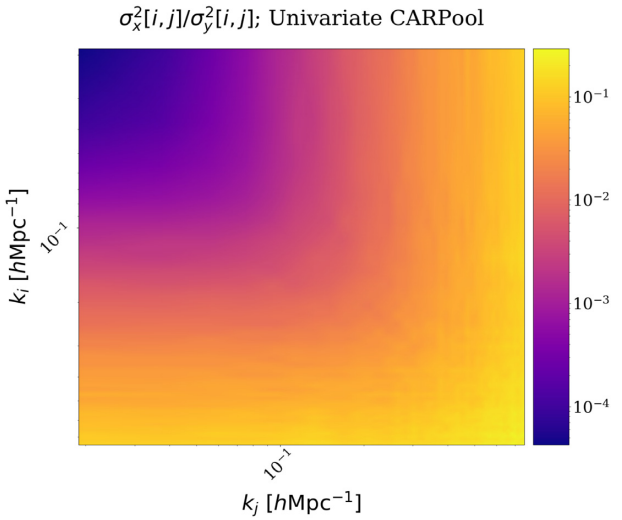

Figure 14. We exhibit the significant variance reduction for the estimated matter reduced bispectrum covariance matrix elements, for the set of equilateral triangles up to $k_1 = k_2 = k_3 \approx 0.75$. The computation of the metric is identical to Fig. 9.

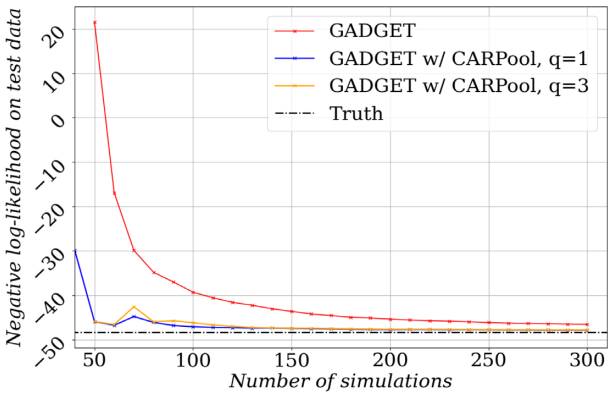

Figure 15. Negative log-likelihood on test data, evaluated for an increasing number of available N-body simulations that intervene in the estimation of the covariance matrix of the reduced matter bispectrum.

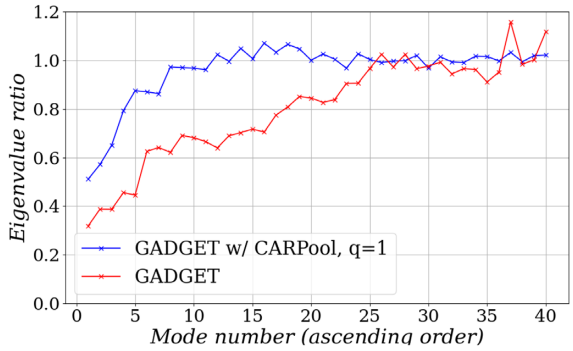

Figure 16. Same as Fig. 6, but for the reduced matter bispectrum of equilateral triangles covariance; note that in this case, only 100 simulations were used in both the standard and the CARPool estimates.

CARPool estimate of the mean PDF. Since the densities of structures in the COLA surrogates do not match the densities of the corresponding structures in the simulations, underestimating the density in haloes and overestimating underdensities in voids, fluctuations in density bins of the surrogate are correlated to fluctuations in other density bins of the simulation. A larger variance reduction would be obtained with a brute-force dense control matrix $\boldsymbol{\beta}$, though this would require a large number of simulations to estimate the control matrix and thus defeat the point of the approach. An alternative would be to define a pre-processing function to map the average density PDFs of the surrogates to approximately match the average PDF of the simulation, as in Leclercq et al. (2013). This would likely increase the bin-to-bin correlations for diagonal control matrix and therefore improve the CARPool estimates. We did not pursue these

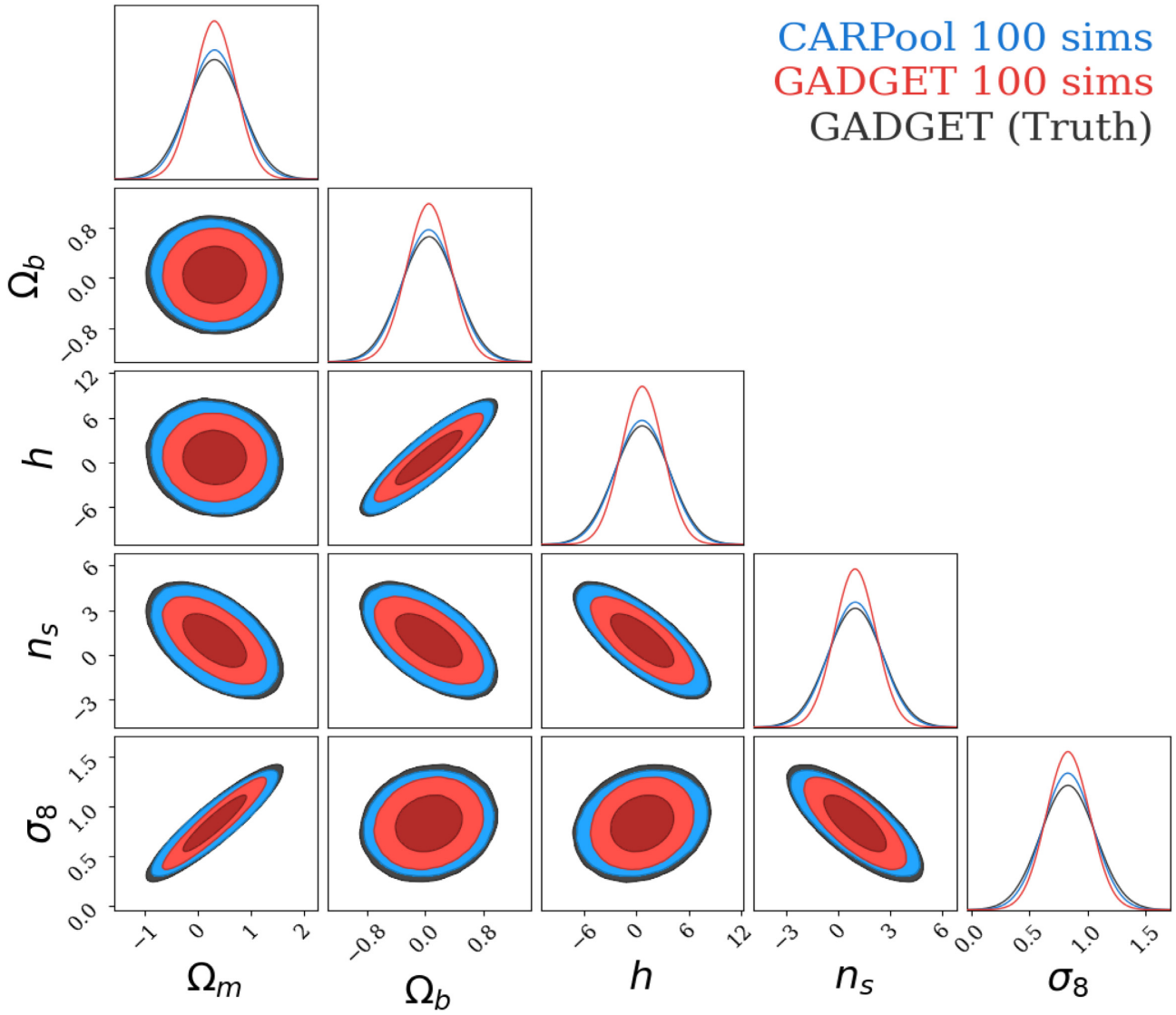

Figure 17. Same as Fig. 12, but based on the estimated equilateral matter bispectrum covariance matrix.

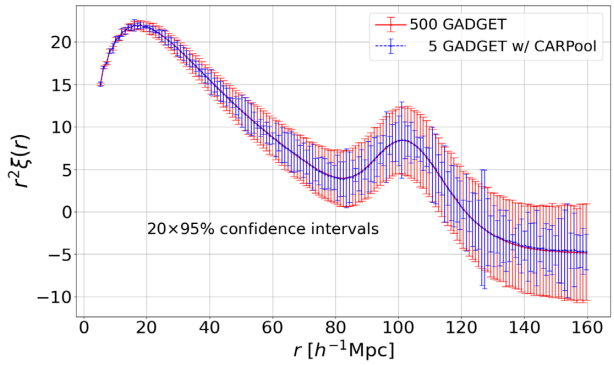

Figure 18. Estimated mean of the matter correlation function with 500 *N*-body simulations versus five pairs of '*N*-body + cheap' simulations. The estimated 95 per cent confidence intervals are computed with the Student *t*-score for CARPool bias-corrected and accelerated (BCa) bootstrap for the sample mean of simulations only.

ideas further in order to test the CARPool approach without designer pre-processing.

We found that as a consequence of the small variance reduction and the near-degeneracy of the covariance, the eigenspectrum of the resulting covariance estimates is not improved and in fact not positive-definite for all test realizations. We will discuss possible remedies to this issue below.

4 DISCUSSION AND CONCLUSION

We explore the problem of estimating covariance matrices as a new application of the CARPool principle, *i.e.*, of combining simulations and surrogates to accelerate Monte Carlo convergence introduced in CWAV20 and demonstrate it on multiple *N*-body simulation outputs. Our generalization uses a matrix vectorization procedure (*cf.* Fig. 2). All that is required to CARPool is a surrogate solver that can rapidly generate particle snapshots with minimal computational effort and a surrogate statistic which is strongly correlated with the one computed by the simulation code. As in CWAV20, CARPooling guarantees unbiased results for whatever quantity it is used to

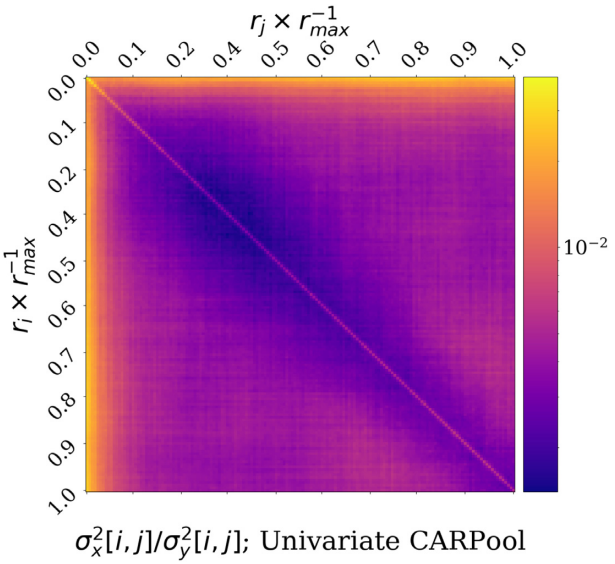

Figure 19. Variance reduction for the matter correlation function covariance matrix elements, for $r \in [5.0, 160.0]\,h^{-1}$ Mpc. The computation is similar to that of Fig. 4.

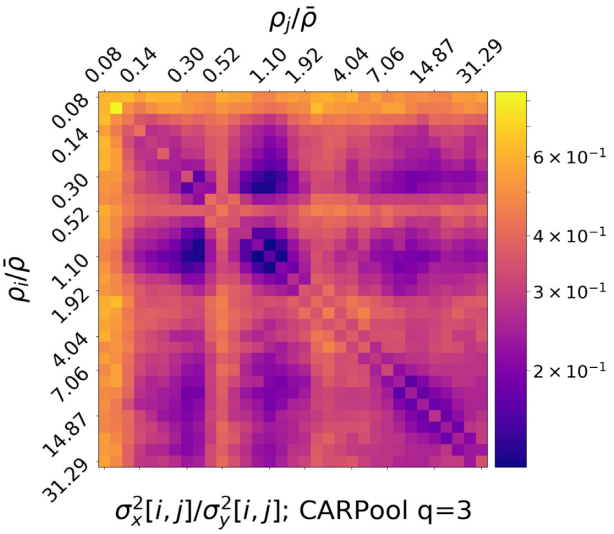

Figure 20. Variance reduction for the one dimensional matter PDF, downsampled by a factor 2. The computation is similar to that of Fig. 4.

estimate, by construction, even if the surrogate is highly biased. No modifications to the simulation codes are required. In this paper, we pair GADGET and COLA to CARPool the elements of the covariance matrix of statistics derived from GADGET N-body simulations. As in CWAV20, we do not perform any pre-processing of the statistics to improve correlations to study the raw performance of CARPool. For particular applications, there are likely physically motivated approaches that would improve the variance reduction, as discussed, *e.g.*, in Section 3.3.6.

Our presentation focuses on the relative reduction of the Monte Carlo variance on the covariance matrix and the resulting reduction in computational cost. An alternative but equivalent perspective is to view CARPool as a technique to de-biases sets of fast surrogates using a limited number of full simulations. Rather than merely accelerating convergence, CARPool can be an enabling technology to help reach accuracy requirements that would have otherwise been out of reach with a fixed amount of computation, since that is likely to be the limited resource.

We assess the impact of the CARPool variance reduction using derived properties of the covariance matrix, such as the inverse covariance (or precision) matrix, the eigenvalues, a log-likelihood statistic, and the (inverse) Fisher matrix. [8]

For the power spectrum and both equilateral and squeezed triangle configurations of the bispectrum, the variance of the covariance matrix elements is reduced by more than order of magnitude and up to four orders of magnitude. These improvements translate to significantly more accurate log-likelihoods computed from test data, eigenvalues, and estimated confidence regions for cosmological parameters when compared to computations using a reference covariance matrix based on 12 000 N-body realizations.

CARPool also gave significant reductions of variance for the covariance matrices of the matter correlation function and the matter one-point PDF, two real-space clustering statistics. In this instance, we found that these improvements did not translate to significant improvements in the derived (test) quantities. For example, in these cases, the eigenvalue ratios and Fisher matrices computed using the CARPool covariance show only marginal improvement compared to those using the sample covariance matrix based on the same number of simulations. These examples have severely ill-conditioned covariance matrices, sometimes causing the smallest eigenvalues of the CARPool estimator to scatter negative. This can be understood because our matrix generalization of CARPool reduces variance on the elements of the covariance matrix but does not explicitly guarantee positive (semi-)definiteness of the covariance matrix. A possible way to address this issue would be to apply techniques designed to improve the condition of the covariance matrix such as tapering (Paz & Sánchez 2015) and shrinkage estimators (Schäfer & Strimmer 2005; Pope & Szapudi 2008; Joachimi 2017), albeit at the cost of giving up the strict unbiasedness of the covariance matrix estimates. Alternatively, we could seek to impose positive definiteness in our matrix generalization of the CARPool method; this will be subject of future work.

More generally, combining CARPool with other fast estimators of the covariance matrix listed in Section 1 will likely lead to further improvements since CARPool relies on the entirely different principle of combining simulations and surrogates. Similarly, we explored only a single surrogate. Combinations of surrogates or better surrogates, such as those mentioned in Section 1 could well lead to further acceleration. We leave an exploration of such combined estimators to future work.

ACKNOWLEDGEMENTS

NC acknowledges funding from LabEx ENS-ICFP (PSL). BDW acknowledges support by the ANR BIG4 project, grant ANR-16-CE23-0002 of the French Agence Nationale de la Recherche; and the 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. The Flatiron Institute is supported by the Simons Foundation.

[8]The guaranteed lack of bias does not automatically also apply to these transformed quantities. An alternative approach would have been directly to CARPool the derived quantities since the unbiasedness guarantee would then apply. Since the CARPool approach is very flexible, we leave the choice of what quantity to apply it to up to the user.

DATA AVAILABILITY

The data underlying this article are available through *globus.org*, and instructions can be found at https://github.com/franciscovillaescusa/Quijote-simulations. Additionally, a PYTHON3 package with code examples and documentation is provided at https://github.com/CompiledAtBirth/pyCARPool to experiment with CARPool.

REFERENCES

Alsing J., Wandelt B., 2018, MNRAS, 476, L60
Angulo R. E., Pontzen A., 2016, MNRAS, 462, L1
Angulo R. E., Zennaro M., Contreras S., Aricò G., Pellejero-Ibañez M., Stücker J., 2020, MNRAS, 507, 5869
Bai Z. D., Yin Y. Q., 1993, The Anna. Prob., 21, 1275
Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, Phys. Rept., 367, 1
Blot L. et al., 2019, MNRAS, 485, 2806
Blot L., Corasaniti P. S., Alimi J.-M., Reverdy V., Rasera Y., 2014, MNRAS, 446, 1756
Blot L., Corasaniti P. S., Amendola L., Kitching T. D., 2016, MNRAS, 458, 4462
Carron J., 2013, A&A, 551, A88
Chartier N., Wandelt B., Akrami Y., Villaescusa-Navarro F., 2020, MNRAS, 503, 1897
Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621
Colavincenzo M. et al., 2019, MNRAS, 482, 4883
Dai B., Seljak U., 2020, Proc. Nat. Acad. Sci., 118, 1
DeRose J. et al., 2019, ApJ, 875, 69
Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rept., 733, 1
Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537
Eifler T., Schneider P., Hartlap J., 2009, A&A, 502, 721
Escoffier S. et al., 2016, preprint (arXiv:1606.00233)
Favole G., Granett B. R., Silva Lafaurie J., Sapone D., 2020, MNRAS, 505, 5833
Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, MNRAS, 463, 2273
Friedrich O. et al., 2021, MNRAS, 508, 3125
Friedrich O., Eifler T., 2018, MNRAS, 473, 4150
Garrison L., 2019, PhD thesis, Univ. Washington
Glynn P. W., Szechtman R., 2002, in Fang K.-T., Niederreiter H., Hickernell F. J., eds, Monte Carlo and Quasi-Monte Carlo Methods 2000, Springer-Verlag, Berlin, p. 27
Habib S. et al., 2016, New A, 42, 49
Hall A., Taylor A., 2019, MNRAS, 483, 189
Harnois-Déraps J., Pen U.-L., 2013, MNRAS, 431, 3349
Harnois-Déraps J., Vafaei S., Van Waerbeke L., 2012, MNRAS, 426, 1262
Harnois-Déraps J., Giblin B., Joachimi B., 2019, A&A, 631, A160
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
He S., Li Y., Feng Y., Ho S., Ravanbakhsh S., Chen W., Póczos B., 2019, Proc. Nat. Acad. Sci., 116, 13825
Heavens A. F., Jimenez R., Lahav O., 2000, MNRAS, 317, 965
Hikage C., Takahashi R., Koyama K., 2020, Phys. Rev. D, 102, 083514
Howlett C., Manera M., Percival W., 2015, Astron. Comput., 12, 109
Ishiyama T., Fukushige T., Makino J., 2009, PASJ, 61, 1319
Izard A., Crocce M., Fosalba P., 2016, MNRAS, 459, 2327
Joachimi B., 2017, MNRAS, 466, L83
Kasim M. F. et al., 2020, preprint (arXiv:2001.08055v2)
Kitaura F. S., Yepes G., Prada F., 2014, MNRAS, 439, L21
Kodi Ramanah D., Charnock T., Villaescusa-Navarro F., Wandelt B. D., 2020, MNRAS, 495, 4227
Kodwani D., Alonso D., Ferreira P., 2019, Open J Astrophys., 2, 3
Lavenberg S., Welch P., 1981, Manag. Sci., 27, 322
Leclercq F., Jasche J., Gil-Marín H., Wandelt B., 2013, J. Cosmol. Astropart. Phys., 2013, 048
Leclercq F., Faure B., Lavaux G., Wandelt B. D., Jaffe A. H., Heavens A. F., Percival W. J., 2020, A&A, 639, A91
Li Y., Singh S., Yu B., Feng Y., Seljak U., 2019, J. Cosmol. Astropart. Phys., 2019, 016
Lippich M. et al., 2019, MNRAS, 482, 1786
McClintock T. et al., 2019a, preprint (arXiv:1907.13167)
McClintock T. et al., 2019b, ApJ, 872, 53
Modi C., Lanusse F., Seljak U., 2021, Astronomy and Computing, 37, 100505
Mohammed I., Seljak U., 2014, MNRAS, 445, 3382
Mohammed I., Seljak U., Vlah Z., 2017, MNRAS, 466, 780
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Paz D. J., Sánchez A. G., 2015, MNRAS, 454, 4326
Pearson D. W., Samushia L., 2016, MNRAS, 457, 993
Percival W. J. et al., 2014, MNRAS, 439, 2531
Percival W. J., Friedrich O., Sellentin E., Heavens A., 2021, MNRAS, preprint(arXiv:2108.10402)
Philcox O. H. E., Eisenstein D. J., 2019, MNRAS, 490, 5931
Philcox O. H. E., Eisenstein D. J., O'Connell R., Wiegand A., 2020, MNRAS, 491, 3290
Philcox O. H. E., Ivanov M. M., Zaldarriaga M., Simonović M., Schmittfull M., 2021, Phys. Rev. D, 103, 043508
Planck Collaboration et al., 2020, A&A, 641, A6
Pontzen A., Slosar A., Roth N., Peiris H. V., 2016, Phys. Rev. D, 93, 103519
Pope A. C., Szapudi I., 2008, MNRAS, 389, 766
Potter D., Stadel J., Teyssier R., 2017, Comput. Astrophys. Cosmol., 4, 2
Rubinstein R. Y., Marcus R., 1985, Oper. Res., 33, 661
Schäfer J., Strimmer K., 2005, Stat. App. Gen. Molec. Biol., 4, Article32
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Sellentin E., Heavens A. F., 2018, MNRAS, 473, 2355
Springel V., 2005, MNRAS, 364, 1105
Taffoni G., Monaco P., Theuns T., 2002, MNRAS, 333, 623
Takahashi R. et al., 2009, ApJ, 700, 479
Tassev S., Zaldarriaga M., 2012, J. Cosmol. Astropart. Phys., 2012, 013
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, J. Cosmol. Astropart. Phys., 2013, 036–036
Tassev S., Eisenstein D. J., Wand elt B. D., Zaldarriaga M., 2015, preprint (arXiv:1502.07751)
Taylor A., Joachimi B., 2014, MNRAS, 442, 2728
Taylor A., Joachimi B., Kitching T., 2013, MNRAS, 432, 1928
Villaescusa-Navarro F. et al., 2018, ApJ, 867, 137
Villaescusa-Navarro F. et al., 2020, ApJS, 250, 2
Wadekar D., Ivanov M. M., Scoccimarro R., 2020, Phys. Rev. D, 102, 123521
Warren M. S., 2013, Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis Association for Computing Machinery
White M., Tinker J. L., McBride C. K., 2014, MNRAS, 437, 2594
Zhai Z. et al., 2019, ApJ, 874, 95

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