CARPool: fast, accurate computation of large-scale structure statistics by pairing costly and cheap cosmological simulations

Nicolas Chartier, Benjamin Wandelt, Yashar Akrami, Francisco Villaescusa-Navarro

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **503,** 1897–1914 (2021)
Advance Access publication 2021 February 17

doi:10.1093/mnras/stab430

CARPool: fast, accurate computation of large-scale structure statistics by pairing costly and cheap cosmological simulations

Nicolas Chartier,[1,2]★ Benjamin Wandelt,[2,3] Yashar Akrami [1,4] and Francisco Villaescusa-Navarro[3,5]

[1]*Laboratoire de Physique de l'École Normale Supérieure, ENS, Universite PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France*
[2]*Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, 75014 Paris, France*
[3]*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*
[4]*Observatoire de Paris, Université PSL, Sorbonne Université, LERMA, F-75014 Paris, France*
[5]*Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey, NJ 08544, USA*

Accepted 2021 February 9. Received 2021 February 9; in original form 2020 October 29

ABSTRACT

To exploit the power of next-generation large-scale structure surveys, ensembles of numerical simulations are necessary to give accurate theoretical predictions of the statistics of observables. High-fidelity simulations come at a towering computational cost. Therefore, approximate but fast simulations, *surrogates*, are widely used to gain speed at the price of introducing model error. We propose a general method that exploits the correlation between simulations and surrogates to compute fast, reduced-variance statistics of large-scale structure observables *without model error* at the cost of only a few simulations. We call this approach Convergence Acceleration by Regression and Pooling (CARPool). In numerical experiments with intentionally minimal tuning, we apply CARPool to a handful of GADGET-III *N*-body simulations paired with surrogates computed using COmoving Lagrangian Acceleration. We find ∼100-fold variance reduction even in the non-linear regime, up to $k_{\rm max} \approx 1.2\,h{\rm Mpc}^{-1}$ for the matter power spectrum. CARPool realizes similar improvements for the matter bispectrum. In the nearly linear regime CARPool attains far larger sample variance reductions. By comparing to the 15 000 simulations from the *Quijote* suite, we verify that the CARPool estimates are unbiased, as guaranteed by construction, even though the surrogate misses the simulation truth by up to 60 per cent at high k. Furthermore, even with a fully configuration-space statistic like the non-linear matter density probability density function, CARPool achieves unbiased variance reduction factors of up to ∼10, without any further tuning. Conversely, CARPool can be used to remove model error from ensembles of fast surrogates by combining them with a few high-accuracy simulations.

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

1 INTRODUCTION

The next generation of galaxy surveys will provide a detailed chart of cosmic structure and its growth on our cosmic light cone. These include the *Euclid* space telescope (Laureijs et al. 2011; Euclid Collaboration 2020), the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration 2016a, b), the Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019; LSST Science Collaboration 2009; LSST Dark Energy Science Collaboration 2018), the Square Kilometre Array (SKA; Yahya et al. 2015; Square Kilometre Array Cosmology Science Working Group 2020), the Wide Field InfraRed Survey Telescope (WFIRST; Spergel et al. 2015), the Subaru Hyper Suprime-Cam (HSC) and Prime Focus Spectrograph (PFS) surveys (Aihara et al. 2018; Tamura et al. 2016), and the Spectro-Photometer for the History of the Universe, Epoch of Reionization, and Ices Explorer (SPHEREx; Doré et al. 2014, 2018). These data sets will provide unprecedented statistical power to constrain the initial perturbations, the growth of cosmic structure, and the cosmic expansion history. To access this information requires accurate theoretical models of large-scale structure statistics, such as power spectra and bispectra. While analytical work, such as standard perturbation theory (Jain & Bertschinger 1994; Goroff et al. 1986), Lagrangian perturbation theory (LPT; Bouchet et al. 1995; Matsubara 2008), renormalized perturbation theory (Crocce & Scoccimarro 2006), and effective field theory (Carrasco, Hertzberg & Senatore 2012; Vlah, White & Aviles 2015; Perko et al. 2016), has made great strides [see also Bernardeau et al. (2002), Desjacques, Jeong & Schmidt (2018) for reviews], the reference models for large-scale structure are based on computationally intensive *N*-body simulations that compute the complex non-linear regime of structure growth. In recent years, the BACCO simulation project (Angulo et al. 2020), the Outer Rim Simulation (Heitmann et al. 2019), the Aemulus project I (DeRose et al. 2019), the ABACUS Cosmos suite (Garrison et al. 2018), the Dark Sky Simulations (Skillman et al. 2014), the MICE Grand Challenge (MICE-GC; Crocce et al. 2015), the Coyote Universe I (Heitmann et al. 2010), and the Uchuu simulations (Ishiyama et al. 2020), among others, involved generation of expensive *N*-body simulations.

While analytical methods compute expectation values of large-scale structure statistics, a simulation generates a single realization and its output therefore suffers from sample variance. Reducing this variance to a point where it is subdominant to the observational error therefore requires running ensembles of simulations.

★ E-mail: nicolas.chartier@phys.ens.fr

Computational cosmologists have been tackling the challenge of optimizing *N*-body codes and gravity solvers for a growingly larger number of particles. Widely used codes include the parallel Tree Particle-Mesh (TreePM or TPM) codes GADGET-II by Springel (2005) and GREEM by Ishiyama, Fukushige & Makino (2009), the adaptive treecode 2HOT by Warren (2013), the GPU-accelerated ABACUS code originated from Garrison (2019), the Hardware/Hybrid Accelerated Cosmology Code (HACC) developed by Habib et al. (2016), and the distributed-memory and GPU-accelerated PKDGRAV3, based on Fast Multipole Methods and adaptive particle timesteps, from Potter, Stadel & Teyssier (2017). The memory and CPU time requirements of such computations are a bottleneck for future work on new-generation cosmological data sets. As an example, the 43 100 runs in the *Quijote* simulations from Villaescusa-Navarro et al. (2020), of which the data outputs are public and used in this paper, required 35 million CPU-core-hours.

The search for solutions has led to alternative, fast, and approximate ways to generate predictions for large-scale structure statistics. The COmoving Lagrangian Acceleration (COLA) solver of Tassev, Zaldarriaga & Eisenstein (2013) is a PM code that solves the particle equations of motion in an accelerated frame given by LPT. Particles are nearly at rest in this frame for much of the mildly non-linear regime. As a consequence, much larger timesteps can be taken, leading to significant time savings. The *N*-body solver $\texttt{FastPM}$ of Feng et al. (2016) operates on a similar principle, using modified kick and drift factors to enforce the Zel'dovich approximation in the mildly non-linear regime. The spatial COLA (sCOLA) scheme (Tassev et al. 2015) extends the idea of using LPT to guide the solution in the spatial domain. Leclercq et al. (2020) have carefully examined and implemented these ideas to allow splitting large *N*-body simulations into many perfectly parallel, independently evolving small simulations.

In a different family of approaches, but still using LPT, Monaco et al. (2013) proposed a parallelized implementation of the PINpointing Orbit Crossing-Collapsed HI-erarchical Objects (PINOCCHIO) algorithm from Taffoni, Monaco & Theuns (2002). Chuang et al. (2015) developed a physically motivated enhancement of the Zel'dovich approximation called EZmocks. Approximate methods and full *N*-body simulations can also be jointly used. For instance, Tassev & Zaldarriaga (2012) proposed a statistical linear regression model of the non-linear matter density field using the density field given by perturbation theory, for which the random residual error is minimized.

Recently, so-called *emulators* have been of great interest: they predict statistics in the non-linear regime based on a generic mathematical model whose parameters are trained on simulation suites covering a range of cosmological parameters. An emulator is trained by Angulo et al. (2020) on the BACCO simulations; similarly, the Aemulus project contributions II, III, and IV (McClintock et al. 2019a, b; Zhai et al. 2019), respectively, construct an emulator for the halo mass function, the galaxy correlation function, and the halo bias using the Aemulus I suite (DeRose et al. 2019). Not only do emulators that map cosmological parameters to certain outputs need large numbers of simulations for training, they also do not guarantee unbiased results with respect to full simulation codes, especially outside the parameter range used during training.

Recent advances in deep learning have allowed training emulators that reproduce particle positions or density fields starting from initial conditions therefore essentially emulating the full effect of a low-resolution cosmological *N*-body code – these include the Deep Density Displacement Model (D^3M) of He et al. (2019) stemming from the U-NET architecture (Ronneberger, Fischer & Brox 2015). Kodi Ramanah et al. (2020) describe a complementary deep learning tool that increases the mass and spatial resolution of low-resolution *N*-body simulations using a variant of Generative Adversarial Networks (Goodfellow et al. 2014).

None of these fast approximate solutions exactly reproduce the results of more computationally intensive codes. They trade computational accuracy for computational speed, especially in the non-linear regime. In this vein, the recent series of papers by Lippich et al. (2019), Blot et al. (2019), and Colavincenzo et al. (2019) compare the covariance matrices of clustering statistics given by several low-fidelity methods to those of full *N*-body codes and find statistical biases in the parameter uncertainties by up to 20 per cent.

A different approach to this problem is to reduce the stochasticity of the initial conditions, thereby modifying the statistics of the observables in such a way as to reduce sample variance. This is the spirit of the method of fixed fields invented and first explored by Pontzen et al. (2016) and Angulo & Pontzen (2016). They found in numerical experiments that a large variety of statistics retain the correct mean, and analytically showed that pairing and fixing, while changing the initial distributions, only impact a measure-zero set of correlations when the errors are not smothered by the large number of available modes. While this approach does not guarantee that any given statistic will be unbiased, the numerical study by Villaescusa-Navarro et al. (2018) showed that 'fixing' succeeds in reducing variance for several statistics of interest with no detectable bias when comparing to an ensemble of hundreds of full simulations and at no additional cost to regular simulations. Still, it is clear that other statistics must necessarily be biased, for example, the square of any variance-reduced statistic, such as four-point functions. Still in the family of variance reduction methods, Smith & Angulo (2019) built a composite model of the matter power spectrum and managed to cancel most of the cosmic variance on large scales, notably by using the ratio of matched phase initial conditions.

In this paper, we show that it is possible to get the best of both worlds: the speed of fast surrogates *and* the guarantee of full-simulation accuracy.[1] We take inspiration from *control variates*, a classical variance reduction technique that directly and optimally minimizes the variance of any random quantity [see Lavenberg & Welch (1981) for a review, and Gorodetsky et al. (2020) and Peherstorfer, Willcox & Gunzburger (2016) for related recent applications], to devise a way to combine fast but approximate simulations (or *surrogates*) with computationally intensive accurate simulations to vastly accelerate convergence while *guaranteeing* arbitrarily small bias with respect to the full simulation code. We call this Convergence Acceleration by Regression and Pooling (CARPool).[2]

The paper is organized as follows. In Section 2, we explore the theory of univariate and multivariate estimation with control variates and highlight some differences in our setting for cosmological simulations. In Section 3, we briefly discuss both the *N*-body simulation suite and our choice of fast surrogates we use in the numerical experiments presented in Section 4. We conclude in Section 5.

Table 1 lists mathematical notation and definitions used throughout this paper.

[1]As a jargon reminder, the accuracy and precision of an estimate refer, respectively, to the trueness of its expectation (in terms of the statistical bias) and the confidence in the expectation (standard errors, confidence intervals).
[2]We will consider surrogates to be much faster than simulations, so that we only need to consider the number of simulations to evaluate computational cost.

Table 1. Mathematical notation and definitions.

Notation	Description
$\mathcal{S}_N = \{r_1, \ldots, r_N\}$	Set of N random seeds r_n of probability space
$\boldsymbol{y}(r_n) \equiv \boldsymbol{y}_n$	Random column vector of size p at seed r_n
$\mathbb{E}[\boldsymbol{y}] \equiv \boldsymbol{\mu}_{\boldsymbol{y}}$	Expectation value of random vector $\boldsymbol{y}$
$[\![m, n]\!]$	Set of integers from m to n
$\boldsymbol{M}^{\boldsymbol{T}}$	Transpose of real matrix $\boldsymbol{M}$
$\boldsymbol{M}^{\dagger}$	Moore–Penrose pseudo-inverse of matrix $\boldsymbol{M}$
$\det(\boldsymbol{M})$	Determinant of matrix $\boldsymbol{M}$
$\mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\boldsymbol{T}}\right] \equiv \boldsymbol{\Sigma}_{\boldsymbol{xx}}$	Variance-covariance matrix of random vector $\boldsymbol{x}$
$\mathbb{E}\left[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\boldsymbol{T}}\right] \equiv \boldsymbol{\Sigma}_{\boldsymbol{yx}}$	Cross-covariance matrix of random vectors $\boldsymbol{y}$ and $\boldsymbol{x}$
σ_y^2	Variance of scalar random variable y
$\boldsymbol{0}_{p,q}$ and $\boldsymbol{0}_p$	Null matrix in $\mathbb{R}^{p\times q}$ and null vector in $\mathbb{R}^p$
$\boldsymbol{I}_p$	Square $p \times p$ identity matrix

2 METHODS

Let us consider a set of observables y_i we would like to model (e.g. power spectrum or bispectrum bins) and collect them into a random vector $\boldsymbol{y}$ with values in $\mathbb{R}^p$. The standard estimate of the theoretical expectation of $\boldsymbol{y}$, $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\mu}$, from a set of *independent and identically distributed* realizations $\boldsymbol{y}_n$, $n = 1, \ldots N$, is the *sample mean*

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

Then the standard deviation σ_i of each element $\bar{y}_i$ decreases as $\mathcal{O}(N^{-\frac{1}{2}})$, under mild regularity conditions (principally that σ_i exists).

Our goal is to find a more precise – i.e. lower variance – and unbiased estimator of $\mathbb{E}[\boldsymbol{y}]$ with a much smaller number of simulations $\boldsymbol{y}_n$. The means by which we achieve this is to construct another set of quantities that are fast to compute such that (i) their means are small enough to be negligible, and (ii) their errors are anticorrelated with the errors in the $\boldsymbol{y}_n$,[3] and add some multiple of these to $\bar{\boldsymbol{y}}$ to cancel some of the error in the y_n. This is the *control variates* principle.

2.1 Theoretical framework

In what follows we will use the word *simulation* to refer to costly high-fidelity runs and *surrogate* for fast but low-fidelity runs.

2.1.1 Introduction with the scalar case

Let us consider a scalar simulated observable y, such that $\mathbb{E}[y] = \mu$, and a surrogate c of y with $\mathbb{E}[c] = \mu_c$. Note that $\mu \neq \mu_c$ in general. For any $\beta \in \mathbb{R}$, the quantity

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

is an unbiased estimator of μ by construction. The optimal value for β is determined by minimizing the variance of the new estimator,

$$\sigma^2_{x(\beta)} = \beta^2\sigma_c^2 - 2\beta\mathrm{cov}(y, c) + \sigma_y^2. \tag{3}$$

The function (3) of β has a strict global minimum point at

$$\beta^{\star} = \underset{\beta\in\mathbb{R}}{\mathrm{argmin}}\,\sigma^2_{x(\beta)} = \frac{\mathrm{cov}(y, c)}{\sigma_c^2}. \tag{4}$$

[3]The intuition behind this principle is that for two random scalars a and b, we have $\sigma^2_{a+b} = \sigma_a^2 + \sigma_b^2 + 2\mathrm{cov}(a, b)$.

Plugging equation (4) into equation (3) allows us to express the variance reduction ratio of control variates as

$$\frac{\sigma^2_{x(\beta)}}{\sigma_y^2} = 1 - \rho^2_{y,c}\,, \tag{5}$$

with $\rho_{y,c}$ the Pearson correlation coefficient between y and c. The latter result shows that no matter how biased the surrogate c might be, the more correlated it is with the simulation y, the better the variance reduction. For the classical control variates method, the choice of c is restricted to cases where μ_c and β are known *a priori*. In Section 2.2 below, we will consider the more general case, typically encountered in practice, where β is not known and we must estimate it from data.

2.1.2 Multivariate control variates

Let $\boldsymbol{y}$ be an unbiased and costly simulation statistic of expectation $\boldsymbol{\mu} \in \mathbb{R}^p$, and $\boldsymbol{c}$ an approximate realization with $\mathbb{E}[\boldsymbol{c}] = \boldsymbol{\mu}_{\boldsymbol{c}} \in \mathbb{R}^q$. Similarly to the scalar case, for any $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$ the control variates estimator is

$$\boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{\beta}\,(\boldsymbol{c} - \boldsymbol{\mu}_{\boldsymbol{c}})\,. \tag{6}$$

$\boldsymbol{\Sigma}_{\boldsymbol{xx}}$, the covariance matrix of the random vector $\boldsymbol{x}(\boldsymbol{\beta})$, is expressed as a function of $\boldsymbol{\beta}$,

$$\boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta}) = \boldsymbol{\beta}\boldsymbol{\Sigma}_{\boldsymbol{cc}}\boldsymbol{\beta}^{\boldsymbol{T}} - \boldsymbol{\beta}\boldsymbol{\Sigma}_{\boldsymbol{yc}}^{\boldsymbol{T}} - \boldsymbol{\Sigma}_{\boldsymbol{yc}}\boldsymbol{\beta}^{\boldsymbol{T}} + \boldsymbol{\Sigma}_{\boldsymbol{yy}}. \tag{7}$$

Optimizing variance reduction here means minimizing the confidence region associated to $\mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta})]$ and represented by the generalized variance $\det(\boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta}))$. Appendix A presents a Bayesian solution to the Gaussian version of this optimization problem.

Here we present an outline of the derivation in de O. Porta Nova & Wilson (1993) and Venkatraman & Wilson (1986). The course by Helwig (2017) provides an overview of canonical correlation analysis that is used in the derivation. The oriented volume of the p-dimensional parallelepiped spanned by the columns of $\boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta})$ is minimized as the analogue of an error bar in the univariate case. Rubinstein & Marcus (1985) proved that

$$\boldsymbol{\beta}^{\star} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p\times q}}{\mathrm{argmin}}\det(\boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta})) = \boldsymbol{\Sigma}_{\boldsymbol{yc}}\boldsymbol{\Sigma}_{\boldsymbol{cc}}^{-\boldsymbol{1}}. \tag{8}$$

Combining equations (8) and (7) gives the generalized variance reduction

$$\begin{aligned}\frac{\det(\boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta}^{\star}))}{\det(\boldsymbol{\Sigma}_{\boldsymbol{yy}})} &= \frac{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{yy}}\left(\boldsymbol{I}_p - \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-\boldsymbol{1}}\boldsymbol{\Sigma}_{\boldsymbol{yc}}\boldsymbol{\Sigma}_{\boldsymbol{cc}}^{-\boldsymbol{1}}\boldsymbol{\Sigma}_{\boldsymbol{yc}}^{\boldsymbol{T}}\right)\right)}{\det(\boldsymbol{\Sigma}_{\boldsymbol{yy}})} \\ &= \prod_{n=1}^{s=\mathrm{rank}(\boldsymbol{\Sigma}_{\boldsymbol{yc}})}\left(1 - \lambda_n^2\right),\end{aligned} \tag{9}$$

where the scalars $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_s^2 \geq 0$ are the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-\boldsymbol{1}}\boldsymbol{\Sigma}_{\boldsymbol{yc}}\boldsymbol{\Sigma}_{\boldsymbol{cc}}^{-\boldsymbol{1}}\boldsymbol{\Sigma}_{\boldsymbol{yc}}^{\boldsymbol{T}}$ and whose square roots are the canonical correlations between $\boldsymbol{y}$ and $\boldsymbol{c}$. More precisely, λ_1 is the maximum obtainable cross-correlation between any linear combinations $\boldsymbol{u}_1^{\boldsymbol{T}}\boldsymbol{y}$ and $\boldsymbol{v}_1^{\boldsymbol{T}}\boldsymbol{c}$,

$$\lambda_1 = \underset{\boldsymbol{u_1}\in\mathbb{R}^p,\boldsymbol{v_1}\in\mathbb{R}^q}{\mathrm{argmax}}\frac{\boldsymbol{u}_1^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{yc}}\boldsymbol{v}_1}{\sqrt{\boldsymbol{u}_1^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{u}_1}\sqrt{\boldsymbol{v}_1^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{cc}}\boldsymbol{v}_1}}, \tag{10}$$

and $\{\lambda_n; n \leq s\}$ are found recursively with the constraint of uncorrelatedness between $\left\{\boldsymbol{u}_{\boldsymbol{n}}^{\boldsymbol{T}}\boldsymbol{y}, \boldsymbol{v}_{\boldsymbol{n}}^{\boldsymbol{T}}\boldsymbol{c}\right\}$ and $\left\{\boldsymbol{u}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{y}, \boldsymbol{v}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{c}, \ldots, \boldsymbol{u}_{\boldsymbol{n-1}}^{\boldsymbol{T}}\boldsymbol{y}, \boldsymbol{v}_{\boldsymbol{n-1}}^{\boldsymbol{T}}\boldsymbol{c}\right\}$. At the end, we have two bases for the transformed vectors $\boldsymbol{u} = \left[\boldsymbol{u}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{y}, \ldots, \boldsymbol{u}_{\boldsymbol{s}}^{\boldsymbol{T}}\boldsymbol{y}\right]^{\boldsymbol{T}}$ and $\boldsymbol{v} = \left[\boldsymbol{v}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{c}, \ldots, \boldsymbol{v}_{\boldsymbol{s}}^{\boldsymbol{T}}\boldsymbol{c}\right]^{\boldsymbol{T}}$ in which their cross-covariance matrix is diagonal i.e. $\boldsymbol{\Sigma}_{\boldsymbol{uv}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_s)$.

2.2 Estimation in practice

In this section, we examine practical implications of the control variates implementation when the optimal control matrix $\boldsymbol{\beta}$ (or coefficients) and the mean of the cheap estimator $\boldsymbol{\mu_c}$ are unknown. We will consider an online approach in order to improve the estimates of (4) or (8) as simulations and surrogates are computed. Estimating $\boldsymbol{\mu_c}$ is done through an inexpensive pre-computation step that consists in running fast surrogates. From now on, to differentiate our use of the control variates principle and its application to cosmological simulations from the theory presented above, we will refer to it as the CARPool technique.

For the purposes of this paper, we will take as our goal to produce low-variance estimates of expectation values of full simulation observables. When we discuss model error, it is therefore only relative to the full simulation. From an absolute point of view the accuracy of the full simulation depends on a number of factors such as particle number, force resolution, timestepping, inclusion of physical effects, etc. The numerical examples of full simulations we give are not selected for their unmatched accuracy, but for the availability of a large ensemble that we can use to validate the CARPool results.

2.2.1 Estimation of $\boldsymbol{\mu_c}$

In the textbook control variates setting, the crude approximation $\boldsymbol{\mu_c}$ of $\boldsymbol{\mu}$ is assumed to be known. There is no reason for this to be the case in the context of cosmological simulations, thus we compute $\bar{\boldsymbol{\mu}}_{\boldsymbol{c}}$ with surrogate samples drawn on a separate set of seeds $\mathcal{S}_M = \{r_1, \ldots, r_M\}$ ($\mathcal{S}_N \cap \mathcal{S}_M = \emptyset$, where $\mathcal{S}_N$ is the set of initial conditions of simulations). What is then the additional variance-covariance of the control variates estimate stemming from the estimation of $\boldsymbol{\mu_c}$?

First, write each cheap-estimator realization as $\boldsymbol{c} = \boldsymbol{\mu_c} + \boldsymbol{\delta}$, with $\mathbb{E}[\boldsymbol{\delta}] = \mathbf{0}_q$,

$$\begin{aligned} \bar{\boldsymbol{\mu}}_{\boldsymbol{c}} &= \boldsymbol{\mu_c} + \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\delta}_i, \\ \boldsymbol{\Sigma}_{\bar{\boldsymbol{\mu}}_{\boldsymbol{c}}\bar{\boldsymbol{\mu}}_{\boldsymbol{c}}} &= \boldsymbol{\Sigma}_{\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}} = \frac{1}{M}\boldsymbol{\Sigma}_{\boldsymbol{cc}}. \end{aligned} \tag{11}$$

Replacing $\boldsymbol{\mu_c}$ by $\bar{\boldsymbol{\mu}}_{\boldsymbol{c}}$ in equation (6) and computing the covariance results in

$$\begin{aligned} \boldsymbol{x}(\boldsymbol{\beta}, \bar{\boldsymbol{\mu}}_{\boldsymbol{c}}) &= \boldsymbol{y} - \boldsymbol{\beta}(\boldsymbol{c} - \boldsymbol{\mu_c}) + \boldsymbol{\beta}\bar{\boldsymbol{\delta}}, \\ \boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta}, \bar{\boldsymbol{\mu}}_{\boldsymbol{c}}) &= \boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta}) + \boldsymbol{\beta}\frac{\boldsymbol{\Sigma}_{\boldsymbol{cc}}}{M}\boldsymbol{\beta}^{\boldsymbol{T}}, \end{aligned} \tag{12}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{xx}}(\boldsymbol{\beta})$ from equation (7). The $\boldsymbol{\beta}\bar{\boldsymbol{\delta}}$ term above is statistically independent of the rest of the sum, since it is computed on a separate set of seeds. As expected, additional uncertainty is brought by $\boldsymbol{\Sigma}_{\boldsymbol{cc}}$ and scaled by the estimated control matrix. See Appendix A for a Bayesian derivation of the combined uncertainty in the Gaussian case while taking into account possible prior information on $\boldsymbol{\mu}$ and/or $\boldsymbol{\mu_c}$.

2.2.2 Estimation of the control matrix

The matrices in equation (8) need to be estimated from data via the bias-corrected sample covariance matrix

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{yc}} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \bar{\boldsymbol{y}}\right)(\boldsymbol{c}_i - \bar{\boldsymbol{c}})^{\boldsymbol{T}}, \\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}} &= \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{c}_i - \bar{\boldsymbol{c}})(\boldsymbol{c}_i - \bar{\boldsymbol{c}})^{\boldsymbol{T}}. \end{aligned} \tag{13}$$

The computational cost of $\boldsymbol{y}$ is the limiting factor for estimating $\boldsymbol{\Sigma}_{\boldsymbol{yc}}$. Therefore, the cross-covariance matrix is estimated online, as our primary motivation is to reduce the computation time: for instance, we certainly do not want to run more costly simulations in a precomputation step like we do for $\boldsymbol{\mu_c}$ with fast simulations. Simply put, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{yc}}$ is updated each time a new simulation pair is available.

Note that for finite N, the inverse of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}}$ in equation (13) is not an unbiased estimator of the precision matrix $\boldsymbol{\Sigma}_{\boldsymbol{cc}}^{-1}$ (Hartlap, Simon & Schneider 2006). Moreover, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}}^{-1}$ is not defined when $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}}$ is rank-deficient, which is guaranteed to happen when N is smaller that p . We have consequently replaced $\boldsymbol{\Sigma}_{\boldsymbol{cc}}^{-1}$ by the Moore–Penrose pseudo-inverse – always defined and unique – $\boldsymbol{\Sigma}_{\boldsymbol{cc}}^{\dagger}$ in equation (8) for the numerical analysis presented in Section 4 to be able to compute multivariate CARPool estimates even when $N < p$.

Since the singular value decomposition exists for any complex or real matrix, we can write $\boldsymbol{\Sigma}_{\boldsymbol{yc}} = \boldsymbol{U}\boldsymbol{V}\boldsymbol{W}^{\boldsymbol{T}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{cc}} = \boldsymbol{O}\boldsymbol{P}\boldsymbol{Q}^{\boldsymbol{T}} = \boldsymbol{O}\boldsymbol{P}\boldsymbol{O}^{\boldsymbol{T}}$ by symmetry. The optimal control matrix now gives $\boldsymbol{\beta}^{\star} = \boldsymbol{U}\boldsymbol{V}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{O}\boldsymbol{P}^{-1}\boldsymbol{O}^{\boldsymbol{T}}$. The product $-\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{O}^{\boldsymbol{T}}$ whitens the centered surrogate vector elements (principal component analysis whitening), $\boldsymbol{O}\boldsymbol{P}^{-\frac{1}{2}}$ restretches the coefficients and returns them to the surrogate basis, and then $\boldsymbol{U}\boldsymbol{V}\boldsymbol{W}^{\boldsymbol{T}}$ projects the scaled surrogate elements into the high-fidelity simulation basis and rescales them to match the costly simulation covariance. It follows that, when using $\hat{\boldsymbol{\beta}}$ in practice, the projections are done in bases specifically adapted to the $\boldsymbol{y}$ and $\boldsymbol{c}$ samples available. With this argument, we justify why we use the same simulation/surrogate pairs to compute $\hat{\boldsymbol{\beta}}$ first (with the Moore–Penrose pseudo-inverse of the surrogate covariance replacing the precision matrix) and estimate the CARPool mean after that.

An online estimation of both $\hat{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{x}}(\hat{\boldsymbol{\beta}})$, considering incoming $\{\boldsymbol{y}_n, \boldsymbol{c}_n\}$ pairs computed on the same seed r_n, amounts to computing a collection of N samples as functions of $\widehat{\boldsymbol{\beta}}$,

$$\boldsymbol{x}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{y}_n - \hat{\boldsymbol{\beta}}(\boldsymbol{c}_n - \bar{\boldsymbol{\mu}}_{\boldsymbol{c}}). \tag{14}$$

We implement equation (6) by taking the sample mean of N such variance-reduced samples,

$$\bar{\boldsymbol{x}}(\hat{\boldsymbol{\beta}}) = \bar{\boldsymbol{y}} - \hat{\boldsymbol{\beta}}(\bar{\boldsymbol{c}} - \bar{\boldsymbol{\mu}}_{\boldsymbol{c}}). \tag{15}$$

This way, equation (15) can be computed each time a simulation/surrogate pair is drawn from a seed in $\mathcal{S}_N = \{r_1, \ldots, r_N\}$, after updating $\hat{\boldsymbol{\beta}}$ according to equation (13).

2.2.3 Multivariate versus univariate CARPool

So far we have not assumed any special structure for $\boldsymbol{\beta}$. If, as in the classical control variates setting, the (potentially dense) covariances on the right-hand side of equation (8) are known *a priori*, then $\boldsymbol{\beta}^{\star}$ is the best solution because it exploits the mutual information between all elements of $\boldsymbol{y}$ and $\boldsymbol{c}$.

In practice, we will be using the online approach discussed in Section 2.2.2 for a very small number of simulations. If we are limited by a very small number of $\{\boldsymbol{y}_n, \boldsymbol{c}_n\}$ pairs compared to the number of elements of the vectors, the estimate of $\boldsymbol{\beta}^{\star}$ can be unstable and possibly worsen the variance of equation (15), though unbiasedness remains guaranteed.

We will demonstrate below that in the case of small number of simulations and a large number of statistics to estimate from the simulations, it is advantageous to impose structure on $\boldsymbol{\beta}$. In the simplest case, we can set the off-diagonal elements to zero. This amounts to treating each vector element separately and results in a decoupled problem with a separate solution (4) for each vector element.

The univariate setting of 2.1.1 applied individually to each vector element (*bin*) will be referred to as 'diagonal $\boldsymbol{\beta}$' or $\boldsymbol{\beta}^{\mathbf{diag}}$, as it amounts to fixing the non-diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{cc}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{yc}}$ to zero in equation (8) and only estimating the diagonal elements

$$\boldsymbol{\beta}^{\mathbf{diag}} = \begin{pmatrix} \frac{cov(y_1,c_1)}{\sigma^2_{c_1}} & & & \\ & \frac{cov(y_2,c_2)}{\sigma^2_{c_2}} & & 0 \\ & & \ddots & \\ 0 & & & \frac{cov(y_p,c_p)}{\sigma^2_{c_p}} \end{pmatrix} \tag{16}$$

The intent of this paper is to show the potential of control variates for cosmological simulations; to this end, we will compare the following unbiased estimators:

(i) GADGET, where we compute the sample mean $\bar{\boldsymbol{y}}$ from N-body simulations only.

(ii) Multivariate CARPool described by equation (6), where we estimate the control matrix $\boldsymbol{\beta}$ online using equations (13), and denote it by $\boldsymbol{\beta}^{\star}$.

(iii) Univariate CARPool, where we use the empirical counterpart of equation (4) as the control coefficient for each element of a vector: we estimate $\boldsymbol{\beta}^{\mathbf{diag}}$.

Other, intermediate choices between fully dense and diagonal $\boldsymbol{\beta}$ are possible and may be advantageous in some circumstances. We will leave an exploration of these to future work, and simply note here that this freedom to tune $\boldsymbol{\beta}$ does not affect the mean of the CARPool estimate.

3 COSMOLOGICAL SIMULATIONS

This section describes the simulation methods that we use to compute the statistics presented in Section 4. The simulations assume a Λ cold dark matter (ΛCDM) cosmology congruent with the *Planck* constraints provided by Planck Collaboration (2020): $\Omega_{\mathrm{m}} = 0.3175$, $\Omega_{\mathrm{b}} = 0.049$, $h = 0.6711$, $n_s = 0.9624$, $\sigma_8 = 0.834$, $w = -1.0$, and $M_\nu = 0.0$ eV.

3.1 *Quijote* simulations at the fiducial cosmology

Villaescusa-Navarro et al. (2020) have publicly released data outputs from N-body cosmological simulations run with the full TreePM code GADGET-III, a development of the previous version GADGET-II by Springel (2005).[4] Available data and statistics include simulation snapshots, matter power spectra, matter bispectra and matter probability density functions. The sample mean of each statistic computed from all available realizations gives the unbiased estimator of $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\mu}$. The fiducial cosmology data set contains 15 000 realizations; their characteristics are grouped in Table 2.

As discussed in Section 2.2, the *Quijote* simulations are selected because we have access to an extensive ensemble of simulations that we can use to validate the CARPool approach. In the following we will look at wavenumbers $k = \sim 1\ h\mathrm{Mpc}^{-1}$ where the *Quijote* simulations may not be fully resolved. This is not important for the purposes of this paper; we will consider the full simulation ensemble as the gold standard that we attempt to reproduce with a much smaller number of simulations plus fast surrogates.

[4]Instructions to access the data are given at https://github.com/franciscovillaescusa/Quijote-simulations

Table 2. Characteristics of GADGET-III simulations.

Characteristic/parameter	Value
Simulation box volume	$(1000\ h^{-1}\mathrm{Mpc})^3$
Number of CDM particles	$N_{\mathrm{p}} = 512^3$
Force mesh grid size	$N_{\mathrm{m}} = 1024$
Starting redshift	$z_i = 127$
Initial conditions	Second-order Lagrangian perturbation theory (2LPT)
Redshift of data outputs	$z \in \{3.0, 2.0, 1.0, 0.5, 0.0\}$

In the next section, we present the chosen low-fidelity simulation code which provides an approximate statistic $\boldsymbol{c}$ for our numerical experiments.

3.2 Choice of approximate simulation method

Any fast solution can be used for $\boldsymbol{c}$, provided that it can be fed with the same initial conditions as of the *Quijote* simulations. To this end, the matter power spectrum from CAMB (Lewis, Challinor & Lasenby 2000) at $z = 0$ is rescaled at the initial redshift $z_i = 127$ to generate the initial conditions, as in Villaescusa-Navarro et al. (2020). In this work, we use the L-PICOLA code developed by Howlett, Manera & Percival (2015), an MPI parallel implementation of the COLA method (Tassev et al. 2013). The core idea of COLA is to add residual displacements computed with a Particle-Mesh N-body solver to the trajectory given by the first- and second-order LPT approximations. If $\boldsymbol{l}$ is the initial Lagrangian position of a particle and $\boldsymbol{x}$ is its Eulerian comoving coordinates, the evolution of the residual displacement field $\boldsymbol{\Psi}_{\mathbf{res}}$ appears by rewriting the equation of motion in a frame comoving with the LPT trajectory,

$$\partial_a^2 \boldsymbol{\Psi}_{\mathbf{res}} = -\nabla_{\boldsymbol{x}} \Phi - \partial_a^2 \boldsymbol{\Psi}_{\mathbf{LPT}}, \tag{17}$$

where a is the cosmological scale factor and

$$\boldsymbol{\Psi}_{\mathbf{res}} \equiv \boldsymbol{\Psi} - \boldsymbol{\Psi}_{\mathbf{LPT}},$$
$$\boldsymbol{x}(\boldsymbol{l}, a) \equiv \boldsymbol{l} + \boldsymbol{\Psi}(\boldsymbol{l}, a).$$

Here, we have omitted the Hubble expansion rate and constants for simplicity, $\boldsymbol{\Psi}_{\mathbf{LPT}}$ is the displacement vector associated to $\boldsymbol{x}_{\mathbf{LPT}}$, the LPT approximation to the Eulerian position $\boldsymbol{x}$ of matter particles, and Φ is the gravitational potential obtained by solving the Poisson equation with $\nabla_{\boldsymbol{x}}$ the gradient operator in Eulerian comoving coordinates. Time integration is performed by discretizing the derivative ∂_a^2 only on the left-hand side of equation (17), while the (second-order) LPT displacements are computed analytically and stored. L-PICOLA has its own initial conditions generator and uses a slightly modified version of the 2LPTIC code.[5]To generate L-PICOLA snapshots and extract statistics, we set the free parameters as presented in Table 3. Justification for these choices, along with more details on COLA and the L-PICOLA implementation, can be found in Appendix C.

4 APPLICATION AND RESULTS

In this section, we apply the CARPool technique to three standard cosmological statistics: the matter power spectrum, the matter bispectrum, and the one-dimensional probability density function (PDF) of matter fractional overdensity. We seek to improve the precision of

[5]The parallelized version of the code is available at http://cosmo.nyu.edu/roman/2LPT/

Table 3. Characteristics of L-PICOLA simulations.

Characteristic/parameter	Value
Number of timesteps	20 (linearly spaced)
Modified timestepping from Tassev et al. (2013)	$nLPT = +0.5$
Force mesh grid size	$N_{\rm m} = 512$
Starting redshift	$z_i = 127$
Initial conditions	Second-order Lagrangian perturbation theory (2LPT)
Redshift of data outputs	$z \in \{1.0, 0.5, 0.0\}$

estimates of theoretical expectations of these quantities as computed by GADGET-III. To assess the actual improvement, we need the sample mean $\bar{\boldsymbol{y}}$ of the *Quijote* simulations on the one hand, and the estimator (15) on the other hand.

Additionally, unless stated otherwise, each test case has the following characteristics:

(i) $N_{\rm max} = 500$ $\{\boldsymbol{y}_i, \boldsymbol{c}_i\}$ simulation pairs are generated, and the cumulative sample mean $\bar{\boldsymbol{y}}$ (resp. $\bar{\boldsymbol{x}}(\boldsymbol{\beta})$) is computed for every other 5 additional simulations (resp. simulation pairs).

(ii) $M = 1,500$ additional fast simulations are dedicated to the estimation of $\boldsymbol{\mu}_{\boldsymbol{c}}$.

(iii) The sample mean of 15 000 N-body simulations, accessible in the *Quijote* database, is taken as the true $\boldsymbol{\mu}$.

(iv) $p = q$ since we post-process GADGET-III and L-PICOLA snapshots with the same analysis codes (e.g. same vector size for $\boldsymbol{y}$ and $\boldsymbol{c}$).

(v) The analysis is performed at redshift $z = 0.5$. The lower the redshift, the more non-linear (and hence more difficult) the structure formation problem. We pick the lowest redshift that is relevant for upcoming galaxy surveys. We expect CARPool to be even more efficient for higher redshifts.

(vi) $\delta(\boldsymbol{x}) \equiv \rho(\boldsymbol{x})/\bar{\rho} - 1$ is the matter density contrast field; the first term designates the matter fractional overdensity field computed with the Cloud-in-Cell (CiC) mass assignment scheme. $\boldsymbol{x}$ exceptionally denotes the three-dimensional comoving grid coordinates here.

(vii) $N_{\rm grid}$ designates the density contrast grid size when post-processing snapshots.

(viii) We use bias-corrected and accelerated (BCa) bootstrap,[6] with $B = 5\,000$ samples with replacement, to compute the 95 per cent confidence intervals of the estimators. Efron & Tibshirani (1994) explain the computation.

The procedure of the method is illustrated in Fig. 1. The first step is to run M fast surrogates to compute the approximate mean $\boldsymbol{\mu}_{\boldsymbol{c}}$. How large M should be depends on the accuracy demanded by the user. Then, for each newly picked initial condition, both the expensive simulation code and the low-fidelity method are run to produce a snapshot pair. Only in this step do we need to run the high-fidelity simulation code N times. The mean (15) can be computed for each additional pair to track the estimate. In the next section, we assess the capacity of CARPool to use less than 10 simulations and a set of fast surrogates to match the precision of a large number of N-body simulations. All the statistics are calculated from the snapshots with the PYTHON 3 module PYLIANS3.[7]

[6]Available at https://github.com/cgevans/scikits-bootstrap

[7]Available at https://github.com/franciscovillaescusa/Pylians3

4.1 Matter power spectrum

This section is dedicated to estimating the power spectrum of matter density in real space at $z = 0.5$, the lower end of the range covered by next-generation galaxy redshift surveys. The density contrast $\delta(\boldsymbol{x})$ is computed from each snapshot with the grid size $N_{\rm grid} = 1024$. The publicly available power spectra range from $k_{\rm min} = 8.900 \times 10^{-3}\,h{\rm Mpc}^{-1}$ to $k_{\rm max} = 5.569\,h{\rm Mpc}^{-1}$ and contain 886 bins. The following analysis is restricted to $k_{\rm max} = 1.194\,h{\rm Mpc}^{-1}$ that results in 190 bins. We simplify our test case by compressing the power spectra into $p = 95$ bins, using the appropriate re-weighting by the number of modes in each k bin given in PYLIANS3. Univariate CARPool gives the best results since we are using the smallest possible number of costly N-body simulations; for this reason, power spectrum estimates using the multivariate framework are not shown here. As we discuss in appendix C, we intentionally run our fast surrogate (COLA) in a mode that produces a power spectrum that is highly biased compared to the full simulations, with a power deficit of more than 60 per cent on small scales.

4.1.1 *CARPool versus* N-*body estimates*

Fig. 2 shows the estimated power spectrum with 95 per cent confidence intervals enlarged by a factor of 20 for better visibility. Only 5 N-body simulations are needed to compute an unbiased estimate of the power spectrum with much higher precision than 500 N-body runs on large scales and on the scale of Baryon Acoustic Oscillations (BAO). On small scales, confidence intervals are of comparable size.[8]

We must verify that these results are not produced by a 'lucky' set of 5 simulation pairs. To this end, we compute 100 CARPool means $\bar{\boldsymbol{x}}(\widehat{\boldsymbol{\beta}^{\rm diag}})$ from distinct sets of five random seeds. The CARPool estimates fall within a sub- per cent accuracy relative to the sample mean from 15 000 N-body simulations, as illustrated by the upper panel of Fig. 3. The GADGET sample mean percentage error of 500 simulations with respect to 15 000 simulations is plotted with 95 per cent confidence intervals. We stress here that every percentage error plot in this paper shows an error with respect to 15 000 N-body simulations. The mean of 500 GADGET realizations is thus not at zero per cent, though the difference is very small.

4.1.2 *Beta smoothing*

Since we use a very small number of simulations, the estimates of the diagonal elements of $\widehat{\boldsymbol{\beta}^{\rm diag}}$ are noisy. This leads to some heavy tailed distributions for the CARPool estimates. Using the freedom we have to modify $\boldsymbol{\beta}$ without affecting unbiasedness, we can exploit the fact that we expect neighboring bins to have similar optimal $\boldsymbol{\beta}$. Convolving the diagonal elements with a five-bin-wide top-hat window slightly reduces the spread at small scales of CARPool estimates computed with only five GADGET power spectra and removes outliers. The comparison of the two panels in Fig. 3 illustrates this point. Using a nine-bin-wide Hanning window for the smoothing yields similar results. We call this technique beta

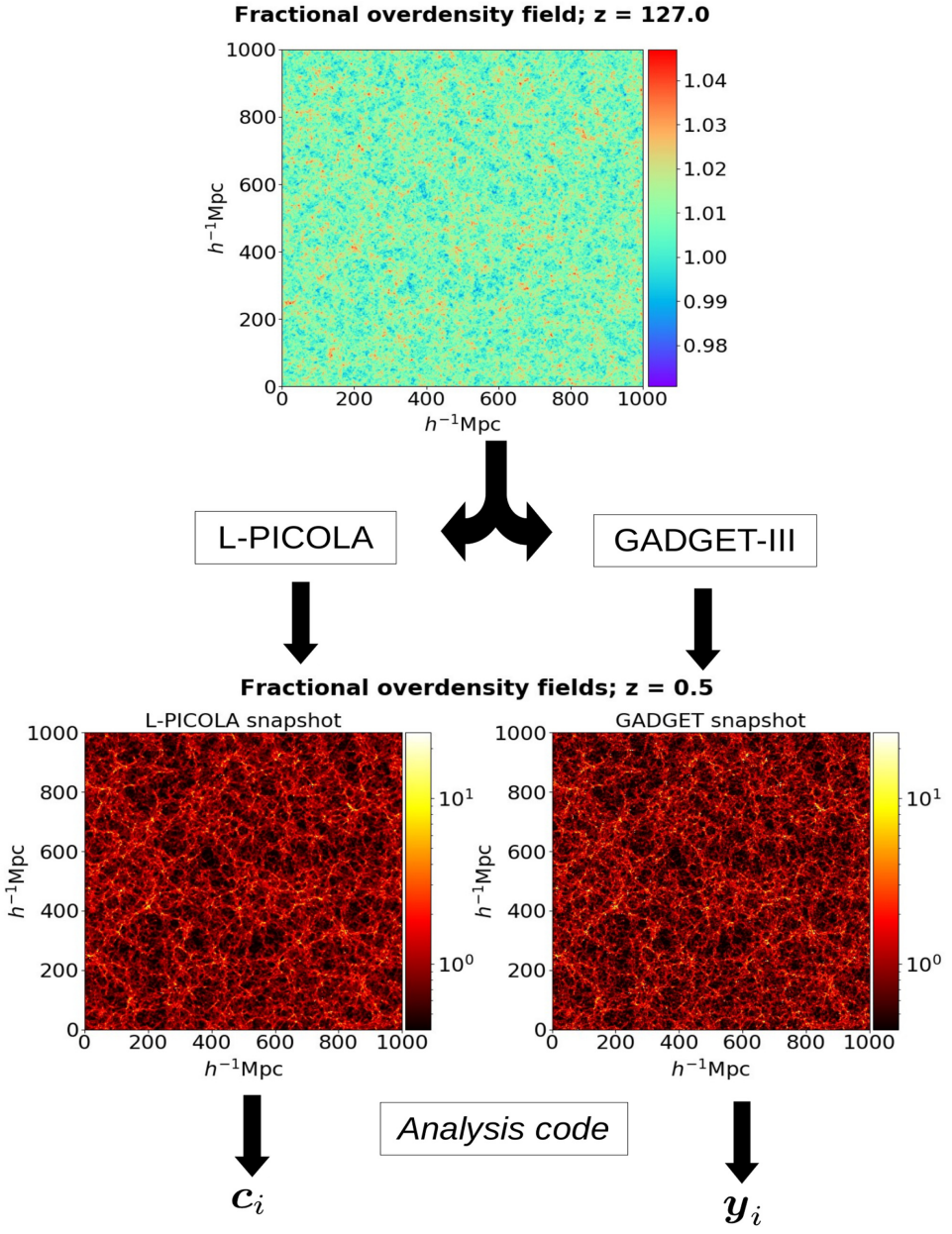
[8]While bootstrap is robust for estimating the 95 per cent error bars of a sample mean with 500 simulation, it is not equally reliable with a very small number of realizations. This leads to large bin-to-bin variations of the estimated CARPool confidence intervals in Fig. 2. An alternative, parametric computation of confidence intervals with very few samples can be found in Appendix B, using Student t-score values.

If $\mathbb{E}[\boldsymbol{c}] = \boldsymbol{\mu_c}$ is unknown:

| Estimate $\bar{\boldsymbol{\mu}}_{\boldsymbol{c}}$ from seeds $\{r_1, \ldots, r_M\} = \mathcal{S}_M$ (Cheap sims)

For $i \in [\![0, N-1]\!]$, $N \ll M$:

Pick initial conditions for $r_i \in \mathcal{S}_N$; $\mathcal{S}_N \cap \mathcal{S}_M = \emptyset$



Evaluate $\hat{\boldsymbol{\beta}}$, $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{c}}$ from $\{\boldsymbol{y}_i, \boldsymbol{c}_i\} \in \mathcal{S}_N$

Compute $\bar{\boldsymbol{x}}(\hat{\boldsymbol{\beta}}) = \bar{\boldsymbol{y}} - \hat{\boldsymbol{\beta}}\,(\bar{\boldsymbol{c}} - \bar{\boldsymbol{\mu}}_{\boldsymbol{c}})$

Figure 1. Flowchart of the practical application of CARPool to cosmological simulations. We highlight the estimation of $\boldsymbol{\mu_c}$ as a precomputation step using M fast simulations. The larger the M, the less impacted the variance/covariance of the control variates estimator, as expressed in (11) and Appendix A. The fractional overdensity images are projected slices of 60 h^{-1}Mpc.

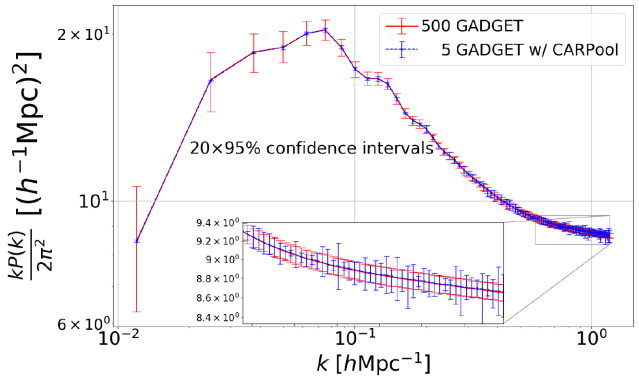


Figure 2. Estimated power spectrum with 500 N-body simulations versus 5 pairs of 'N-body + cheap' simulations, from which $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ is derived. The estimated 95 per cent confidence intervals are computed with the BCa bootstrap. They are enlarged by a factor of 20 for better visibility.

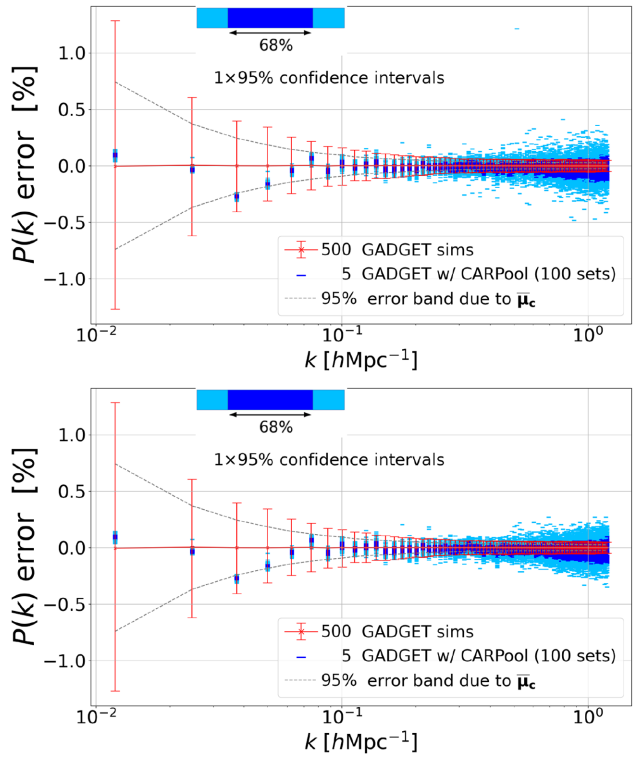


Figure 3. Estimated power spectrum percentage error with respect to 15 000 N-body runs: 500 N-body simulations versus 100 sets of five pairs of 'N-body + cheap' simulations. Each set uses a distinct $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$, calculated with the same seeds used for $\bar{\boldsymbol{x}}$. The upper panel estimate uses $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ while the lower panel convolves the diagonal elements of $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ with a narrow top-hat window. Beta smoothing removes outliers and Gaussianizes the tails by effectively increasing the number of degrees of freedom for each β estimate. Both panels use the same random seeds. The estimated 95 per cent confidence intervals are plotted for the N-body sample mean only, using BCa bootstrap. The dark blue symbols show the 68 per cent percentile of the CARPool estimates ordered by the absolute value of the percentage error; the rest appears in light blue symbols.

smoothing and use it with a five-bin-wide top-hat window in what follows.

Both panels of Fig. 3 show the symmetric 95 per cent confidence intervals of the surrogate mean with grey dashed lines. They represent

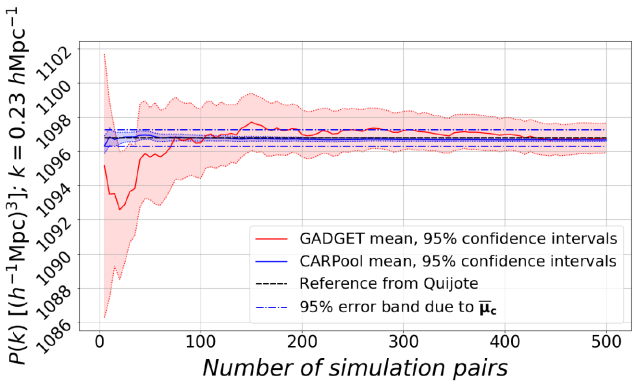


Figure 4. Convergence of a single k-bin at the BAO scale: the cumulative sample mean $\bar{\boldsymbol{y}}$ of N-body simulations versus the sample mean $\bar{\boldsymbol{x}}(\widehat{\boldsymbol{\beta}^{\mathbf{diag}}})$. Confidence intervals take into account that $\boldsymbol{\beta}^{\mathbf{diag}}$ is estimated from the same number of samples used to compute the CARPool estimate of $P(k)$.

the 95 per cent error band likely to stem from the estimation of $\boldsymbol{\mu}_c$, relatively to the mean of 15 000 GADGET simulations, hence the fact that, at large scales especially, the CARPool means concentrate slightly away from the nullpercentage error. Though the unbiased estimator in equation (15) takes a precomputed cheap mean, the practitioner can decide to run more approximate simulations on the fly to improve the accuracy of $\bar{\boldsymbol{\mu}}_c$. Note that the CARPool means with 5 N-body simulations still land withing the 95 per cent confidence intervals from 500 GADGET simulations, even at large scales where the difference due to the surrogate mean is visible.

Fig. 4 exhibits the convergence of one power spectrum bin at the BAO scale as we add more simulations: the 95 per cent error band of the control variates estimate shrinks extremely fast compared to that of the N-body sample mean.

4.1.3 Empirical variance reduction

The left-hand panel of Fig. 5 shows the empirical generalized variance reduction of the CARPool estimate compared to the standard estimate, as defined in equation (9). The vertical axis corresponds to the volume ratio of two parallelepipeds of dimension $p = 95$, in other words the volume ratio of error 'boxes' for two estimators. The determinant $\det\left(\widehat{\boldsymbol{\Sigma}_{yy}}\right)$ is fixed because we take all 15 000 N-body simulations available in *Quijote* to compute the most accurate estimate of $\boldsymbol{\Sigma}_{yy}$ we have access to, whereas $\det\left(\boldsymbol{\Sigma}_{xx}(\hat{\boldsymbol{\beta}})\right)$ changes each time new simulation pairs are run. More precisely, for each data point in Fig. 5, we take the control matrix estimate computed with $5k$, $k \in [\![1, 100]\!]$ simulation pairs and generate 3000 $\boldsymbol{x}$ samples according to (14) to obtain an estimator of $\boldsymbol{\Sigma}_{xx}$. For that, we use 3000 *Quijote* simulations and 3000 additional L-PICOLA surrogates run with the corresponding seeds.

The simpler univariate scheme outperforms the estimation of the optimal $\boldsymbol{\beta}^\star$ for $N = 5k$, $k \in [\![1, 100]\!]$, corroborating the experiments of Section 4.1.1. Furthermore, variance reduction granted by a sub-optimal diagonal $\boldsymbol{\beta}^{\mathbf{diag}}$ improves rapidly and reaches its apparent limit quickly. We suspect that the slight worsening of the variance reduction, when the number of available samples to estimate $\boldsymbol{\beta}^\star$ neighbors the vector size p, is linked to the eigenspectrum of $\boldsymbol{\Sigma}_{c,c}^\dagger$ and could be improved by projecting out the eigenmodes corresponding to the smallest, noisiest eigenvalues.

We depict the scale-dependent performance of CARPool for the matter power spectrum in the right-hand panel of Fig. 5. The vertical

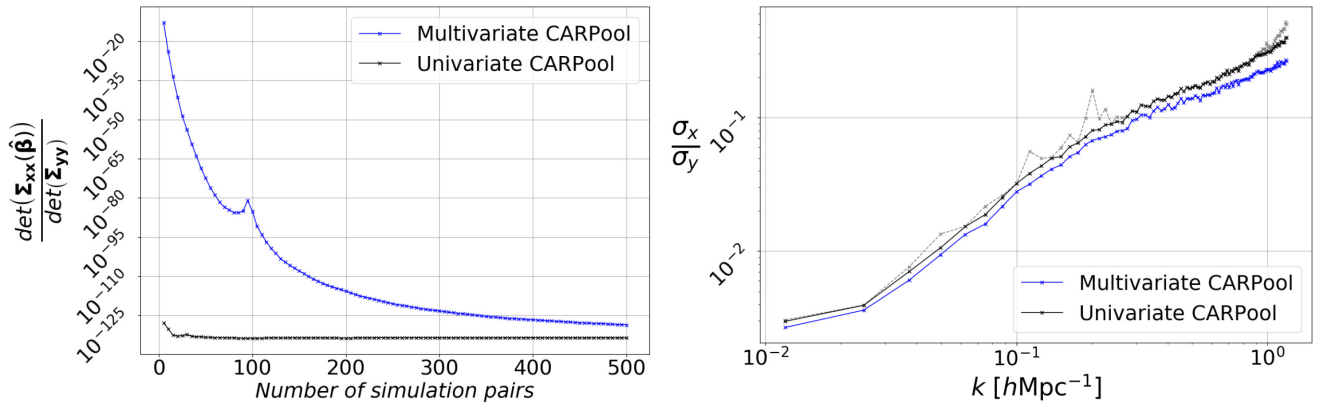


Figure 5. Left-hand panel: Generalized variance ratio for the power spectrum up to $k_{\max} \approx 1.2\ h\mathrm{Mpc}^{-1}$ as a function of the number of available simulations. Each $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ serves to generate 3000 samples according to (14) to estimate the CARPool covariance matrix. Right-hand panel: Standard deviation reduction for each power spectrum bin due to CARPool. The blue and black curves use $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ estimated with 500 samples. The dashed grey curve exhibits the actual standard deviation ratio when we have five samples only to compute $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$. $\mathbf{\Sigma}_{yy}$ is estimated using all 15 000 available power spectra from the *Quijote* simulations.

axis is the variance reduction to expect from the optimal control coefficients (or matrix). Namely, we take the data points of the left panel for 500 simulation/surrogate pairs, extract the diagonal of the covariance matrices, and divide the arrays. The blue and black curves show the variance reduction with respect to the sample mean of N-body simulations using all 500 simulation/surrogate pairs to estimate the control matrix. In practice, we estimate $\boldsymbol{\beta}$ using only five simulation/surrogate pairs; does this noisy $\hat{\boldsymbol{\beta}}$ lead to significant inefficiency? The grey dashed curve shows the actual standard deviation reduction brought by the rough estimate of $\boldsymbol{\beta}^{\mathbf{diag}}$ using five simulation pairs only, with which the results of Figs 2 and 3 are computed. A few k-bins fluctuate high but the variance reduction remains close to optimal, especially considering that only five simulations were used, and we have not attempted any further regularization except for beta smoothing.

4.2 Matter bispectrum

We compute the shot-noise corrected matter bispectrum in real space (Hahn et al. 2020; Villaescusa-Navarro et al. 2020), using PYSPECTRUM[9] with $N_{\mathrm{grid}} = 360$ and bins of width $\Delta k = 3k_{\mathrm{f}} = 1.885 \times 10^{-2}\ h\mathrm{Mpc}^{-1}$, where $k_{\mathrm{f}} = \frac{2\pi}{L}\ h\mathrm{Mpc}^{-1}$ is the fundamental mode depending on the box size L. As in the previous section, we present only the results using $\boldsymbol{\beta}^{\mathbf{diag}}$ instead of $\boldsymbol{\beta}^{\star}$. We examine two distinct sets of bispectrum coefficients: in the first case we study the bispectrum for squeezed isosceles triangles as a function of opening angle only, averaging over scale; in the second case we compute equilateral triangles as a function of k.

4.2.1 *Squeezed isosceles triangles*

We start the analysis by regrouping isosceles triangles ($k_1 = k_2$) and re-weighting the bispectrum monopoles for various k_3/k_1 ratios in ascending order. Only squeezed triangles are considered here: $(k_3/k_1)_{\max} = 0.20$ so that the dimension of $\boldsymbol{y}$ is $p = 98$ (see Table 1).

[9]Available at https://github.com/changhoonhahn/pySpectrum

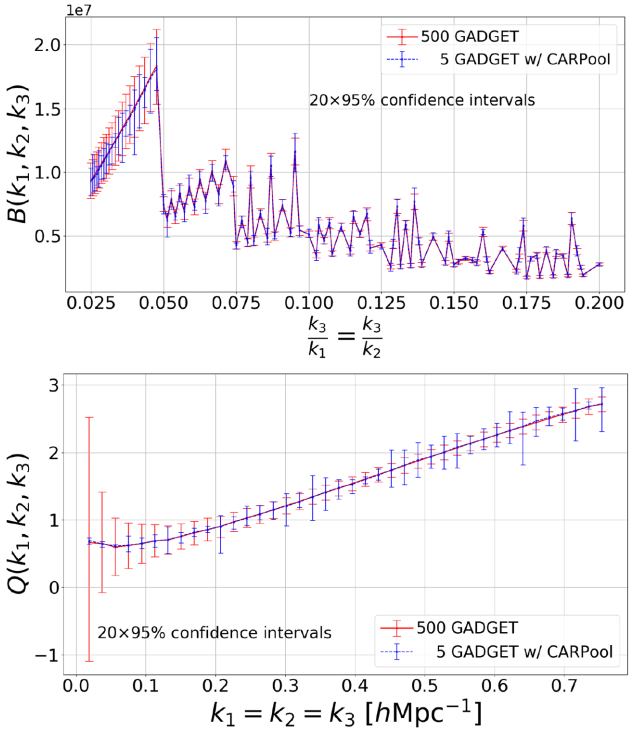


Figure 6. Upper panel: Estimated bispectrum for squeezed isosceles triangles with 500 N-body simulations versus 5 pairs of 'N-body + cheap' simulations, from which the smoothed $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ is derived. The estimated 95 per cent confidence intervals are computed with the BCa bootstrap. They are enlarged by a factor of 20 for better visibility. Lower panel: As in the upper panel, but for the reduced bispectrum of equilateral triangles.

4.2.2 *CARPool versus* N*-body estimates*

On the order of 5 samples are required to achieve a precision similar to that of the sample mean of 500 N-body simulations as we show in Fig. 6 (upper panel). Fig. 7 (upper panel) corroborates the claim by showing thepercentage error of 100 CARPool means using 5 costly simulations each. The reference is the mean of the 15 000 bispectra from the *Quijote* simulations. As in the previous section, we show

the 95 per cent error band due to estimation of the surrogate mean $\boldsymbol{\mu}_c$ with dashed curves.

4.2.3 *Empirical variance reduction*

As for the power spectrum, the upper left-hand panel of Fig. 8 shows that the generalized variance reduction is much more significant when separately estimating control coefficients for each triangle configuration. The right-hand side of the curve suggests an increasing improvement of the multivariate case, but in this range of numbers of required samples the variance reduction scheme loses its appeal. We have used 1800 additional simulations to compute the covariance matrices intervening in the generalized variance estimates. In the upper right-hand panel of the figure, the calculation of the standard deviation ratio for each triangle configuration follows the same logic as in Section 4.1.3. The grey dashed curve corresponds to the standard deviation reduction brought by control coefficients (i.e. the univariate CARPool framework) estimated with 5 simulation/surrogate pairs only.

4.2.4 *Equilateral triangles*

Here, we analyse equilateral triangles with the modulus of $k_1 = k_2 = k_3$ varying up to $k_{\rm max} = 0.75\,h{\rm Mpc}^{-1}$ ($p = 40$). For better visibility, we show the reduced bispectrum monopole $Q(k_1, k_2, k_3)$.

4.2.5 *CARPool versus* N*-body estimates*

Similarly to the previous set of triangle configurations, we compare the precision of the CARPool estimator using 5 N-body simulations with that of the sample mean from 500 GADGET runs. Fig. 6 (lower panel) exhibits the estimated reduced bispectrum with five seeds, while Fig. 7 (lower panel) shows the relative error of various CARPool sets with respect to the reference from 15 000 N-body samples.

4.2.6 *Empirical variance reduction*

In Fig. 8 (lower panels), we observe a trend similar to that of the previous experiments: the univariate control coefficients are much better than the control matrix in terms of generalized variance reduction for a realistic number of full N-body simulations.

4.3 Probability density function of smoothed matter fractional overdensity

The power spectrum and the bispectrum are Fourier-space statistics. How does CARPool fare on a purely direct-space statistic? In the *Quijote* simulations, the probability density function of the matter fractional overdensity, or the *matter PDF*, is computed on a grid with $N_{\rm grid} = 512$, smoothed by a top-hat filter of radius R. There are 100 histogram bins in the range $\rho/\bar{\rho} \in \left[10^{-2}, 10^{2}\right]$. We work with the $R = 5\ h^{-1}$Mpc case and restrict the estimation of the PDF to the interval $\rho/\bar{\rho} \in \left[8 \times 10^{-2}, 5 \times 10^{1}\right]$ that contains $p = 70$ bins. Note that we intentionally do not do anything to improve the correspondence of the surrogate and simulation histograms, an example of which is displayed in Fig. 9.

4.3.1 *Empirical variance reduction*

For the matter PDF, we show the empirical variance reduction results before the actual estimates: Fig. 10 shows that the variance reduction is much milder for the PDF than for the power spectrum

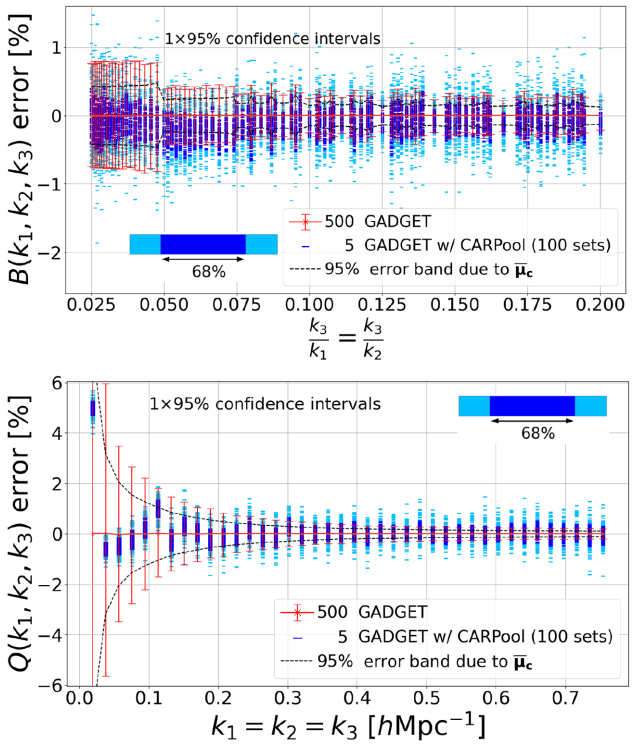


Figure 7. Upper panel: Estimated bispectra percentage error for squeezed isosceles triangles with respect to 15 000 N-body runs: 500 N-body simulations versus 100 sets of 5 pairs of 'N-body + cheap' simulations. Each set uses a distinct $\widehat{\boldsymbol{\beta}^{\rm diag}}$, calculated with the same seeds intervening in $\bar{\boldsymbol{x}}$ and smoothed by a five-bin-wide flat window. The estimated 95 per cent confidence intervals are plotted for the N-body sample mean only, using BCa bootstrap. The dark blue symbols show the 68 per cent percentile of the CARPool estimates ordered by the absolute value of the percentage error; light-blue symbols represent the rest. Lower panel: As in the upper panel, but for the reduced bispectrum of equilateral triangles.

or the bispectrum, both for the univariate and multivariate CARPool frameworks. While the multivariate case does eventually lead to significant gains, CARPool needs $\mathcal{O}(100)$ simulations to learn how to map density contrast in COLA outputs to density contrast in GADGET-III simulations. While COLA places overdense structures close to the right position, their density contrast is typically underestimated, meaning a level sets of the COLA output is informative about a different level set of the GADGET-III simulation.

The right-hand panel none the less proves that it is possible to reduce the variance of the one-point PDF with CARPool, unlike with paired-fixed fields (Villaescusa-Navarro et al. 2018). As for the bispectrum, we took the data outputs of 1800 additional simulations to compute the covariance matrices intervening in the generalized variance and standard error estimates.

4.3.2 *CARPool versus* N*-body estimates*

For the matter PDF we compare CARPool estimates in both the multivariate and univariate settings. Figs 11 and 12 are paired and show the comparable performance at the tails of the estimated PDF for the smoothed $\widehat{\boldsymbol{\beta}^{\rm diag}}$ with 50 samples on the one hand, and the dense $\widehat{\boldsymbol{\beta}}$ matrix obtained with 125 simulations on the other. We can expect $\mathcal{O}(10^1)$ fewer N-body simulations to compute an accurate estimate of the PDF when applying the simple univariate CARPool

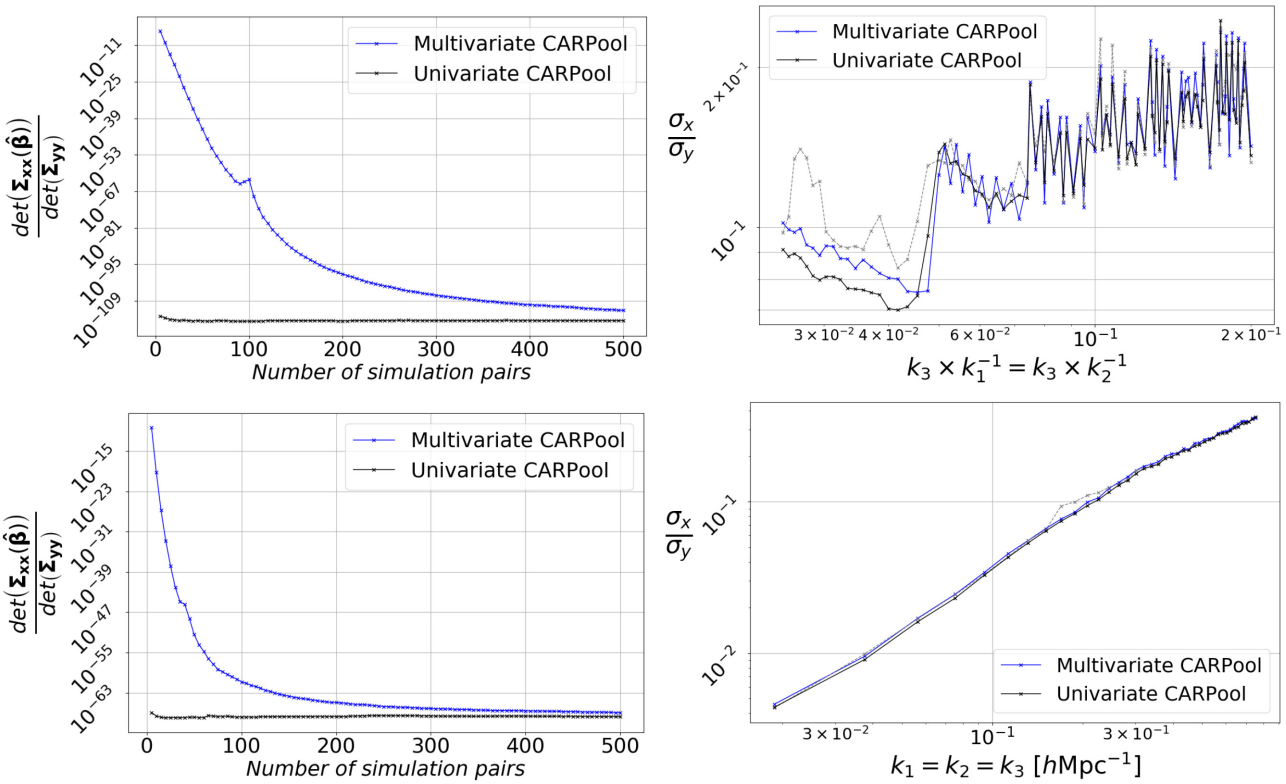


Figure 8. Upper left-hand panel: Generalized variance ratio of bispectrum for squeezed isosceles triangles as a function of the number of available simulations. Each $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ serves to generate 1800 samples according to (14) to estimate the CARPool covariance matrix. Upper right-hand panel: Standard deviation reduction for each squeezed isosceles triangle to expect from CARPool. The blue and black curves respectively use $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$ estimated with 500 samples. The dashed grey curve exhibits the actual standard deviation ratio when we have five samples only to compute $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$. $\boldsymbol{\Sigma}_{yy}$ is estimated with all 15 000 available bispectra from the *Quijote* simulations. Lower panels: As in the upper panels, but for the reduced bispectrum of equilateral triangles.

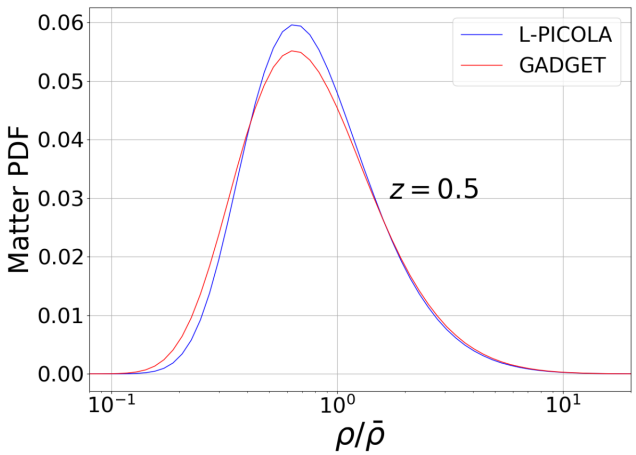


Figure 9. Probability density function of the smoothed matter fractional overdensity of GADGET-III and L-PICOLA snapshots at $z = 0.5$ for the same initial conditions. The characteristics of L-PICOLA are provided in Table 3.

technique (50 instead of 500 here). As discussed above, with enough simulations CARPool can learn the mapping between the density contrasts of COLA and GADGET outputs. Therefore, the matter PDF is a case where the multivariate framework, which involves the estimation of $p \times p$ covariance matrices, shows improvement over the more straightforward univariate case once the number of available simulation pairs passes a threshold.

While we wanted to test the performance of CARPool with minimal tuning, we expect that with some mild additional assumptions and tuning the univariate CARPool approach could be improved and similar gains to the multivariate case could be obtained with a smaller number of simulations. As an example, one could pre-process the COLA outputs to match the PDF (and power spectrum) of GADGET-III using the approach described in Leclercq et al. (2013) to guarantee a close correspondence between bins of density contrast. In addition, a regularizing assumption would be to consider transformations from COLA to GADGET-III density contrasts that are smooth and monotonic.

4.4 Summary of results

Here we present a summary of the variance reduction observed in our numerical experiments. With $M = 1500$ additional fast simulations reserved for estimating the cheap mean $\bar{\boldsymbol{\mu}}_{\boldsymbol{c}}$, and with percentage errors relative to the mean of 15 000 full N-body runs available in *Quijote*, we find:

(i) With only 5 N-body simulations, the univariate CARPool technique recovers the 95-bin power spectrum up to $k_{\max} \approx 1.2\,h\mathrm{Mpc}^{-1}$ within the 0.5 per cent error band, when the control coefficients are smoothed.

(ii) For the bispectrum of 98 squeezed isosceles triangle configurations, the recovery is within 2 per cent when 5 N-body simulations are available, and 1 per cent when we have 10 of them, still with the smoothed $\widehat{\boldsymbol{\beta}^{\mathbf{diag}}}$.

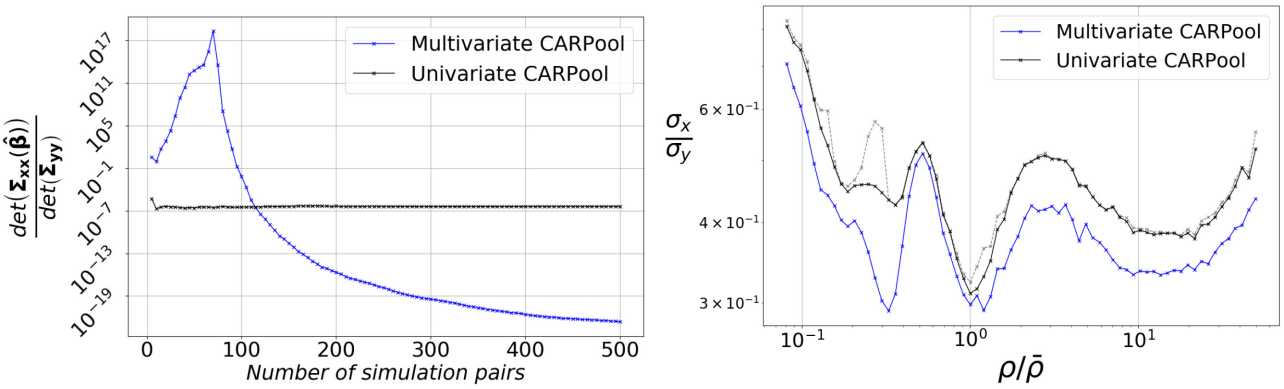


Figure 10. Left-hand panel: Generalized variance ratio of the matter PDF as a function of the number of available simulations. Each $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}^{\mathrm{diag}}}$ serves to generate 1800 samples according to (14) to estimate the CARPool covariance matrix. Right-hand panel: Standard deviation reduction for the PDF bin to expect from CARPool. The blue and black curves respectively use $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}^{\mathrm{diag}}}$ estimated with 500 samples. The dashed grey curve exhibits the actual standard deviation ratio when we have 10 samples only to compute $\widehat{\boldsymbol{\beta}^{\mathrm{diag}}}$. $\mathbf{\Sigma}_{yy}$ is estimated with all 15 000 available PDFs from the *Quijote* simulations.

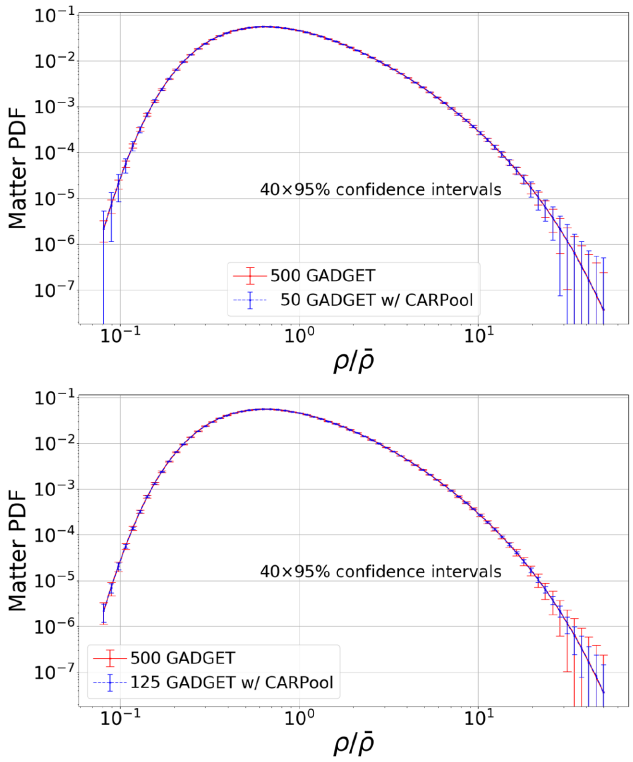


Figure 11. Estimated matter PDF with 500 *N*-body simulations versus CARPool estimates. $\widehat{\boldsymbol{\beta}^{\mathrm{diag}}}$ is used in the upper panel whereas the full control matrix is computed in the lower panel. The estimated 95 per cent confidence intervals are computed with the BCa bootstrap. They are enlarged by a factor of 40 for better visibility.

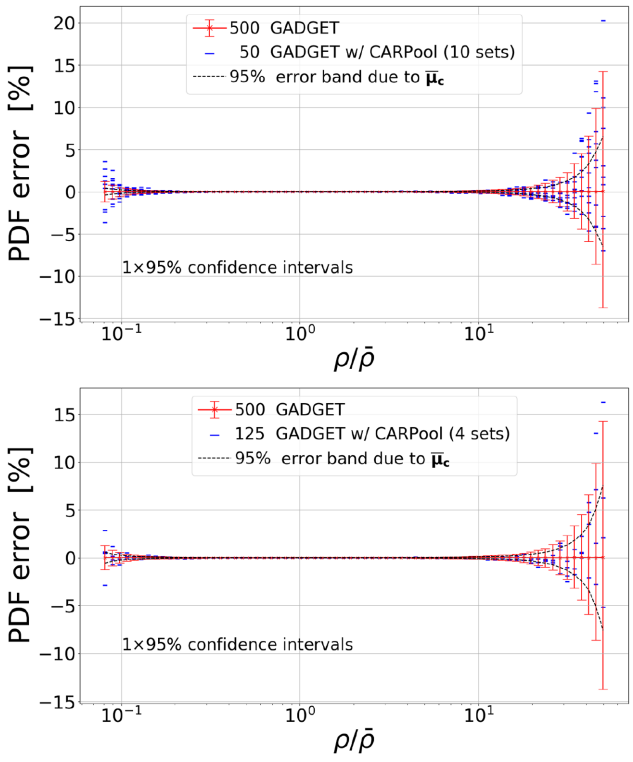


Figure 12. Estimated matter PDF percentage error with respect to 15 000 *N*-body runs: sample mean of 500 *N*-body simulations versus CARPool estimates. In the upper panel, $\widehat{\boldsymbol{\beta}^{\mathrm{diag}}}$ is used for each set and smoothed by a five-bin-wide flat window. In the lower panel, the full control matrix $\widehat{\boldsymbol{\beta}}$ is estimated for each group of seeds. The estimated 95 per cent confidence intervals are plotted for the *N*-body sample mean only, using BCa bootstrap.

(iii) The bispectrum estimator of equilateral triangles on 40 bins falls within the 2 per cent (resp. 1 per cent) error band with 5 simulations (resp. 10) at large k, and performs better than the mean of 500 GADGET simulations at large scales.

(iv) The standard deviation of matter PDF bins can also be reduced with CARPool, by factors between 3 and 10, implying that the number of required costly simulations is lowered by an order of magnitude.

In Appendix B, we provide the power spectrum and bispectrum results when the CARPool means are computed with 10 simulation/surrogate pairs instead of the 5 pairs presented so far.

5 DISCUSSION AND CONCLUSIONS

We presented CARPool, a general scheme for reducing variance on estimates of large-scale structure statistics. It operates on the

idea of forming a combination (pooling) of a small number of accurate simulations with a larger number of fast but approximate surrogates in such a way as to not introduce systematic error (zero bias) on the combination. The result is equivalent to having run a much larger number of accurate simulations. This approach is particularly adapted to cosmological applications where our detailed physical understanding has resulted in a number of perturbative and non-perturbative methods to build fast surrogates for high-accuracy cosmological simulations.

To show the operation and promise of the technique, we computed high-accuracy and low-variance predictions for statistics of GADGET-III cosmological N-body simulations in the ΛCDM model at $z = 0.5$. A large number of surrogates are available; for illustration we selected the approximate particle mesh solver L-PICOLA.

For three different examples of statistics, the matter power spectrum, the matter bispectrum, and the probability density function of the matter fractional overdensity, CARPool reduces variance by factors 10 to 100 even in the non-linear regime, and by much larger factors on large scales. Using only five GADGET-III simulations CARPool is able to compute Fourier-space two-point and three-point functions of the matter distribution at a precision comparable to 500 GADGET-III simulations.

CARPool requires (i) inexpensive access to surrogate solutions, and (ii) strong correlations of the fluctuations about the mean of the surrogate model with the fluctuations of the expensive and accurate simulations. By construction, CARPool estimates are unbiased compared to the full simulations no matter how biased the surrogates might be. In all our examples, we achieved substantial variance reductions even though the fast surrogate statistics were highly biased compared to the full simulations.

So far we have presented CARPool as a way to accelerate the convergence of ensemble averages of accurate simulations. An equivalent point of view would be to consider it a method to remove approximation error from ensembles of fast mocks by running a small number of full simulations. Such simulations often already exist, as in our case with the *Quijote* simulations, not least because strategies to produce fast surrogates are often tested against a small number of simulations.

In some cases there are opportunities to use CARPool almost for free: for instance, using linear theory from the initial conditions as a surrogate model has the advantage that $\boldsymbol{\mu_c}$ (the mean linear theory power spectrum) is perfectly known *a priori*. In addition, the decorrelation between linearly and non-linearly evolved perturbations is well studied, and can be used to set $\boldsymbol{\beta}$. Even for just a single N-body simulation, and without the need to estimate $\boldsymbol{\mu_c}$ from an ensemble of surrogates, this would remove cosmic variance on the largest scales better than in our numerical experiments with L-PICOLA, which are limited by the uncertainty of the $\boldsymbol{\mu_c}$ estimate.

Regardless of the details of the implementation, the reduction of sample variance on observables could be used to avoid having to run ensembles of simulations (or even surrogates) at the full survey volume. This would simplify simulation efforts for upcoming large surveys since memory limitations rather than computational time are currently the most severe bottleneck for full-survey simulations (Potter et al. 2017).

In comparison to other methods of variance reduction, CARPool has the main advantage of guaranteeing lack of model error ('bias') compared to the full simulation. 'Fixing' (Angulo & Pontzen 2016; Pontzen et al. 2016) explicitly modifies the statistics of the generated simulation outputs; which observables are unbiased must be checked on a case-by-case basis, either through theoretical arguments or through explicit simulation (Villaescusa-Navarro et al. 2018). Klypin, Prada & Byun (2020) argue that 'fixed' field initialization is unsuitable for simulation suites to estimate accurate covariance matrices, and they are pessimistic about the possibility of generating mock galaxy catalogues solely with this technique.

Pontzen et al. (2016) and Angulo & Pontzen (2016) also introduce and study the 'pairing' technique. 'Pairing' reduces variance for k-space observables (such as the power spectrum) by a factor of $\mathcal{O}(1)$ by combining two simulations whose initial conditions only differ by an overall minus sign, that is they are *phase-flipped*. This technique can be analysed simply in the control variates framework of CARPool. Consider the phase-flipped simulation as the surrogate for the moment. The mean of an ensemble of phase-flipped simulations is identical to the mean of the unflipped simulations by symmetry. 'Pairing' then amounts to taking $\beta = -1$ to cancel off contributions of odd-order terms in the initial conditions (Angulo & Pontzen 2016; Pontzen et al. 2016) to reduce variance on the simulation output. Inserting this β in equation (2) and taking the expectation shows that 'pairing' is an unbiased estimator of the simulation mean.

Other opportunities of exploiting the control variates principle abound; related ideas have been used in the past. As an example, a very recent study (Smith et al. 2021) succeeds in reducing the variance of the quadrupole estimator of the two-point clustering statistic in redshift space. In this case, the variance reduction is achieved by combining different, correlated lines of sight through the halo catalogue of the Outer Rim simulation. Though not driven by a general theoretical framework that guarantees unbiasedness and optimal variance reduction, for the specific application at hand their approach does not require pre-computation of fast surrogates and uses a control matrix set based on physical assumptions.

While we intentionally refrained from tuning CARPool for this first study, there are opportunities to use physical insight to adapt it for cosmological applications. For instance, the one-point remapping technique proposed by Leclercq et al. (2013) that allows us to increase the cross-correlation between LPT-evolved density fields and full N-body simulations could improve snapshots of a chosen surrogate for CARPool.

In future work, we plan to explore intermediate forms of CARPool between the multivariate and univariate versions we study in this paper. Any given entry of $\boldsymbol{y}$ could be predicted by an optimal combination of a small subset of $\boldsymbol{c}$. In this case, the variance reduction could be improved compared to the univariate case while the reduced dimension of the control matrix would ensure a stable estimate using a moderate number of simulations.

The CARPool setup can be applied to numerous 'N-body code plus surrogate' couples for cosmology. It can be used to make high-resolution corrections to low-resolution simulations, while reducing variance. This will provide an alternative to the procedure suggested by Rasera et al. (2014), where the mass resolution effect is estimated by a polynomial fit of the matter power spectrum ratio, and the work of Blot et al. (2015), where a linear transformation of the low-resolution power spectra preserving the mean and variance is smoothed by a polynomial fit. Furthermore, rather than using a single surrogate, taking advantage of multiple low-fidelity methods for variance reduction is also a possibility to explore, especially if the cost of running a large number of surrogates is non-neglible. For instance, taking the linear theory as a second surrogate in addition to L-PICOLA would have strongly reduced the number of L-PICOLA runs required to match the variance of the $\boldsymbol{\mu_c}$ estimate to the massively reduced variance of $\boldsymbol{y} - \boldsymbol{\beta}\,(\boldsymbol{c} - \boldsymbol{\mu_c})$. In this regard, the multifidelity Monte Carlo scheme of Peherstorfer et al. (2016) and the approximate control variates framework of Gorodetsky et al. (2020) are recent techniques that reduce variance with multiple surrogates for a fixed computational budget. We can also combine CARPool with other techniques. For instance, if the paired-fixed fields initialization of

Angulo & Pontzen (2016) is found to be unbiased in practice for a particular statistic, then one can combine it with CARPool for further variance reduction.

The simplicity of the theory behind CARPool makes the method attractive for various applications both in and beyond cosmology, as long as the conditions given above are satisfied. Our results suggest that CARPool allows estimating the expectation values of any desired large-scale structure correlators with negligible variances from a small number of accurate simulations, thereby providing a useful complement to analytical approaches such as higher-order perturbation theory or effective field theory. We are planning to explore a number of these applications in upcoming publications.

ACKNOWLEDGEMENTS

We thank Martin Crocce, Janis Fluri, Cullan Howlett, and Hans Arnold Winther for their advice on COLA, and Boris Leistedt for stimulating discussions. We are grateful to Pier-Stefano Corasaniti, Eiichiro Komatsu, Marius Millea, Andrew Pontzen, Yann Rasera, and Matias Zaldarriaga for stimulating comments on an earlier version of the manuscript. Nicolas Chartier acknowledges funding from LabEx ENS-ICFP (PSL). Benjamin Wandelt 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. Yashar Akrami is supported by LabEx ENS-ICFP: ANR-10-LABX-0010/ANR-10-IDEX-0001-02 PSL∗. Francisco Villaescusa-Navarro acknowledges funding from the WFIRST program through NNG26PJ30C and NNN12AA01C.

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 and code examples are provided at https://github.com/CompiledAtBirth/pyCARPool to reproduce some results presented in this study.

REFERENCES

Aihara H. et al., 2018, PASJ, 70, S4
Angulo R. E., Pontzen A., 2016, MNRAS, 462, L1
Angulo R. E., Zennaro M., Contreras S., Aricò G., Pellejero-Ibañez M., Stücker J., 2020, preprint (arXiv:2004.06245)
Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
Blot L. et al., 2019, MNRAS, 485, 2806
Blot L., Corasaniti P. S., Alimi J. M., Reverdy V., Rasera Y., 2015, MNRAS, 446, 1756
Bouchet F., Colombi S., Hivon E., Juszkiewicz R., 1995, A&A., 296, 575
Carrasco J. J. M., Hertzberg M. P., Senatore L., 2012, J. High Energy Phys., 09, 082
Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621
Colavincenzo M. et al., 2019, MNRAS, 482, 4883
Crocce M., Scoccimarro R., 2006, Phys. Rev. D, 73, 063519
Crocce M., Castander F. J., Gaztañaga E., Fosalba P., Carretero J., 2015, MNRAS, 453, 1513
de O. Porta Nova A. M., Wilson J. R., 1993, Eur. J. Oper. Res., 71, 80
DeRose J. et al., 2019, ApJ, 875, 69
DESI Collaboration, 2016a, preprint (arXiv:1611.00036)
DESI Collaboration, 2016b, preprint (arXiv:1611.00037)
Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rep., 733, 1
Doré O. et al., 2014, preprint (arXiv:1412.4872)
Doré O. et al., 2018, preprint (arXiv:1805.05489)
Efron B., Tibshirani R. J., 1994, An Introduction to the Bootstrap. Chapman & Hall, New York
Euclid Collaboration, 2020, Astronomy & Astrophysics, 642, A191
Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, MNRAS, 463, 2273
Garrison L., 2019, PhD thesis, University Of Washington
Garrison L. H., Eisenstein D. J., Ferrer D., Tinker J. L., Pinto P. A., Weinberg D. H., 2018, ApJS, 236, 43
Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, preprint (arXiv:1406.2661)
Gorodetsky A. A., Geraci G., Eldred M. S., Jakeman J. D., 2020, J. Comput. Phys., 408, 109257
Goroff M. H., Grinstein B., Rey S. J., Wise M. B., 1986, ApJ, 311, 6
Habib S. et al., 2016, New Astron., 42, 49
Hahn C., Villaescusa-Navarro F., Castorina E., Scoccimarro R., 2020, JCAP, 2020, 040
Hartlap J., Simon P., Schneider P., 2006, A&A, 464, 399
He S., Li Y., Feng Y., Ho S., Ravanbakhsh S., Chen W., Póczos B., 2019, Proc. Natl. Acad. Sci., 116, 13825
Heitmann K. et al., 2019, ApJS, 245, 16
Heitmann K., White M., Wagner C., Habib S., Higdon D., 2010, ApJ, 715, 104
Helwig E. N., 2017, Canonical Correlation Analysis. http://users.stat.umn.edu/ helwig/notes/cancor-Notes.pdf
Howlett C., Manera M., Percival W., 2015, Astron. Comput., 12, 109
Ishiyama T. et al., 2020, preprint (arXiv:2007.14720)
Ishiyama T., Fukushige T., Makino J., 2009, PASJ, 61, 1319
Ivezić Ž. et al., 2019, ApJ, 873, 111
Izard A., Crocce M., Fosalba P., 2016, MNRAS, 459, 2327
Jain B., Bertschinger E., 1994, ApJ, 431, 495
Klypin A., Prada F., Byun J., 2020, MNRAS, 496, 3862
Kodi Ramanah D., Charnock T., Villaescusa-Navarro F., Wandelt B. D., 2020, MNRAS, 495, 4227
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Lavenberg S., Welch P., 1981, Manag. Sci., 27, 322
Leclercq F., Jasche J., Gil-Marín H., Wandelt B., 2013, JCAP, 2013, 048
Leclercq F., Faure B., Lavaux G., Wandelt B. D., Jaffe A. H., Heavens A. F., Percival W. J., Noûs C., 2020, Astronomy & Astrophysics, 639, A91
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Lippich M. et al., 2019, MNRAS, 482, 1786
LSST Dark Energy Science Collaboration, 2018, preprint (arXiv:1809.01669)
LSST Science Collaboration, 2009, preprint (arXiv:0912.0201)
Matsubara T., 2008, Phys. Rev. D, 77, 063530
McClintock T. et al., 2019a, preprint (arXiv:1907.13167)
McClintock T. et al., 2019b, ApJ, 872, 53
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Peherstorfer B., Willcox K., Gunzburger M., 2016, SIAM J. Sci. Comput., 38, A3163
Perko A., Senatore L., Jennings E., Wechsler R. H., 2016, preprint (arXiv:1610.09321)
Planck Collaboration, 2020, Astronomy and Astrophysics, 641, A6
Pontzen A., Slosar A., Roth N., Peiris H. V., 2016, Phys. Rev. D, 93, 103519
Potter D., Stadel J., Teyssier R., 2017, Comput. Astrophys. Cosmol., 4, 2
Quinn T., Katz N., Stadel J., Lake G., 1997, preprint (astro–ph/9710043)
Rasera Y., Corasaniti P. S., Alimi J. M., Bouillot V., Reverdy V., Balmès I., 2014, MNRAS, 440, 1420
Ronneberger O., Fischer P., Brox T., 2015, preprint (arXiv:1505.04597)
Rubinstein R. Y., Marcus R., 1985, Oper. Res., 33, 661
Skillman S. W., Warren M. S., Turk M. J., Wechsler R. H., Holz D. E., Sutter P. M., 2014, preprint (arXiv:1407.2600)
Smith R. E., Angulo R. E., 2019, MNRAS, 486, 1448
Smith A., de Mattia A., Burtin E., Chuang C.-H., Zhao C., 2021, MNRAS, 500, 259
Spergel D. et al., 2015, preprint (arXiv:1503.03757)
Springel V., 2005, MNRAS, 364, 1105

Square Kilometre Array Cosmology Science Working Group, 2020, PASA, 37, e007
Taffoni G., Monaco P., Theuns T., 2002, MNRAS, 333, 623
Tamura N. et al., 2016, in Proc. SPIE, Ground-based and Airborne Instrumentation for Astronomy VI. SPIE, Bellingham, p. 99081M
Tassev S., Zaldarriaga M., 2012, JCAP., 2012, 013
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, JCAP, 2013, 036
Tassev S., Eisenstein D. J., Wandelt B. D., Zaldarriaga M., 2015, preprint (arXiv:1502.07751)
Venkatraman S., Wilson J. R., 1986, Oper. Res. Lett., 5, 37
Villaescusa-Navarro F. et al., 2018, ApJ, 867, 137
Villaescusa-Navarro F. et al., 2020, ApJS, 250, 2
Vlah Z., White M., Aviles A., 2015, JCAP, 09, 014
Warren M. S., 2013, preprint (arXiv:1310.4502)
Yahya S., Bull P., Santos M. G., Silva M., Maartens R., Okouma P., Bassett B., 2015, MNRAS, 450, 2251
Zhai Z. et al., 2019, ApJ, 874, 95

APPENDIX A: ANALYTICAL DERIVATION: A BAYESIAN APPROACH

There is an elegant Bayesian derivation of the optimal form of the control variates estimator for the Gaussian case. The result coincides with the minimum variance estimator even in the non-Gaussian case. As in the derivation by Rubinstein & Marcus (1985), the covariance matrices of the full simulations $\boldsymbol{y}$ and of the fast simulations $\boldsymbol{c}$ are assumed to be known. In the main text, we use non-parametric approaches to estimate uncertainties since $\boldsymbol{\beta}$ is not known *a priori* but estimated from the same simulations that we use to estimate $\boldsymbol{\mu}_y$.

For notational simplicity, we will use $\boldsymbol{y}$ for the empirical mean of the brute-force simulations, $\boldsymbol{c}$ for the empirical mean of cheap simulations, and $\boldsymbol{t}$ for the target, the unknown mean of $\boldsymbol{y}$. These quantities can be related in a linear model,

$$\boldsymbol{y} = \boldsymbol{t} + \boldsymbol{\epsilon}_y, \tag{A1}$$

$$\boldsymbol{c} = \boldsymbol{m} + \boldsymbol{\epsilon}_c. \tag{A2}$$

We model the quantities on the right-hand side as

$$\boldsymbol{t} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{tt}), \tag{A3}$$

$$\boldsymbol{\epsilon}_y \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma}_{yy}/N), \tag{A4}$$

$$\boldsymbol{\epsilon}_c \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma}_{cc}/N), \tag{A5}$$

$$\boldsymbol{m} \sim N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_{mm}), \tag{A6}$$

which express, respectively, any prior information on $\boldsymbol{t}$ from previous runs, the noise terms for $\boldsymbol{y}$ and $\boldsymbol{c}$ after averaging over N simulations, and prior information on $\boldsymbol{m}$ from a separate run of fast simulations of $\boldsymbol{c}$. In addition, the basis of our methods is to exploit correlation between the Monte Carlo noise $\boldsymbol{y}$ and $\boldsymbol{c}$, so $\mathrm{cov}(\boldsymbol{y}, \boldsymbol{c}) \equiv \boldsymbol{\Sigma}_{yc}/N$.

Gathering these together in a single vector gives $\boldsymbol{z} = (\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{m})^T$ with multivariate normal density $p(\boldsymbol{z}) = p(\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{m})$. This joint vector $\boldsymbol{z}$ is a multivariate Gaussian $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_c \\ \boldsymbol{\mu}_c \end{pmatrix} \tag{A7}$$

and

$$C = \begin{pmatrix} \boldsymbol{\Sigma}_{tt} & \boldsymbol{\Sigma}_{tt} & \boldsymbol{0}_{p,p} & \boldsymbol{0}_{p,p} \\ \boldsymbol{\Sigma}_{tt} & \boldsymbol{\Sigma}_{tt} + \boldsymbol{\Sigma}_{yy}/N & \boldsymbol{\Sigma}_{yc}/N & \boldsymbol{0}_{p,p} \\ \boldsymbol{0}_{p,p} & \boldsymbol{\Sigma}_{yc}{}^T/N & \boldsymbol{\Sigma}_{mm} + \boldsymbol{\Sigma}_{cc}/N & \boldsymbol{\Sigma}_{mm} \\ \boldsymbol{0}_{p,p} & \boldsymbol{0}_{p,p} & \boldsymbol{\Sigma}_{mm} & \boldsymbol{\Sigma}_{mm} \end{pmatrix}. \tag{A8}$$

The diagonal covariances are the block marginals, representing prior information; e.g. $\boldsymbol{\Sigma}_{mm}$ expresses the uncertainty in $\boldsymbol{m}$ obtained from a prior, independent simulation set of the fast surrogate. For that reason $\boldsymbol{\Sigma}_{ym} = \boldsymbol{\Sigma}_{tm} = \boldsymbol{\Sigma}_{tc} = \boldsymbol{0}_{p,p}$.

We are interested in the posterior $p(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{c})$; this expresses the information we have about our target $\boldsymbol{t}$ when we have obtained the set of correlated sample pairs $(\boldsymbol{y}, \boldsymbol{c})$. Based on our assumptions, we know the posterior $p(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{c})$ to be Gaussian with mean

$$\begin{aligned} \boldsymbol{\mu}_{t|y,c} &= \boldsymbol{\mu}_y + (\boldsymbol{\Sigma}_{tt} \quad \boldsymbol{0}_{p,p}) \begin{pmatrix} \boldsymbol{\Sigma}_{tt} + \boldsymbol{\Sigma}_{yy}/N\boldsymbol{\Sigma}_{yc}/N \\ \boldsymbol{\Sigma}_{yc}{}^T/N\boldsymbol{\Sigma}_{mm} + \boldsymbol{\Sigma}_{cc}/N \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y} - \boldsymbol{\mu}_y \\ \boldsymbol{c} - \boldsymbol{\mu}_c \end{pmatrix} \\ &= \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{tt}[\boldsymbol{\Sigma}_{tt} + \frac{1}{N}(\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yc}(N\boldsymbol{\Sigma}_{mm} + \boldsymbol{\Sigma}_{cc})^{-1}\boldsymbol{\Sigma}_{yc}{}^T)]^{-1} \\ &\quad ((\boldsymbol{y} - \boldsymbol{\mu}_y) - \boldsymbol{\Sigma}_{yc}(N\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{cc})^{-1}(\boldsymbol{c} - \boldsymbol{\mu}_c)) \end{aligned} \tag{A9}$$

and covariance

$$\begin{aligned} \boldsymbol{\Sigma}_{t|y,c} &= \boldsymbol{\Sigma}_{tt} - (\boldsymbol{\Sigma}_{tt} \quad \boldsymbol{0}_{p,p}) \begin{pmatrix} \boldsymbol{\Sigma}_{tt} + \boldsymbol{\Sigma}_{yy}/N\boldsymbol{\Sigma}_{yc}/N \\ \boldsymbol{\Sigma}_{yc}{}^T/N\boldsymbol{\Sigma}_{mm} + \boldsymbol{\Sigma}_{cc}/N \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{tt} \\ \boldsymbol{0}_{p,p} \end{pmatrix} \\ &= \boldsymbol{\Sigma}_{tt} - \boldsymbol{\Sigma}_{tt}[\boldsymbol{\Sigma}_{tt} + \frac{1}{N}(\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yc}(N\boldsymbol{\Sigma}_{mm} + \boldsymbol{\Sigma}_{cc})^{-1}\boldsymbol{\Sigma}_{yc}{}^T)]^{-1}\boldsymbol{\Sigma}_{tt} \\ &= [\boldsymbol{\Sigma}_{tt}{}^{-1} + N(\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yc}(N\boldsymbol{\Sigma}_{mm} + \boldsymbol{\Sigma}_{cc})^{-1}\boldsymbol{\Sigma}_{yc}{}^T)^{-1}]^{-1}. \end{aligned} \tag{A10}$$

These results generalize the earlier equations by (i) including $\boldsymbol{\Sigma}_{mm}$, the error estimate of $\boldsymbol{\mu}_c$ from a prior run of fast simulations, (ii) allowing for information from previous runs to be included by specifying prior mean $\boldsymbol{\mu}_t$ and prior covariance $\boldsymbol{\Sigma}_{tt}$, and (iii) giving analytical uncertainty estimates for the accelerated estimates.

To make contact with equation (15), we will consider special cases of this expression. Without prior information on $\boldsymbol{\mu}_y$ (i.e. $\boldsymbol{\Sigma}_{tt} \to \infty$) we obtain

$$\boldsymbol{\mu}_{t|y,c} = \boldsymbol{y} - \frac{1}{N}\boldsymbol{\Sigma}_{yc}(\boldsymbol{\Sigma}_{mm} + \frac{1}{N}\boldsymbol{\Sigma}_{cc})^{-1}(\boldsymbol{c} - \boldsymbol{\mu}_c) \quad \text{(no prior on } \boldsymbol{y}) \tag{A11}$$

and

$$\boldsymbol{\Sigma}_{t|y,c} = +\frac{1}{N}(\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yc}(N\boldsymbol{\Sigma}_{mm} + \boldsymbol{\Sigma}_{cc})^{-1}\boldsymbol{\Sigma}_{yc}{}^T) \quad \text{(no prior on } \boldsymbol{y}). \tag{A12}$$

For the case where the error on $\boldsymbol{m}$ can be neglected (i.e. $\boldsymbol{\Sigma}_{mm} \to \boldsymbol{0}_{p,p}$) but prior information is included, we obtain

$$\begin{aligned} &\boldsymbol{\mu}_{t|y,c,\boldsymbol{\Sigma}_{mm}\to\boldsymbol{0}_p} \\ &\quad = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{tt}(\boldsymbol{\Sigma}_{tt} + \boldsymbol{\Sigma}_{t|y,c,\boldsymbol{\Sigma}_{mm}\to\boldsymbol{0}_{p,p}})^{-1} \\ &\qquad ((\boldsymbol{y} - \boldsymbol{\mu}_y) - \boldsymbol{\Sigma}_{yc}\boldsymbol{\Sigma}_{cc}{}^{-1}(\boldsymbol{c} - \boldsymbol{\mu}_c)) \quad (\boldsymbol{\mu}_c \text{ known}) \end{aligned} \tag{A13}$$

and

$$\begin{aligned} &\boldsymbol{\Sigma}_{t|y,c,\boldsymbol{\Sigma}_m\to\boldsymbol{0}_{p,p}} \\ &\quad = [\boldsymbol{\Sigma}_{tt}{}^{-1} + N(\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yc}\boldsymbol{\Sigma}_c{}^{-1}\boldsymbol{\Sigma}_{yc}{}^T)^{-1}]^{-1} \quad (\boldsymbol{\mu}_c \text{ known}). \end{aligned} \tag{A14}$$

In the absence of prior information and assuming that $\boldsymbol{\mu}_c$ is perfectly known (i.e. $\boldsymbol{\Sigma}_{tt} \to \infty$ and $\boldsymbol{\Sigma}_{mm} \to \boldsymbol{0}_{p,p}$), equation (A10) simplifies to match the result of equation (15) from Rubinstein & Marcus (1985),

$$\boldsymbol{\mu}_{t|y,c} = \boldsymbol{y} - \boldsymbol{\Sigma}_{yc}\boldsymbol{\Sigma}_{cc}{}^{-1}(\boldsymbol{c} - \boldsymbol{\mu}_c)) \quad (\boldsymbol{\mu}_c \text{ known, no prior on } \boldsymbol{y}) \tag{A15}$$

and

$$\boldsymbol{\Sigma}_{t|y,c} = \frac{1}{N}(\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yc}(\boldsymbol{\Sigma}_{cc})^{-1}\boldsymbol{\Sigma}_{yc}{}^T) \quad (\boldsymbol{\mu}_c \text{ known, no prior on } \boldsymbol{y}). \tag{A16}$$

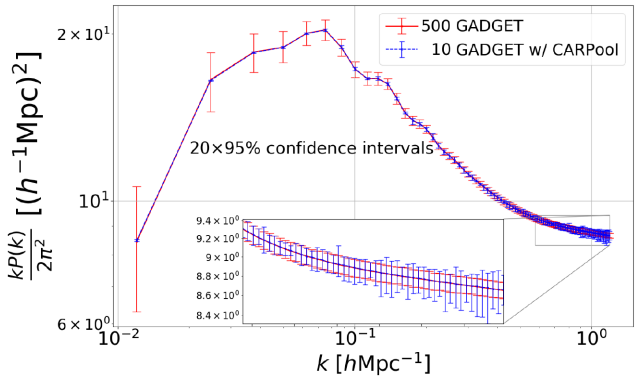


Figure B1. As in Fig. 2, but with 10 *N*-body simulations used for the CARPool estimate.

APPENDIX B: ADDITIONAL INSIGHT ON RESULTS AND CONFIDENCE INTERVALS

We start with a reminder about confidence intervals. The '1σ rule of thumb' (same for two and three) is a direct application of the *central limit theorem* (CLT) when estimating a random variable with the sample mean of N realizations,

$$\bar{y} \pm \gamma \frac{\widehat{\sigma_y}}{\sqrt{N}}, \tag{B1}$$

where γ is the z-score – e.g. from a normal distribution – associated to a given confidence band. The 95 per cent symmetric confidence intervals correspond to $\gamma \approx 1.96$, hence the name 2σ rule.' With a very small number of samples, the CLT is not really 'working,' so it is common practice to penalize the confidence intervals by taking

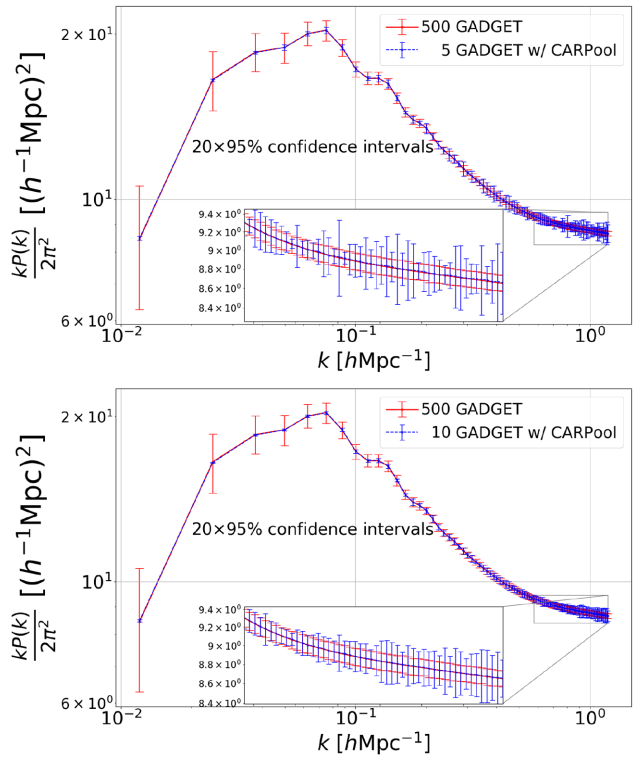


Figure B2. The upper panel shows the same data as in Fig. 2 and the lower panel is paired with Fig. B1, except that the confidence intervals come from *t*-score values with 4 and 9 d.o.f., respectively.

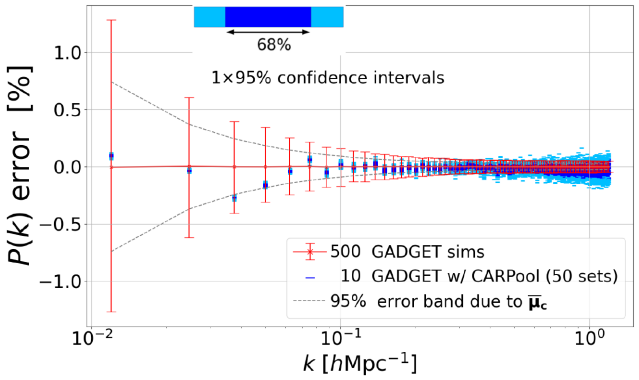


Figure B3. As in the lower panel of Fig. 3, but with 10 *N*-body simulations used for the CARPool estimate.

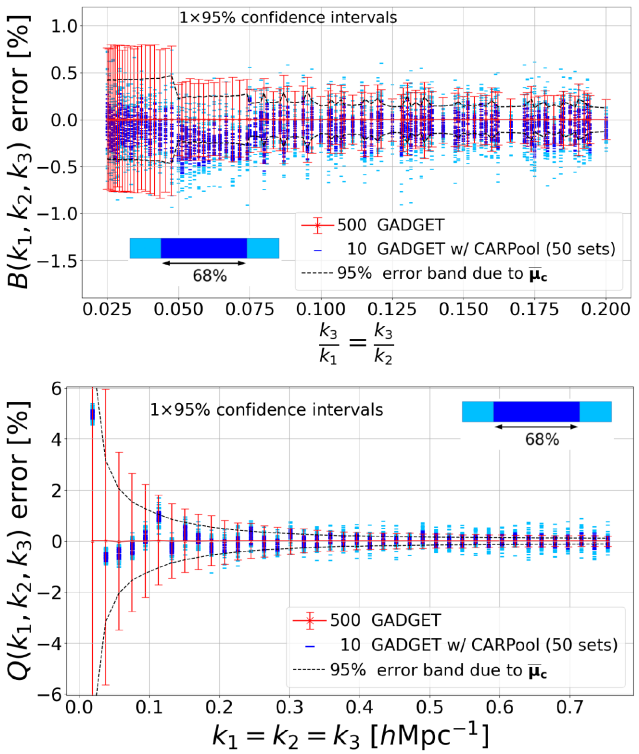


Figure B4. As in Fig. 7, but with 10 *N*-body simulations used for the CARPool estimate.

γ from a *t*-score table i.e. from a Student distribution with $N - 1$ degrees of freedom, which has fatter tails. For instance, for $N = 10$, $\gamma \approx 2.26$ for the 95 per cent confidence band.

Because the trustworthiness of confidence intervals for a sample mean with very few realizations is debatable, we provide here, by way of an example for the power spectrum only, Fig. B1 with bootstrap confidence intervals of 10 CARPool samples and Fig. B2 for CARPool with 5 and 10 *N*-body simulations but with *t*-score intervals accordingly to equation (B1). The latter figure is to compare with Figs 2 and B1 (exact same data except for the blue CARPool confidence intervals). We have agreement between the paired plots, and we notice that the symmetric confidence intervals from *t*-score tend to be larger. Additionally, for the two- and three-point clustering statistics, we present in Figs B3 (power spectrum) and

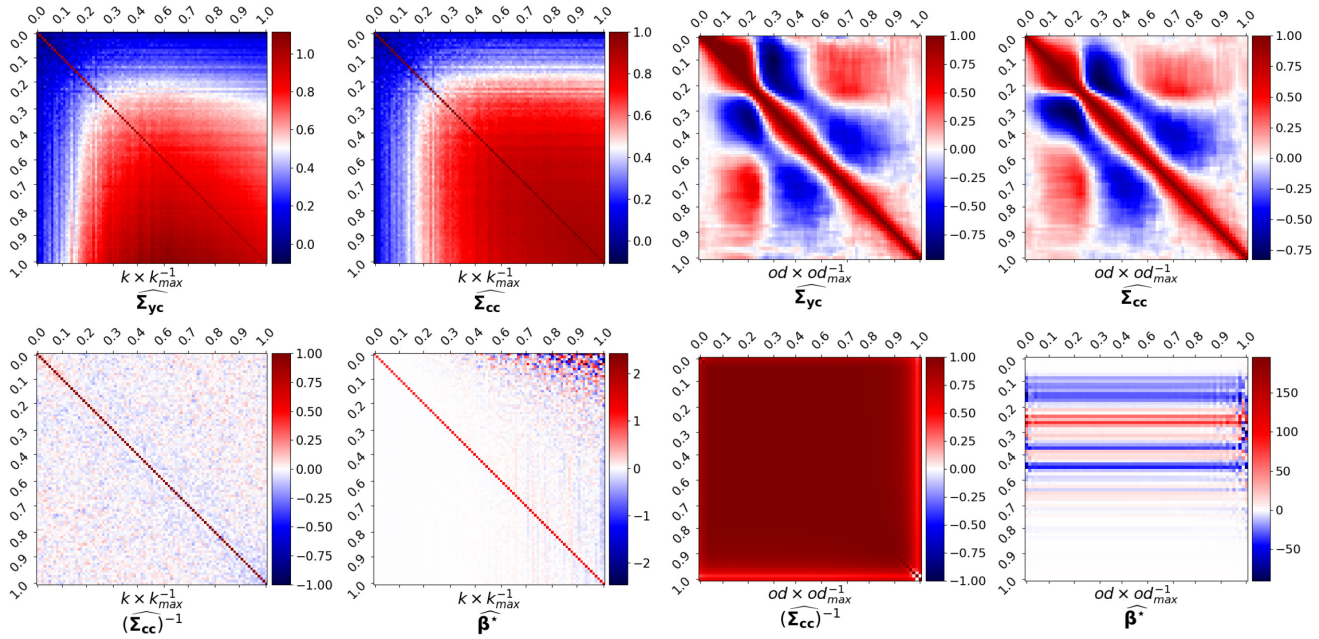


Figure B5. Estimated matrices intervening in $\boldsymbol{\beta}^{\star}$ for the matter power spectrum (left) and the matter PDF (right). The cross-covariance, covariance, and precision matrices are normalized i.e. we display $\boldsymbol{D}^{-1}\widehat{\boldsymbol{\Sigma}}\boldsymbol{D}^{-1}$ with $\boldsymbol{D} = \sqrt{\mathrm{diag}\left(\widehat{\boldsymbol{\Sigma}}\right)}$. 'od' denotes the fractional overdensity bin $\rho/\bar{\rho}$. For better visibility, the diverging colour scale is not forced to be centered at 0.0 for the $\boldsymbol{\Sigma}_{yc}$ and $\boldsymbol{\Sigma}_{cc}$ estimates in the upper left corner (power spectrum). All matrices are estimated using 500 simulation pairs, and represent the 'close to optimal' $\boldsymbol{\beta}^{\star}$ towards which the control matrix estimator tends in the multivariate setting.

B4 (bispectrum) the percentage error of CARPool means with 10 simulations that are not shown in the main part of the paper.

We provide also in Fig. B5 an overview of the optimal control matrix $\boldsymbol{\beta}^{\star}$ from equation (8) for the matter power spectrum and matter PDF test cases.

APPENDIX C: COLA TIMESTEPPING AND CROSS-CORRELATION COEFFICIENTS

We briefly explain our choice of timestepping strategy to generate a collection of low-fidelity snapshots at $z = 0.5$. In COLA, the cosmological scale factor a is used to discretize the time derivative of the left-hand side of the COLA equation of motion (17),

$$\begin{aligned}\boldsymbol{v}_{i+\frac{1}{2}} &= \boldsymbol{v}_{i-\frac{1}{2}} - \Delta a_1 \partial_a^2 \boldsymbol{\Psi}_{\mathbf{res}},\\ \boldsymbol{r}_{i+\frac{1}{2}} &= \boldsymbol{r}_i + \boldsymbol{v}_{i+\frac{1}{2}}\Delta a_2 + \Delta D_1 \boldsymbol{\Psi}_{\mathbf{1}} + \Delta D_2 \boldsymbol{\Psi}_{\mathbf{2}}.\end{aligned} \tag{C1}$$

$\Delta D_l = D_{l,i+1} - D_{l,i}$ with $l \in \{1, 2\}$ are the changes of linear (or Zel'dovich) and second-order growth factors between the timesteps, normalised such that $D_1(a = 1) = D_2(a = 1) = 1$. $\boldsymbol{\Psi}_{\mathbf{1}}$ and $\boldsymbol{\Psi}_{\mathbf{2}}$ are, respectively, the linear (or Zel'dovich) and second-order LPT (or 2LPT) displacement fields at $a = 1$. We have enabled the timestepping scheme from Tassev et al. (2013) in which the time intervals Δa_i, $i \in \{1, 2\}$ are given by

$$\begin{aligned}\Delta a_1 &= \frac{H_0}{nLPT}\frac{a_{i+\frac{1}{2}}^{nLPT} - a_{i-\frac{1}{2}}^{nLPT}}{a_i^{nLPT-1}},\\ \Delta a_2 &= \frac{H_0}{a_{i+\frac{1}{2}}^{nLPT}}\int_{a_i}^{a_{i+1}} \frac{a^{nLPT-3}}{H(a)} da.\end{aligned} \tag{C2}$$

Here, *nLPT* is an additional free parameter that should be tuned experimentally for every simulation setting, as Tassev et al. (2013), Howlett et al. (2015), and Izard, Crocce & Fosalba (2016) already stressed. The Kick-and-Drift/Leapfrog algorithm of Quinn et al. (1997) can also be used in L-PICOLA.

Before generating our ensemble of fast surrogates, we tested the sensitivity of the cross-correlation coefficients ζ_{yc} between the full N-body dark matter density contrast field $\boldsymbol{\delta}_{\boldsymbol{y}}$ and $\boldsymbol{\delta}_{\boldsymbol{c}}$ produced by L-PICOLA,

$$\zeta_{yc} = \frac{\mathbb{E}\left[\delta_y(\boldsymbol{k})\delta_c(\boldsymbol{k})^*\right]}{\sqrt{\mathbb{E}\left[\mid \delta_y(\boldsymbol{k}) \mid^2\right]\mathbb{E}\left[\mid \delta_c(\boldsymbol{k}) \mid^2\right]}} = \frac{P_{yc}(\boldsymbol{k})}{\sqrt{P_y(\boldsymbol{k})P_c(\boldsymbol{k})}}, \tag{C3}$$

to the choice of timestepping.

The numerator in (C3) is the cross power spectrum between the two aforementioned density contrast fields. $\delta(\boldsymbol{k})$ is the Fourier transform of the real-space density contrast $\delta(\boldsymbol{x})$. Note that these coefficients serve as a proxy for the correlation between the COLA and GADGET snapshots, but do not provide an estimation of the canonical cross-correlations of (10) between the statistics $\boldsymbol{y}$ and $\boldsymbol{c}$ computed from these snapshots. Having tested different schemes, we concluded that choosing linearly-spaced timesteps yields a better cross-correlation than with logarithmic ones, and that the fewer the timesteps, the more influential the modified timestepping parameter *nLPT* in terms of cross-correlation coefficients (in the case of this study, with a very high starting redshift of $z_i = 127$). Fig. C1 shows an example with 10 and 20 linearly-spaced timesteps and $nLPT \in \{-2.5, +0.5\}$ (the fiducial value and our experimentally 'best' value, respectively). Although $\zeta_{yc}(k = 1.0\ h\mathrm{Mpc}^{-1}) \approx 0.96$ with 10 timesteps exceeds $\zeta_{yc}(k = 1.0\ h\mathrm{Mpc}^{-1}) \approx 0.94$ with 20 timesteps for $nLPT = +0.5$, we still chose to generate our L-PICOLA snapshots with 20 timesteps between $z_i = 127$ and $z = 0.0$, again, to avoid tuning L-PICOLA for any one particular statistic. In any case, even with 20 timesteps the L-PICOLA surrogates are much faster than full GADGET-III simulations.

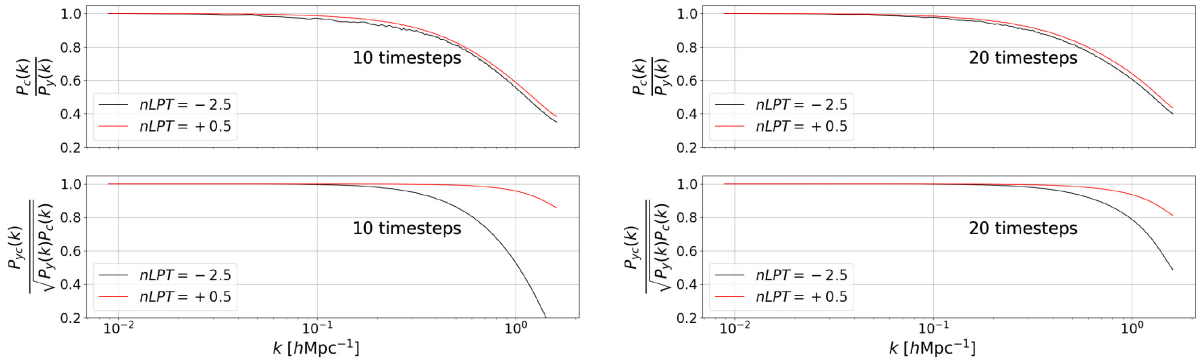


Figure C1. Power spectrum recovery ratio (top) and cross power spectrum coefficients (bottom) at $z = 0.5$ between a specific L-PICOLA snapshot computed with 10 (left) and 20 (right) linearly-spaced timesteps and the corresponding N-body snapshot derived from the same initial conditions at $z_i = 127$.

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