Covmos: A new Monte Carlo approach for galaxy clustering analysis

Philippe Baratta1, Julien Bel2, Sylvain Gouyou Beauchamps1,3,4, and Carmelita Carbone5

1 Aix Marseille Université, CNRS/IN2P3, CPPM, IPhU, Marseille, France
e-mail: baratta@cppm.in2p3.fr
2 Aix Marseille Univ, Université de Toulon, CNRS, CPT, Marseille, France
3 Institut d’Estudis Espacials de Catalunya (IEEC), Carrer Gran Capitá 2-4, 08034 Barcelona, Spain
4 Istituto di Astrofisica Spaziale e Fisica cosmica Milano, Via A. Corti 12, 20133 Milano, Italy

Received 13 December 2022 / Accepted 9 March 2023

ABSTRACT

We validate the Covmos method introduced in our previous study allowing for the fast simulation of catalogues of different cosmological field tracers (e.g. dark matter particles, halos, galaxies, etc.). The power spectrum and one-point probability distribution function of the underlying tracer’s density field are set as inputs of the method and arbitrarily chosen by the user. To evaluate the validity domain of Covmos at the level of the produced two-point statistics covariance matrix, we chose to target these two input statistical quantities from realistic N-body simulation outputs. We performed this cloning procedure based on the ACDM as well as on a massive neutrino cosmology, for five redshifts in the range of $z \in [0, 2]$. First, we validated the output real-space two-point statistics (both in the configuration and Fourier space) and estimated over 5000 Covmos realisations per redshift and per cosmology, with a volume of $1$ Gpc$^3$ and $10^5$ particles each. This validation was performed against the corresponding N-body measurements, estimated from 50 simulations. We found the method to be valid up to $k \sim 0.26$ Mpc$^{-1}$ for the power spectrum and down to $r \sim 20$ Mpc$^{-1}$ for the correlation function. Then, we extended the method by proposing a new modelling of the peculiar velocity distribution, aimed at reproducing the redshift-space distortions both in the linear and mildly non-linear regimes. After validating this prescription, we finally compared and validated the resulting redshift-space two-point statistics covariance matrices in the same range of scales. We released the Python code associated with this method on a public repository, which allows for the production of tens of thousands of realisations in record time. In particular, Covmos is intended for any study involving large-scale galaxy-survey science that requires a large number of mock realisations.

Key words. large-scale structure of Universe – catalogs

1. Introduction

Understanding the dark sector of the universe constitutes the main challenge of future large galaxy surveys such as Euclid (Laureijs et al. 2011), the Legacy Survey of Space and Time (LSST, LSST Science Collaboration 2009), or the Dark Energy Spectroscopic Instrument (DESI, DESI Collaboration 2016). These surveys have been designed to provide a large amount of data, with the ability to disentangle and constrain several cosmological parameters with a high level of accuracy. Such parameters include the dark energy equation of state and the neutrino mass, as well as more alternative cosmological scenarios, through galaxy clustering, weak lensing, and their cross-correlations.

However, even with infinitely precise and voluminous data, discriminating among different cosmological models is not guaranteed, as it also requires the development of specific analysis methods. The complexity of galaxy survey data analyses mainly relies on the fact that we ought to combine several types of correlated cosmological probes and identify numerous systematic errors and biases. Thus, such analyses must lay on robust statistical tools including a good estimation of correlations in the data and expected error bars. In particular, the cosmological model extraction is based on the evaluation of a likelihood, requiring an accurate covariance matrix for the considered observables (Taylor et al. 2013).

Although predicting the galaxy clustering covariance in a fully analytical way (Wadkar & Scoccimarro 2020; Lacasa 2018; Krause & Eifler 2016) is possible, when one needs to cross-correlate various observables, alternatives to analytical covariances must be developed. As a matter of fact, most of the covariance matrices exploited in galaxy survey analysis today rely on large samples of cosmological simulations (Kitaura et al. 2016; Zhao et al. 2021) and even when an analytical covariance can be used, such as in the recent DES Y3 results (Friedrich et al. 2021), this has to be validated against a covariance sampled from accurate numerical simulations.

In addition to the fact that the construction of such large samples of mock catalogues is highly CPU and time consuming, the major problem is that the covariance is model-dependent: at least one covariance matrix per tested cosmological model is needed (when assuming a covariance matrix constant with respect to the parameters to infer).

Different approaches have been investigated to bypass these issues, called approximate methods. A category of them is designed to speed up the N-body evolution by either approximating the dynamics of the global dark matter field or by focusing on the evolution of early identified halos such as ICE-COLA (Tassev et al. 2013; Izard et al. 2018), PINOCCHIO (Monaco et al. 2002, 2013), or PEAK PATCH (Bond & Myers 1996). Although highly competitive with respect to N-Body simulations (at least up to four-point statistics, see...
Lippich et al. 2019; Blot et al. 2019; Colavincenzo et al. 2019),
they still remain CPU and time-demanding. Other types of meth-
ods, such as PATCHY (Kitaura et al. 2013, 2015) or HALOGEN
(Avila et al. 2015) are faster than the above methods since they
do not require any field evolution. They directly produce the
dark matter field at the considered redshift, before applying a bias-
ning scheme (using second order Lagrangian perturbation theory
fitted to \(N\)-body products to obtain halo catalogues. However,
these methods require a certain level of calibration and there-
fore remain indirectly expensive to adapt from one cosmological
model to another.

Finally, a third category of approximate methods has already
been studied in the specific case of log-normal density fields
(e.g. Coles & Jones 1991; Greiner & Enßlin 2015). These meth-
ods consist in targeting, at a given redshift, both the clustering
amplitude, usually through the power spectrum, and the proba-
bility distribution function (PDF) of the simulated density field.
A discretisation of this field, based on the so-called local Poisson
process approximation, can eventually be applied to obtain a dis-
crete set of objects. Because of the rapidity and the simplicity of
implementation of such methods, a vast amount of literature has
investigated the log-normal distributions (e.g. Xavier et al. 2016;
Alonso et al. 2014; Agrawal et al. 2017). However, it has been
shown by Carron (2011) that the log-normal statistics poorly
reproduce the true data correlations when entering the non-linear
regime, due to the limited range of validity of such approximated
PDF forms (see Tosone et al. 2020, for an extension).

In this context, we introduced in Baratta et al. (2020) a new
Monte Carlo method fully suited for simulating catalogues of
objects defined by an arbitrary power spectrum and a one-point
PDF. In the present work, we aim at validating this method
against the DEM Hunl (Carbone et al. 2016; Castorina et al. 2015;
Parimbelli et al. 2022) set of \(N\)-body simulations. We thus evalu-
ate the validity domain of the two-point statistic covariance
matrices produced with the COV\`{M}OS\textsuperscript{1} public code.

This paper is organised as follows. In Sect. 2, we detail the
core of the method and generalise it to the simulation of power
spectra with non-linear prescriptions. In Sect. 3, after introduc-
ing the DEM Hunl \(N\)-body simulations, we compare and discuss
the results for the produce two-point statistics (the power spec-
trum and the two-point correlation function) in real space. We
then extend the Monte Carlo method by introducing our pecu-
liar velocity assignment procedure in Sect. 4, describing how
we reproduced redshift-space distortions effects both in the lin-
ear and mildly non-linear regimes. We compare the auto- and
cross-covariances in the redshift space for the multipoles of the
anisotropic power spectrum and two-point correlation function
with those of the \(N\)-body reference in Sect. 5. In Sect. 6 we
present our conclusions.

2. Theoretical pipeline in real space

In this section, we recall the basics of the pipeline that enables
the generation of a discrete set of objects defined by a given,
arbitrary one-point PDF and density power spectrum. For a more
detailed description of the method, we refer to Baratta et al.
(2020).

In the first stage, we focus on the generation of the continu-
ous density field. Then we discuss how to circumvent the limit
of the method, that is, the production of negative, non-physical
power spectra, when the target power spectrum is the non-linear

\[ \delta(x) = L(\nu(x)), \]

where the resulting \(\delta\)-field must represent the final cosmological,
non-Gaussian contrast density field. The transformation, \(L\), can
be found using standard probability transformation rules, as long
as a target \(\delta\)-PDF is set. In particular, using the local conservation
of probabilities between \(\nu\) and \(\delta\), \(L\) in Eq. (1) can directly be
identified as:

\[ \delta = C_\nu^{-1}[C_\delta(\nu)], \]

where \(C_\nu\) is defined as the cumulative function of the field \(Y\).
The results of Eq. (1) not only impact the moments of \(\nu\), but also
its two-point statistics. By consistently using probability conserv-
ation properties, we can quantify the effect of applying a local
transformation, \(L\), to the Gaussian field through the mapping, \(\lambda\),
via:

\[ \xi_L \equiv \langle \delta_1 \delta_2 \rangle = \int \nu_1 \nu_2 L(\nu_1)L(\nu_2)P_\nu(\nu_1, \nu_2, \xi_\delta) \equiv \Lambda(\xi_\delta), \]

where \(P_\nu(\nu_1, \nu_2, \xi_\delta)\) is the bivariate normal distribution of \(\nu_1\) and \(\nu_2\),
its corresponding two-point correlation function. \(\xi_L\) stands for the
two-point correlation function of the cosmological field.

Since the power spectrum, \(P(k)\), is the Fourier transform of
the two-point correlation function, we can rely on the mapping,
\(\lambda\), in order to predict the input power spectrum of the Gaussian
field, \(\nu\). Indeed, according to Appendix C of Bel et al. (2016),
with \(|\xi_L| < 1\), \(\lambda\) turns out to be continuous and monotonic in
this interval. This implies the existence of the reciprocal function
\(\lambda^{-1}\), ensuring that Eq. (3) is inverted as \(\xi_\delta = \lambda^{-1}(\xi_L)\).

As a matter of fact, once \(\xi_\delta\) is known (provided that \(P_\delta\) is the
target power spectrum), being able to compute \(\xi_\delta\), is equiva-
 lent to finding the power spectrum \(P_\delta\) to be assigned to the
initial Gaussian field. This specific field is then generated such
that, once transformed under the local mapping, \(L\), the result-
ing \(\delta\)-field follows the targeted power spectrum, \(P_\delta\). Formally
speaking, an overview of the pipeline is expressed as:

\[ P_\delta(k) = \mathcal{F} \left[ \lambda^{-1} \mathcal{F}^{-1} [P_\delta(k)] \right], \]

where \(\mathcal{F}\) and \(\mathcal{F}^{-1}\) stand respectively for the Fourier and inverse
Fourier transforms of the power spectra. The presented inversion
scheme relies on our ability to evaluate the transformation \(\lambda\) of
the two-point correlation function. Therefore, we need to num-
merically compute Eq. (3).\textsuperscript{2}

\textsuperscript{1} Publicly available in github.com/PhilippeBaratta/COV\`{M}OS

\textsuperscript{2} It can be shown that it is possible to switch the 2D integral of Eq. (3)
into a sum of finite 1D integrals, quickly converging and therefore less
computationally expensive. As a result, computing the Hermite expan-
sion of the local mapping, \(L\), allows us to write down the Taylor expan-
sion of the two-point correlation function, \(\xi_\delta\). For more details, we refer
to Appendix C of Bel et al. (2016).
2.2. Practical implementation: Constraints and limits

Here, we consider a periodic, cubical comoving volume of a size, \(L\). We assume that the different fields are sampled on a regularly spaced grid in this box, characterised by the sampling parameter \(N_s\), representing the number of grid nodes per dimension. In this setting, we must now consider a three-dimensional description of the various fields and two-point statistics. In particular, the quantities described in Sect. 2.1 become \(\nu(x) \rightarrow \nu^{(n)}(x)\), \(\delta(x) \rightarrow \delta^{(n)}(x)\), \(P(k) \rightarrow P^{(n)}(k)\), and so on.

Since generating a non-Gaussian density field in the configuration space requires us to start from an initial Gaussian field in Fourier space\(^3\), the Monte Carlo method extensively uses three-dimensional (3D) fast Fourier transforms (FFT) and their inverse (IFFT) in order to be efficient. As a result, we have to sample the generated fields on a regular grid in the configuration space, producing (in the Fourier space) a well known effect called aliasing (Jing 2005). Basically, the FFT of a sampled field is the standard Fourier transform of the field itself, to which replicas (or aliases) of it are added. This produces an extra power on scales close to the Nyquist frequency \(k_N = \pi/(L/N_s)\), also affecting the Fourier phases. Assuming that \(\text{IFFT} | \text{FFT}[Y] | = Y\) (or equivalently that the IFFT of an aliased field is the ‘right’ field in the configuration space), the targeted power spectrum, \(P^{(n)}(k)\), used in Eq. (4) must be artificially aliased according to\(^4\)

\[
P_{\text{aliased}}^{(n)}(k) = \sum_i P_i(|k - 2n\pi/k_N|),
\]

so that the method may be self-consistent. Here, \(n \equiv (n_1, n_2, n_3) \in \mathbb{Z}^3\). Simply stated, the set of fields we generate in Fourier space must necessarily be aliased. This ensures the fields in configuration space to be unaffected by aliasing.

Moreover, as the cosmological fields are sampled on a finite grid, this naturally introduces a smoothing scale of their statistics, related to the quantity \(L/N_s\). As a consequence, prior to the application of Eq. (4), the two statistical inputs (power spectrum and PDF) must be smoothed consistently with the grid. This is natural for the PDF if its targeted shape is directly estimated on the grid point (NGP) window; see Sefusatti et al. (2016).

\(\nu\) and \(\delta\) quantities described in Sect. 2.1 become \(\nu^{(n)}(x) \rightarrow \nu^{(n)}(x)\), \(\delta(x) \rightarrow \delta^{(n)}(x)\), \(P(k) \rightarrow P^{(n)}(k)\), and so on.

account or not. Such an outcome is at this stage not understood. As an example, with a high resolution grid of \(N_s = 1024\) and \(L = 1000\) Mpc \(h^{-1}\), corresponding to a precision of \(~1\) Mpc \(h^{-1}\), about \(-45\%\) of negative elements are produced for a CDM non-linear power spectrum at redshift \(z = 0\). They are mainly located in the Fourier volume in the vicinity of the Nyquist frequency. This number of negative values is reduced for \(N_s = 512\) (\(-20\%) and disappears for \(N_s = 256\). However, choosing such poor grid setting would inevitably reduce the range of usable simulated modes.

On the one hand, the variance, \(\sigma^2 = k_F^3 \sum_l P_l(k)\), estimated from the power spectrum of the Gaussian field (with \(k_F = 2\pi/l\) being the fundamental frequency of the periodic box) is actually fully unitary (as required by the procedure), while on the other hand, these non-physical negative elements prevent the simulation of the Gaussian field, \(\nu_k \sim \sqrt{P(k)}\). To sidestep this restriction, a corrective method can be adopted: this consists of clipping to zero all negative values. As a direct consequence, the variance computed on the non-Gaussian power spectrum automatically deviates from unity with \(\sigma^2 > 1\). In order to recover the procedure condition, \(\sigma^2 = 1\), an additional 3D filtering can be applied to match the unit variance. Here, we propose the 3D filtering version of Eq. (6), that is:

\[
F(k) = \exp\left[-(kR_2)^2\right].
\]

Naturally, this procedure has the effect of reducing the maximum mode up to which the power spectrum is reliably simulated. While \(R_1\) and \(R_2\) are constrained by the previously discussed conditions on the variance, \(i\) and \(j\) should be chosen in such a way that the simulated power spectrum matches the targeted one up to a maximum wave-mode. Their values are discussed in Sect. 3.

We go on to briefly summarise the method to provide a global understanding. We generated, in Fourier space, a Gaussian field, \(\nu\), on a 3D grid defined by the power spectrum \(P^{(n)}(k)\) (Eq. (4)), after being corrected using Eq. (7). We then perform an inverse FFT of \(\nu^{(n)}(x)\) to get \(\nu(x)\) and locally applied the non-linear transform, \(L\), to it, as in Eq. (1). The obtained field, \(\delta(x)\), is now expected to both follow the target density PDF (the exact one) and (filtered) power spectrum.

2.3. From continuous field to discrete catalogues

In order to finalise the procedure above, we can turn the obtained density field, \(\delta(x)\), into a discrete catalogue of objects by applying a local Poisson sampling (Lazyer 1996) of it. Provided a mean number density, \(\rho_0\), expected in a catalogue of volume, \(V\), and a simulated local density contrast field, \(\delta(x)\), we can estimate the number of objects, \(N\), which is locally expected within a small volume element, \(\Delta\nu = (L/N_s)^3\), as \(N = \rho_0 \Delta \nu (1 + \delta)\).

Of course, this expected number \(N\) is not an integer. However, it can be seen as the expectation value that we should see if the sampling was made several times at a position \(x\). Thus, the number of observed objects in the small volume, \(\Delta\nu\), can be drawn from a Poisson distribution with an expectation value \(N\).

We can think about two kinds of interpolation schemes when placing the corresponding number of particles within the grid cell. The first is the so-called top hat scheme, where particle coordinates follow a uniform distribution. It is fast, but the density fields will appear structurally discontinuous. A natural extension of this method that is able to fix this issue is the tri-linear scheme, implemented here by populating the grid cell with particles following the tri-linearly interpolated density field distribution. In Appendix A, we detail the method that allows for

\(^3\) Generating a Gaussian field with a given power spectrum is greatly simplified in the Fourier space.
\(^4\) We consider here an aliasing correction in the context of a nearest grid point (NGP) window; see Sefusatti et al. (2016).
such assignment of particles in voxels. We note that both methods have an impact on the shape of the resulting Poisson power spectrum, which fortunately can be accurately predicted even beyond the Nyquist frequency (see Baratta et al. 2020).

To anticipate our outcomes, we find that the choice of the interpolation scheme has no impact on the estimated covariance matrix in the considered wave-mode range. Nonetheless, in the following, we only use the three-linear interpolation scheme.

2.4. Discussion

The method can be viewed as a cloning procedure of a power spectrum and a PDF obtained from an original reference catalogue. For the method to be optimal, the ideal choice for the MAS used to estimate the PDF must be compatible with the type of interpolation scheme used to assign coordinates to particles. For example, in the case of the tri-linear interpolation process, a cloud in cell (CIC) MAS should have been used to estimate the PDF. This would allow for a better reproduction of the 1-p PDF at the end of the chain. However, for technical reasons, it is often impossible to estimate this kind of PDF without it being heavily affected by noise. It is therefore generally more convenient to resort to a piecewise cubic spline (PCS) estimated PDF, which is less affected by noise. COVMOS not proposing the equivalent of the PCS to place particles, the 1-p PDF of the cloned catalogue cannot be expected to be identical to the original reference simulation and will seem over-smoothed. However, such a discrepancy is materialised for scales smaller than the grid size, affecting only the scales beyond our scales of interest.

Moreover, before going through the validation of the pipeline in Sect. 3, it must be mentioned the fact that the clipping procedure of the negative elements of the power spectrum introduced in Eq. (7) has some key statistical consequences. First, there is the fact that the produced mean density, \( \bar{\rho} \), is not rigorously zero but may systematically be shifted to a level of \( \sim 10^{-3} \). This excess of mean density cannot simply be subtracted since it correlates with the associated realised power spectra at all \( k \) modes. A non-trivial consequence of this additional correlation is the biasing of the two-point statistics covariances. Appendix B shows in details how this bias can be identified and subtracted. We note that all results involving covariances in the following of this article has been corrected for this bias.

3. Application in real space

The method introduced above allows us to simulate, within a cubical volume, objects whose nature (dark matter particles, dark matter halos, galaxies, etc.) only depends on the choice of the input density power spectrum and PDF. In this section, for simplicity, we simulate dark matter particles in real space when targeting the statistics of a set of \( N \)-body simulations. We evaluate the range of reliability of the produced two-point statistics for a reasonable (computationally speaking) grid sampling precision of \( N = 1024 \).

3.1. DEMNUni\_cov N-body simulations

Originally developed for testing different probes in the presence of massive neutrinos and dynamical dark energy, the Dark Energy and Massive Neutrino Universe (DEMNUni\_6; Carbone et al. 2016) experiment is a suite of \( N \)-body simulations (Roncarelli et al. 2015; Castorina et al. 2015; Moresco et al. 2017; Fabbian et al. 2018; Zennaro et al. 2018; Ruggeri et al. 2018; Bel et al. 2019; Kreisch et al. 2019; Schuster et al. 2019; Verza et al. 2019, 2022b,a; Paribelli et al. 2021, 2022; Guidi et al. 2022; Hernández-Moliner et al. 2023) in ACDM, plus different \( \nu W_0 \nu_0 \) CDM cosmological scenarios. The largest simulations follow the evolution of 2048\(^3\) CDM and (if present) 2048\(^3\) neutrino particles, in a box of size \( L = 2 \text{ Gpc} \ h^{-1} \). They have been produced on various high-performance computing (HPC) machines provided by the super computing centre CINECA, Italy, by exploiting the GADGET-3 code of Springel (2005), further developed by Viel et al. (2010) for the inclusion of massive neutrino particles.

All the simulations of the DEMNUni have in common the following cosmological parameters:

\[
\Omega_k = 0, \quad \Omega_m = 0.32, \quad \Omega_b = 0.05, \quad h = 0.67, \quad n_s = 0.96, \quad A_s = 2.1265 \times 10^{-9}.
\]

The initial conditions have been set at redshift \( z = 99 \) using the Zeldovich approximation (Zeldovich 1970) and the public code described in Zennaro et al. (2016, 2018, 2019).

The evolution of cosmic structures is simulated down to redshift \( z = 0 \), while outputting 63 particle snapshots with equally spaced logarithmic scale factor intervals. Among these snapshots, we restrict the analysis of this work to the five following redshifts:

\[ z = [0, 0.48551, 1.05352, 1.45825, 2.05053]. \]

Regarding our study oriented toward the estimation of covariances, only the DEMNUni subset, hereafter called DEMNUni\_cov Paribelli et al. (2021), constitutes the reference basis for the validation tests of our Monte-Carlo method. This subset is made of two ACDM cosmologies, one without massive neutrinos and the other one with massive neutrinos (added at the level of massive particles inside the simulation) of a total mass, \( M_\nu = 0.16 \text{ eV} \). In the following, these two cosmologies are referred to as ‘ACDM’ and ‘\( \nu \nu \)’, respectively. The main objective in using these two suites of simulations is that they are run 50 times each with different realisations of the initial conditions. In this way, they offer the possibility of estimating the covariance matrices to some extent.

Each simulation consists of 1024\(^3\) dark matter particles of mass \( m_p = 8 \times 10^{10} M_\odot h^{-1} \), and 1024\(^3\) massive neutrino particles (in the case of the massive neutrino cosmology) of mass \( m_\nu = 1 \times 10^3 M_\odot h^{-1} \) in a cubical volume of size \( L = 1000 \text{ Mpc} \ h^{-1} \) with periodic boundary conditions. For each snapshot of each realisation, we estimated the PDF and the power spectrum of the matter field using a PCS MAS on a grid of sampling parameter \( N = 1024 \). The power spectrum was estimated using.

---

5 Note that a the target power spectrum smoothed using \( i = 2.5 \) in Eq. (6) matches the PCS window.

6 https://www.researchgate.net/project/DEMUN-Universe-DEMNUni

7 cineca.it

8 For simplification in the following, these five redshifts will be referred to the round numbers \( z = [0, 0.5, 1, 1.5, 2] \).
3.2. Comparing the two-point statistics

In the first place, for the Monte Carlo simulations to be run, we must fix the values of the filtering parameters \(i\) and \(j\) (see Eqs. (6) and (7)). For various combinations of them, with \(i\) taking several intermediate values in \([2, 6]\) and \(j\) in \([2, 10]\), we can predict the output \(P_{\delta}\) of the method by explicitly computing \(\text{FFT}[\alpha(\xi_i)]\) (see Eq. (3)). We can then average the resulting 3D power spectra in shells of width \(K_F\) and compare them to the targeted one. The results, presented in Fig. 1, show that, in order to maximise the range of well simulated modes, \(i\) must be equal to 3 for redshifts \(z = [0, 0.5]\), then \(i = 4\) for \(z = [1, 1.5]\) and, finally, \(i = 5\) for \(z = 2\). On the other hand, an arbitrarily large value for \(j\) can be chosen. Figure 1 only presents the results for the \(\Lambda\)CDM scenario. We note that in the present case, as well as for the following figures, only one cosmology will be discussed if the same results are obtained also for the other one.

Moreover, from this figure, we can see the successive effects of the two filtering, Eqs. (6) and (7). The difference between the dotted black line and zero is due to the first filter. We recall that this smoothing is exclusively due to the fact that the fields are set on a finite grid. The difference between the dotted black line and the coloured ones shows the impact of the second filter, required to solve issue of the negative elements of the spectrum.

For the present analysis, in the following, we use the same doublet parameters \((i, j)\) for all redshifts and cosmologies, fixing them to \((i = 2.5, j = 8)\). This choice was first motivated by the fact that we want to carry out a comparative study of the method between the redshifts, without adding extra tuning to enhance the output statistics (i.e. maximising the range of validity of the produced power spectra). Second, we recall that the PDF is obtained from a PCS MAS. This has the particularity of acting as a filter on the power spectrum fairly well reproduced by Eq. (6), with

\[
\text{PDF on the power spectrum fairly well reproduced by Eq. (6)},
\]

the nBodyKit\(^9\) (Hand et al. 2018) software, applying the interlacing method to reduce the aliasing (Sefusatti et al. 2016).

The averaged PDF and filtered power spectra (following the prescription in Eq. (6)) of the 50 simulations can then be directly used as targets for the Monte Carlo code. It is to be stressed that the method requires a high level of precision in the estimation of the PDF that is targeted. In practice, we estimate a logarithmically spaced histogram of the \(\delta\)-field from \(-1\) to the maximal value of \(\delta\) obtained from a catalogue smoothed on the grid following a defined MAS (\(\delta \sim 10^3\) in the present case of a PCS on the DEMUNi\_cov catalogues). We obtained about \(10^6\) values.

A bad estimate (in terms of precision) of the PDF leads to a bad computation of the transformation, \(\mathcal{L}\), linking the Gaussian field to the non-Gaussian fields (see Eq. (1)) and, as a consequence, the two-point correlation function of the initial Gaussian field that systematically leads to a higher number of negative elements in the power spectrum. However, this can, in practice, be avoided by resorting to a fit of a given analytical PDF. For our purposes, we checked that having 50 simulations was enough to avoid issues with extreme peaks (related to a bad estimate of the tail of the PDF). Finally, in the case where the randomly drawn Gaussian fields hits the 0th percentile in the \(\delta\)-PDF, nothing pathological can affect the field, typically because the actual probability of finding \(-1\) is zero. On the other hand, for the 100th percentile, this operation would be similar to clipping because a mode cannot suppot the maximum density contrast seen in the 50 DEMUNi\_cov simulations.

\[
\text{i = 2.5. Given such a choice, we try to keep a good consistency between these two statistics. Moreover, since we can’t anticipate an eventual effect of a doublet (i, j) on the produced covariances, we chose for it to be the same no matter the redshift, to ease the interpretation of the various output statistics.}
\]

We ran 5000 simulations of COVHOS catalogues per redshift in the \(\Lambda\)CDM and 16\(\mu\)m cosmologies. Each of them consists of \(10^8\) dark matter particles in boxes of size \(L = 1000\ Mpc\ h^{-1}\). We recall that the target statistics of the method are the ones estimated on the DEMUNi\_cov simulations.

After removing the relatively low shot noise contribution to the estimated power spectra \((P_{5\%} \sim 0.04 h^{-3} \mathrm{Mpc}^3)\), we compare them to the ones estimated from the DEMUNi\_cov in Fig. 2. As expected, we can observe a smaller range of reliable simulated wave modes when compared to its grid equivalent in Fig. 1. Indeed, as already discussed, the Poisson sampling gives rise to a smoothing of the cosmic field that visually starts to dominate around \(k \sim 0.2\ h\ Mpc^{-1}\). This smoothing has the direct consequence to reduce the maximum accurately simulated Fourier mode to around \(k \sim 0.25\ h\ Mpc^{-1}\) (in the 1-\sigma limit). We recall that this result is obtained assuming a grid precision of \(N_x = 1024\). Improving this setting would lower the effect of the smoothing induced by the applied assignment schemes.

\[\text{nbodykit.readthedocs.io}\]
In galaxy surveys, the observed positions of tracers of the matter field are systematically affected by their peculiar velocities. As a consequence, the power spectrum is affected at all scales, this is the so-called redshift-space distortion (RSD) effect. Indeed, it acquires an angular dependence with respect to the line of sight, losing the property of being isotropic. The amplitude of its angular average is enhanced in the linear regime due to the coherent infall of objects toward high density regions; this is the so-called Kaiser effect (Kaiser 1987). In the non-linear regime (on smaller scales), the density fluctuations and velocities are coupled in a non-trivial way (Scoccimarro 2004) which make the RSD effects on the power spectrum difficult to be predicted accurately.

However, because of these negative spectrum problems to be corrected, it can be expected that the method (whatever the grid accuracy employed) may be limited to a minimum scale. It is also noted that this filtering effect is similar, regardless of the redshift. This is due to the fact that we have defined the parameters $i$ and $j$ as constants.

When switching to the configuration space, we estimated the two-point correlation functions over the 50 DEMNUni $\_c$ov realisations at $z = 0$ and for the 16nu cosmology only. The outputs were compared to a set of 100 correlation functions estimated on the COVMOS set of simulations. We restricted this number given the high computational cost of estimating such statistics. Figure 3 shows this comparison in the range of scales $[0.1, 125]$ Mpc $h^{-1}$, depicting a high level of agreement between the two methods for all scales higher than 7 Mpc $h^{-1}$. Below this limit, correlations seem to be over-produced in the range of $[1.5, 5]$ Mpc $h^{-1}$ and under-produced for scales of $<1.5$ Mpc $h^{-1}$.

### 4. Theoretical pipeline in the redshift space

In galaxy surveys, the observed positions of tracers of the matter field are systematically affected by their peculiar velocities. As a consequence, the power spectrum is affected at all scales, this is the so-called redshift-space distortion (RSD) effect. Indeed, it acquires an angular dependence with respect to the line of sight, losing the property of being isotropic. The amplitude of its angular average is enhanced in the linear regime due to the coherent infall of objects toward high density regions; this is the so-called Kaiser effect (Kaiser 1987). In the non-linear regime (on smaller scales), the density fluctuations and velocities are coupled in a non-trivial way (Scoccimarro 2004) which make the RSD effects on the power spectrum difficult to be predicted accurately.

In the present section, we show our practical implementation of peculiar velocities in the COVMOS code. The main idea is to first generate a velocity field correlated with the density field so that we make sure that we are able to reproduce the Kaiser effect on large scale.

In principle, applying the Monte Carlo pipeline (as detailed in Sect. 2) to generate a continuous velocity field (as long as a corresponding power spectrum and a PDF are provided) is perfectly feasible. However, several subtle points must be addressed. The first is related to the fact that we expect the divergence of the peculiar velocity field, $v_p$, and the density field, $\delta$, to be correlated to satisfy the linear continuity equation:

\[
\bar{\nabla} \cdot v_p = -aH(a)f\delta(x),
\]

where $f$ is the growth rate of structures. This implies that the velocity field cannot be simulated in an independent way from the density field, $\delta$.

Beside, once the velocity field has been generated on a regular grid in space, we need to carefully associate a velocity to the previously generated tracers anywhere in the simulated volume. In the following, we detail these two main steps in order to simulate a catalogue of objects together with their peculiar velocities.
4.1. Simulating the velocity field

The method that we developed to generate the velocity field is based on the fact that when estimating the velocity field with a Delaunay tessellation in DEMNUni simulations (as in Bel et al. 2019), the PDF of each velocity component appears to be close to Gaussian (see the black dashed line in Fig. 5, as compared to the Gaussian model in red dashed). In addition, we want to reproduce the power spectrum of the divergence of the velocity field and, finally, as stated before, we need to satisfy the continuity equation.

Let us define the scaled velocity field $\mathbf{u}(x)$ as:

$$ u(x) \equiv -\nu \rho(x) \frac{1 + z}{100E(z)}, \quad (9) $$

where $E(z) \equiv H(z)/H_0$, such that the observed comoving distance, $s$, of an object can be deduced from the true one $r$ as $s = r - u_x (u_x$ being the line of sight projection of $\mathbf{u}$). The idea is then to generate the divergence, $\theta(x) = \nabla \cdot \mathbf{u}$, of the scaled velocity field in Fourier space according to the power spectrum of the divergence of the velocity field and to assume that the velocity field, $\mathbf{u}$, is curl-free:

$$ \theta_k = -i \frac{k}{k^2} \theta_k. \quad (10) $$

This way it is only necessary to generate the divergence, $\theta_k$, in Fourier space with a real and an imaginary part following a centred Gaussian distribution with a variance given by the power spectrum of the divergence $P_{\theta\theta}(k)$. We note, however, that in order to satisfy to the continuity equation, we impose the condition that the Gaussian field, $\theta$, used to generate the density field is also used to generate the divergence. This means that we impose strictly the continuity equation only in the linear regime. To do so, we can rely on the fitting functions given in Bel et al. (2019):

$$ P_{\theta\theta}^{lin}(k) = k_0^2 P_{\theta\theta}^{cb}(k), \quad (11) $$

$$ P_{\theta\theta}(k) = P_{\theta\theta}^{cb}(k) \exp(-\alpha_1 k^2 - \alpha_2 k^4 - \alpha_3 k^6), \quad (12) $$

where $P_{\theta\theta}^{cb}(k)$ stands for the linear cold dark matter plus baryons power spectrum (numerically predicted by Boltzmann codes) and

$$ a_1 = -0.817 + 3.198 \sigma_{8, m}, \quad (13) $$

$$ a_2 = +0.877 - 4.191 \sigma_{8, m}, \quad (14) $$

$$ a_3 = -1.199 + 4.629 \sigma_{8, m}, \quad (15) $$

where $\sigma_{8, m}$ is the rms of the linear matter fluctuations smoothed on spheres of radius $8 \, \text{Mpc} \, h^{-1}$ and must be predicted for the total matter (accounting for massive neutrinos).

In summary, we start by simulating, on a Fourier grid, the Gaussian $\theta$-field using the same random Fourier phases as for the density field, defined by the non-linear $\theta - \theta$ power spectrum in Eq. (12). This field is then converted in the 3-dimensional velocity field in Eq. (10), which we apply an inverse Fourier transform to so that we can obtain the velocity field in configuration space, $\mathbf{u}(x)$, taken from Eq. (9).

4.2. Drawing peculiar velocities

In order to properly reproduce redshift-space distortions, it is necessary to assign velocities to the objects inside the simulated catalogue. To this aim we could interpolate the regularly spaced velocity grid $\mathbf{u}(x)$ at all the positions of the objects, however, this would not accurately reproduce the non-linear redshift space distortions. Indeed, on small scales, we would expect that the angular average power spectrum is suppressed by the fact that inside halos, the velocities are not coherent; instead, they are close to be random thanks to the virial theorem. That is the reason why we drew the velocity of each object out of a Gaussian distribution centred on the interpolated value of the velocity field, $\mathbf{u}$, at the position of the object and with a given dispersion, $\Sigma$. This is a similar method to the one used by de la Torre & Peacock (2013) in order to assign velocities to halos below the mass resolution of a simulation.

Aimed at being as realistic as possible, we allow the velocity dispersion to depend on the underlying matter density. In fact, we would expect that more massive halos have a large velocity dispersion (Elahi et al. 2018). Using the DEMNUni$_{\text{cov}}$ simulations, we estimated the relation between the local density $1 + \delta$ and the velocity variance $\sigma_V^2$. By fixing the grid sampling to $N_t = 1024$, we can run an estimation of the local velocity variance as a function of the local density contrast value over the 50 simulations in the two cosmologies, ΛCDM and 16uw, for the five redshifts $z = [0, 0.5, 1, 1.5, 2]$. This is represented in Fig. 4 where (in the upper panel) we show that no matter what the considered redshift, there is a clear relation between the velocity dispersion and the density. Indeed, this relation exhibits a very low level of stochasticity, meaning that we can further assume there is an analytical relation between the local dispersion and the density. In addition, the lower panel of Fig. 4 shows that the presence of massive neutrinos induces changes in this relationship. As expected, the dispersion in a cosmology with massless neutrinos is systematically higher.

When inspecting Fig. 4, we can see that the global behaviour of the velocity dispersion-density relation can be approximated as a power law. Even if this does not seem to be accurate for low values of the density, we further assume a power law relationship. This is justified for two reasons. First when the density is low, so is the dispersion. Thus, the way we model the relation is expected to affect less the final velocities of the various objects in the catalogues. Second, when the density is low, we estimate both the density and the velocity dispersion with a low number of objects, thus the shot noise can introduce a bias. Thus, in the end, we adopt the following analytical form

$$ \Sigma(x) \equiv \beta \rho^\alpha, \quad (16) $$

where $\alpha$ is fixed from the velocity dispersion-density relation estimated from DEMNUni$_{\text{cov}}$ and listed in the right column of Table 1. The second parameter, $\beta$, is not taken from the simulations, but it needs to be set in order to make sure that the one-point velocity dispersion $\sigma_{1P}$ of the catalogues is matching the one of DEMNUni$_{\text{cov}}$ (this will be explained more in details in the following).

We note also that the exponent $\alpha$ of the power law can be related to dynamics inside over-densities and thus (by extension) inside halos. In fact, we checked that if one uses halos to estimate the velocity dispersion-density relation, then the same value for $\alpha$ is measured but with a different normalisation $\beta$.

The measured one-point velocity variances are listed in the left column of Table 1, where we can see that it decreases with the redshift and that adding massive neutrinos is reducing it. As anticipated, we fix the normalisation $\beta$ so that the one-point velocity variance, $\sigma_{1P}^2$, is the same as the one-point velocity variance measured in the DEMNUni$_{\text{cov}}$ catalogues.

Furthermore, matching the one-point velocity variance allows us to provide a physical description of the amplitude parameter, $\beta$. Thus, we can show how the one-point variance
is related to the prescription in Eq. (16) when going from the large scale velocity to the velocity of individual objects. Indeed, the peculiar velocity, $V_p$, assigned to a particle, can be formally expressed as

$$V_p = w + \Sigma(\rho)v,$$

(17)

where $w = -aHu$ and $v$ is drawn from an uncorrelated tri-variate centred reduced Gaussian distribution. This means that the conditional probability distribution, $\mathcal{P}(V_p|w_\rho, \rho)$, of a single component, $V_c$, of the vector, $V_p$, is given by:

$$\mathcal{P}(V_c|w_\rho, \rho) = \frac{1}{\sqrt{2\pi} \Sigma(\rho)} e^{-\frac{(w_\rho - \rho V_c)^2}{2\Sigma^2(\rho)}}.$$

(18)

Since we define the variance of a single velocity component as the arithmetic mean of the square of the velocity component, it turns out that:

$$\sigma^2_v = \langle \rho V_c^2 \rangle = \iint dV_c dw_\rho d\rho V_c^2 \mathcal{P}(V_c|w_\rho, \rho),$$

(19)

due to the fact that the density of points is not uniformly distributed, thus, regions with higher density will provide a larger contribution to the expectation value.

From Eq. (19), in order to predict the one-point velocity variance, $\sigma^2_v$, we need to express the one-point probability distribution of the object velocity component, $V_c$, which depends on the density and the velocity, $w_\rho$$, as

$$\mathcal{P}(V_c|w_\rho, \rho) \equiv \mathcal{P}(V_c|w_\rho)\mathcal{P}(\rho).$$

(20)

Substituting Eqs. (18) and (20) in the expression in Eq. (19) and using the fact that the second-order moment of the Gaussian distribution is reduce to the sum of its variance and the squared of its expectation value, Eq. (19) is simplified to:

$$\sigma^2_v = \int dw_\rho d\rho \mathcal{P}(V_c|w_\rho)\rho \left[\langle w_\rho^2 \rangle + \Sigma^2(\rho)\right] \equiv \sigma^2_w + \sigma^2_\rho.$$

(21)

In the above expression the first term, $\sigma^2_w \equiv \int dw_\rho d\rho \rho^2 \mathcal{P}(V_c|w_\rho, \rho)$, is the variance of the interpolated velocity field that can be estimated independently from the choice of the function $\Sigma(\rho)$, while this is not the case for the second marginalised term, defined as $\sigma^2_\rho \equiv \int d\rho \Sigma^2(\rho)$. Then, coming back to our simple model in Eq. (16), the $\beta$ parameter can be obtained using Eq. (21) as:

$$\beta = \frac{\langle \rho \rangle - \langle \rho \rangle^2}{\sigma^2_\rho},$$

(22)

where $\sigma^2_\rho \equiv \langle \rho \rangle^2$ is directly estimated on the interpolated density field at the object coordinates.

In Fig. 5, we compare the distribution of the velocity components of the regularly sampled velocity field $w$, estimated from the DEMNUni_cov, with the same distribution obtained from COV莫斯. We see that the Gaussian approximation assumes to generate the velocity field is indeed a good approximation. In addition, we compare the one-point velocity PDF estimated from a COVmos realisation of (exceptionally) $\sim 10^9$ dark matter particles with the one estimated from particles in a DEMNUni_cov realisation. We can see that, assuming a power law relation between the velocity dispersion and the matter density (Eq. (16)) and fixing the normalisation of the relation by requiring that the one-point velocity dispersion of COVmos matches the one of DEMNUni_cov (Eq. (22)), allowing us to reproduce the non-Gaussian distribution of the particle velocity. Indeed, by matching only the second order cumulant moment, we can also reproduce the exponential tails of the distribution. In the next section, we show the efficiency of our approximations for generating the velocity field at the level of the power spectrum in the redshift space.

5. Applications in the redshift space

In this section, we describe how we applied the whole pipeline for simulating particles (described in Sect. 2) and their associated peculiar velocities (Sect. 4). We study to which extent the various two-point statistics, associated error bars and correlations are reliably simulated.

5.1. Validating the two-point statistics

With the same settings applied in Sect. 3, we simulated 5000 catalogues of $10^8$ dark matter particles with associated peculiar velocities, for all five redshifts and for the two cosmologies. We estimate on each realisation the multipoles of the power spectrum of the peculiar velocity field, expressed in terms of the peculiar velocity, $\bar{V}$, and the peculiar density contrast, $\delta$.

Table 1. Averaged one-point velocity rms (in Mpc/h) of the one-point velocity distribution as measured from particles in the 50 DEMNUni_cov simulations, per redshift and per cosmology.

| $z$ | $\sigma_v$/ACDM | $\sigma_v$/16nu | $\alpha$/ACDM | $\alpha$/16nu |
|-----|-----------------|-----------------|----------------|----------------|
| 0   | 3.9412          | 3.7870          | 0.7437         | 0.7544         |
| 0.5 | 3.6393          | 3.5032          | 0.8277         | 0.8368         |
| 1   | 3.1772          | 3.0640          | 0.9032         | 0.9086         |
| 1.5 | 2.8930          | 2.7936          | 0.9363         | 0.9414         |
| 2   | 2.5650          | 2.4818          | 0.9594         | 0.9554         |

Notes. The $\alpha$ parameters are inferred from the same set of simulations using linear regressions in ranges $\delta + 1 \in [50, 500]$.
Fig. 5. Comparison of the one-point and smoothed PDFs estimated from the COVMOS and DEMNUni__cov experiments. Top: estimated one-point (solid lines) and smoothed (dashed lines) velocity PDF from one COVMOS simulation in red and one DEMNUni__cov simulation in black. The smoothed velocity PDF is estimated from a Delaunay tessellation of the particle coordinates of the DEMNUni__cov_01 experiment. This map is then interpolated on a grid of sampling parameter $N_s = 512$. Here the cosmology is ΛCDM. Bottom: residual between the DEMNUni__cov and COVMOS PDF.

Fig. 6. Redshift-space power spectrum comparison between the COVMOS and the DEMNUni__cov experiments. Top panel: average of the estimated redshift-space power spectrum monopoles over 5000 realisations of COVMOS catalogues in the $10h^{-1}$m cosmology. The dispersion on each realisation is represented by the shaded area. The same quantities are represented in black for the 50 DEMNUni__cov realisations. Bottom panels: relative deviation between the averaged COVMOS and DEMNUni__cov outputs in dashed black lines with error bars following the same colour-coding.

The ratio $\hat{\beta}(\ell) = 2\ell + 1 \overline{d\mu_k P(k,\mu_k) \ell^\ell(\mu_k)}$, where $\ell^\ell(\mu_k)$ are defined as the Legendre polynomials on the order of $\ell$, $\mu_k \equiv \cos(\phi_k)$ and $\phi_k$ is the angle between the line of sight and the wave vector, $\mathbf{k}$. We adopt the plane parallel limit which consists in choosing the line of sight as being a fixed direction in the simulated co-moving volume.

Generating the entire set of simulations and measuring their associated real- and redshift-space power spectra took nearly 20 days, running on about 940 processors of 2.4 GHz. We note that reducing the precision of the COVMOS grids to $N_s = 512$, three days are necessary to simulate and estimate the power spectrum of 100 000 catalogues.

In Figs. 6 and 7, we compare the power spectra of COVMOS to those estimated in the reference DEMNUni__cov, N-body simulations. An initial observation of Fig. 6 leads us to the very same conclusions as those drawn for its real-space equivalent Fig. 2. The ratios $\hat{P}_{\text{COVMOS}}(k)/\hat{P}_{\text{DEMNUni}}(k)$ in both spaces do indeed seem similar. The maximal wave-mode at which the power spectra are reliably simulated are conserved. We can, however, note a gain in power of the redshift-space power spectrum at $z = 0$, with respect to its real-space equivalent. Such effects can be due to an overestimation of the β parameter in Eq. (22), which, in our code implementation, is averaged over a tens of realisations before becoming constant for the whole set of simulations. To correct for this effect, the beta parameter should be obtained from a larger number of realisations. Our new version of COVMOS averages 100 values for β, estimated on single realisations.

In order to quantify the robustness of the proposed velocity model, we can see the redshift-space to real-space averaged monopole ratio in Fig. 7. In the left panels, we can see that the effects of RSD on the power spectrum are well reproduced in COVMOS when compared to DEMNUni__cov. For both the enhancement of the apparent clustering in the linear regime ($\leq 0.1h^{-1}\text{Mpc}^{-1}$) or its suppression in the mildly non-linear regime ($\geq 0.1h^{-1}\text{Mpc}^{-1}$), the velocity model allows us to reproduce (with high fidelity) the large-scale limit (Kaiser boost) as well as the transition to smaller scales.

Finally, looking at the ratio between the previous quantity in the COVMOS and DEMNUni__cov cases allows us to extract and discuss the validity domain of the proposed velocity modelling (right panel of Fig. 7). While the model seems to perfectly recover the RSD effects in the linear regime at the percent level, a gradual departure can be observed in the non-linear regime. We observe a lack of power for $k \geq 0.6h^{-1}\text{Mpc}^{-1}$, which can be due to the severe filtering of the COVMOS fields in this regime. Since we cannot expect the peculiar velocity assignment method to
be adapted to a highly suppressed clustering amplitude, the grid precision here seems to be ill-suited with regard to validating the velocity model on highly non-linear scales \((k \geq 0.6 \, h \, \text{Mpc}^{-1})\).

For completeness, in Fig. 8, we compare the higher order multipoles of the power spectrum between COVMOS and DEMUn4i_cov. As in the case of the monopole, in the linear regime, both the amplitude and the dispersion of the simulated multipoles are in good agreement. In addition, it appears that in the mildly non-linear regime, these multipoles are not impacted by the various filterings affecting the monopoles (see Fig. 6). Indeed, none of these multipoles show a clear lack of power for \(k > 0.2 \, h \, \text{Mpc}^{-1}\), so that this anisotropic part of the power spectrum turns out to be reliably simulated on a wider range than the redshift-space monopole.

As in real space, we can also make a comparison in redshift-space at the level of the two-point correlation function. We ran the estimation of the configuration space correlation function multipoles (applying periodic boundary conditions) over a smaller sample of 100 COVMOS realisations, in the 16nu cosmology only at \(z = 0\). In addition, we estimated the two-point correlation functions on the complete set of 50 DEMUn4i_cov simulations. Figure 9 displays the averaged monopole, quadrupole, and hexadecapole, compared between COVMOS and DEMUn4i_cov on the scale range of \(r \in [0, 125] \, \text{Mpc} \, h^{-1}\). We can see that the redshift space multipoles produced by COVMOS are in agreement with what is expected from DEMUn4i_cov at least down to 40 Mpc \(h^{-1}\). Indeed, in that regime, it is not only the amplitude that is matching, but also the dispersion of the measurements. We can also notice that despite the fact that we made this verification at redshift \(z = 0\), the monopole is well reproduced on even smaller scales \((-20 \, h^{-1}\text{Mpc})\) than the other multipoles. This serves as a validation of our scheme to assign velocities in configuration space as well.

5.2. Comparing the real and redshift-space covariance matrices

In this section we compare the covariance matrices produced by the two methods, obtained from the standard estimator:

\[
\hat{C}_{ij} = \frac{1}{N-1} \sum_{i=1}^{N} (P^s(k_i) - \mu_i)(P^s(k_j) - \mu_j),
\]

\[
\mu_i = \frac{1}{N} \sum_{i=1}^{N} P^s(k_i),
\]

where \(P^s(k_i)\) is the power spectrum of the \(s\)-th simulation (over a total of \(N\) evaluated at mode \(k_i\)).
The statistics, according to Scoccimarro et al. (1999), can be decomposed into two terms:

$$C_{ij} = \frac{\hat{P}(k_i)^2}{M_k} \delta_{ij} + k_i^2 \bar{T}(k_i, k_i),$$

(25)

referred to as the Gaussian and non-Gaussian contributions, respectively. The former is related to the amplitude of the power spectrum, normalised by the number of independent averaged modes $M_k$ within shells centred around $k_i$, only affecting the diagonal of the matrix. The second affects the whole matrix and depends on the cross-shell-averaged trispectrum:

$$\bar{T}(k_i, k_j) = \int \int T(k_1, -k_1, k_2, -k_2) \frac{d^3k_1}{V_{k_1}} \frac{d^3k_2}{V_{k_2}},$$

(26)

where the trispectrum can be constructed from the density field as

$$\langle \delta k_1 \delta k_2 \delta k_3 \delta k_4 \rangle = \delta^D(k_1 + k_2 + k_3 + k_4) T(k_1, k_2, k_3).$$

(27)

Expected to be zero in the case of a Gaussian realisation, this second contribution directly weighs the non-Gaussian nature of the simulated cosmic fields.

While the first term (in the case of the COVMOS method) is expected to be suppressed when the various filtering starts to act ($k > 0.2h \, \text{Mpc}^{-1}$ for a grid precision of $N_r = 1024$), the second term is not so obvious to assume. The main approximation of COVMOS is to suppose that $T(k_i, k_j)$ can be (to some extent) properly reconstructed if both the power spectrum and the PDF are accurate. However, such a connection remains difficult to foresee. In any case, this term is expected to dominate the non-linear regime and to become raised in power as the targeted PDF follows a more clustered, skewed shape.

With regard to the estimator of the covariance, Eq. (24) will be associated in the following an analytical error-bar, which is predicted when assuming a Gaussian multivariate distribution of the data (Wishart 1928):

$$V[\hat{C}_{ij}] = \frac{C_{ij}^2 + C_i C_j}{N - 1}.$$

(28)

In Fig. 10, we show the contributions to the variance, both in the COVMOS and the DEMNu\textsubscript{i} cov cases, for all redshifts and in the 16u cosmology. The figure shows that for the five redshifts, the full variances, $\hat{C}_{ij}$, are matching within Gaussian error bars. This observation can also be extended to modes well beyond those for which the power spectrum is correctly reproduced. This effect can be understood when decomposing the variance in Gaussian and non-Gaussian contributions (Eq. (25)): the former is computed from the average power spectrum over the realisations, whereas the latter is obtained by subtracting the Gaussian contribution from the total variance. In doing so, we can see that in the linear regime, the Gaussian term that dominates the second term (by about one order of magnitude) unsurprisingly follows the same trend as discussed for the power spectrum. It is accurately reproduced, but starts to disagree with the $N$-body reference around $k = 0.2h \, \text{Mpc}^{-1}$. On the other hand, as the damping happens at scales where the trispectrum term starts to dominate, it allows us to recover this lack of power and produces an overall variance with correct amplitude. Nevertheless, we can observe at $z = 0$, a systematic overestimate of this term, from $k \sim 0.25/h \, \text{Mpc}^{-1}$ both in the real and redshift space. This could be due to the insufficiency of our approximations at very low redshift.

We compare all terms of the covariance matrices obtained for the $N$-body and the COVMOS method for the three redshifts $z = \{0, 1, 2\}$ in Figs. 11–13, respectively. This comparison is conducted in real-space for the monopole and in the redshift space for all multipoles, where we also show the cross-covariances between multipoles. We represent the residual of the COVMOS and DEMNu\textsubscript{i} cov matrices in standard deviation, namely:

$$R_{ij} = \frac{\hat{C}_{ij} - \hat{C}_{ij}}{\sqrt{\sigma_{ij}^2 + \sigma_{ij}^2}},$$

(29)

where the $\sigma_{ij}^2$ terms are computed using Eq. (28). Looking at these three figures, we can first remark a large number of over- or under-estimations of the covariance terms – without any apparent global trend. They constitute the only observable feature that we can see in redshift space at $z = 1$ and $z = 2$ and in real
space at $z = 2$. This is explained by the stochastic noise, unrelated to the validity of the method, expected to follow a centered and reduced normal distribution. Thus, drawing the distribution of $R_I$ against its prediction helps in the discussion of the validity of the produced covariances. To avoid overloading the figures, this is shown only at $z = 0$ in Fig. 14, taking a maximal mode $k_{\text{max}} = 0.25h\text{Mpc}^{-1}$. In this regime in the redshift space, we can see that the measured distribution is in good agreement with the prediction, demonstrating that no bias in the COVMOS matrices can be detected. This is not as strong in the real space, where we can detect a shift of the distribution toward larger $\sigma$ values. This depicts an overall over-estimation of the covariance elements that can be easily spotted in the top panels of Fig. 11.

Secondly, we can see in real space a migration of an over-correlation pattern toward higher scales for decreasing redshift. These over-correlations are maximised around the diagonal at $z = 0$ from $k \sim 0.2h\text{Mpc}^{-1}$ to $k \sim 0.5h\text{Mpc}^{-1}$ and no change can be noticed when switching from one cosmology to another. Once again, this can be attributed to the weak validity of the COVMOS approximations at low redshift $z \sim 0$.

On the other hand, it turns out that the redshift-space covariance terms are systematically better simulated than their equivalents in real space. This effect is more apparent at $z = 0$ where the systematic over-estimate of the correlations in real space are significantly reduced in the redshift space. This improvement is in agreement with Fig. 10, showing a slight increase of the range of reliability of the trispectrum in redshift than in real space. According to Eq. (25), the off-diagonal terms of the covariance matrix are only related to the trispectrum terms, explaining why the whole matrices in the redshift space are more accurate than in real-space. Further investigation is required to determine why the trispectrum is better simulated in the redshift space as compared to real space. It is evident that the presented method cannot reproduce the full trispectrum in real space, thus limiting us in that regard. However, in the redshift space, the velocity field features contribute to a significant portion of the trispectrum shape, and the proposed method might perform better due to the relationship between the density and local velocity dispersion.

Concerning the other multipoles and their cross-covariance terms in the redshift space, no particular over or under-correlation pattern throughout the simulated redshifts can be observed beyond the noise, as depicted in Fig. 14. At these probes redshifts, both error bars and correlations of the multipoles are in good agreement with the $N$-body outputs.

Finally, Figs. 15 and 16 show the results for the variance and the covariance of the two-point correlation function multipoles, respectively, both in the real and the redshift spaces.

Focusing on real space, the variance of the two-point correlation function monopole seems to be reproduced in a reliable way down to $\sim 6 \text{Mpc} h^{-1}$, before being successively over- $(2 < r < 6 \text{Mpc} h^{-1})$ and under- $(r < 2 \text{Mpc} h^{-1})$ estimated, which is in agreement with the observation made directly on the two-point correlation function monopole in Fig. 9. These observations can also be made for the whole monopole covariance in the upper panel of Fig. 16: a majority of the matrix elements fall in the
Fig. 11. Difference in $\sigma$-deviation between the cross-covariance matrices of the power spectrum in Fourier space obtained from 5000 COVMONO and 50 DEMURNI-cov realisations at $z = 0$. This difference is normalised by the predicted Gaussian error in Eq. (28); also see Eq. (29). Upper panels show the monopoles in real space for the two cosmologies, while the bottom panels stand for the cross-covariance terms in the redshift space, for the 16nu cosmology only.
Fig. 12. Difference in σ-deviation between the cross-covariance matrices of the power spectrum in Fourier space obtained from 5000 COVMOS and 50 DEMUn1_cov realisations at \( z = 1 \). This difference is normalised by the predicted Gaussian error in Eq. (28); also see Eq. (29). Upper panels give the monopoles in real space for the two cosmologies, while the bottom panels stand for the cross-covariance terms in the redshift space, for the 16nu cosmology only.
Fig. 13. Difference in $\sigma$-deviation between the cross-covariance matrices of the power spectrum in Fourier space obtained from 5000 COVMOS and 50 DEMUNi_cov realisations at $z = 2$. This difference is normalised by the predicted Gaussian error Eq. (28), see Eq. (29), in Eq. (28); also see Eq. (29). Upper panels give the monopoles in real space for the two cosmologies, while the bottom panels stand for the cross-covariance terms in the redshift space, for the $16nu$ cosmology only.
Fig. 14. Normalised distributions of the residual of the covariance matrices normalised by Gaussian error (the $\sigma$-deviation distribution, see Fig. 11) for the multipoles of the power spectrum at $z = 0$ in blue. Its Gaussian fit (on the two first moments) is represented in black and is compared to the standardised normal distribution in orange. Upper panels presents the distributions in real space for the two cosmologies, while the bottom panels give the results in the redshift space, for the 16$\nu$ cosmology only. In these panels, the maximal Fourier mode is set to $k_{\text{max}} = 0.25 h \text{Mpc}^{-1}$.

$\langle |\sigma| \rangle$ deviation ($>68\%$) in the regime $>\sim 6 \text{ Mpc} h^{-1}$ before getting over- or under-estimated above this limit.

In the redshift space, as for the power spectrum analysis, a significant improvement in the simulation of the errors and correlations are observed on small scales. The variance of the monopole is now reliable down to $\sim 3 \text{ Mpc} h^{-1}$, while the under-estimation of the correlation between scales (observed in the real space) is damped in this case. Despite the fact that we have a less
significant number of realisations in the case of the two-point correlation function, we can see that in general the covariance and cross-covariance of the two-point correlation multipoles is well reproduced. More realisations would be needed to assess the significance of the 2-sigma deviations around 30 Mpc\(^{-1}\) in the cross-covariance between the monopole and the quadrupole and in the covariance of the quadrupole. Indeed, we started with 50 realisations only and so, generating 50 more greatly improved the comparison.

6. Conclusion and discussion

Although it is apparently straightforward, the reliance on covariance matrices for two-point statistics estimators evaluated from mock catalogues generally leads to two main issues. The first is the sampling noise that propagates down to the parameter estimation matrices for two-point statistics estimators evaluated from mock catalogues generally leads to two main issues. The first is the sampling noise that propagates down to the parameter estimation matrices. Indeed, it is faster to generate a realisation than to estimate the power spectrum of that realisation, and the diagonal covariance, is in good agreement with the one obtained from 50 DEMNUni\_cov realisations in black and 100 COVMOS realisations in red. Assuming a Gaussian distribution of the data, errors bars are computed from Eq. (28).

As a result, it is necessary to specify a set of statistical quantities and relations as inputs of the code. These are used as targets for the pipeline. These include: (i) the power spectra of the density and (ii) velocity fields, (iii) the one-point probability distribution function of the density field, and (iv) a relation between the velocity dispersion and the density.

In addition, we validated the whole COVMOS pipeline in the real and redshift spaces, for two-point statistics (the power spectrum and the correlation function) and their associated covariance matrices, against state-of-the-art N-body simulations (DEMNUnii). We investigated the performance of two cosmological models: a standard ΛCDM model and one including massive neutrinos. In both cosmological scenarios, we show that for a spatial resolution of the order of 1 h\(^{-1}\)Mpc, in a comoving volume of 1 h\(^{-3}\)Gpc, the matter power spectrum multipoles in the redshift space are reliably reproduced up to \(k = 0.25 \, h \, \text{Mpc}^{-1}\) (down to \(r = 10 h \, \text{Mpc}\) for the two-point correlation function).

We extended the validation of the method to 4-point statistics in Fourier space. More precisely, we verified that the shell-averaged trispectrum, \(T\), which enters the power spectrum covariance, is in good agreement with the one obtained from 50 N-body simulations. This results in an accurate reproduction of the power spectrum covariance matrices. Although it is slightly over-estimated in real space for \(k \sim 0.15 \, h \, \text{Mpc}^{-1}\), it turns out that in the redshift space, we can reach \(k \sim 0.2 \, h \, \text{Mpc}^{-1}\) for the monopole power spectrum.

Having validated the method against N-body simulation, a major advantage of the code is its numerical efficiency in terms of computational cost. As an example, for this work, we generated 5 \times 5000 catalogues of 10\(^8\) objects in 20 days on a HPC cluster\(^\text{10}\), using 940 processors of 2.4 GHz. Thus, we were able to generate an extremely large number (25k) of realisations in less than 20k cpu-hr, in order to produce noise-free covariance matrices. Indeed, it is faster to generate a realisation than to estimate the power spectrum of that realisation, and the difference is even more important when considering the estimation of the two-point correlation function. The method may need to be further investigated to reach a higher level of precision and accuracy before a potential application in the final data analysis of large galaxy surveys. We note that the precision of the method is governed by two factors. The first is the grid precision set by the user that has to have the effect of smoothing the statistics at a certain scales. The second constraint is less controllable. It is related to the production of negative elements in some of the pipeline power spectra. A scale limitation is therefore to be expected and would require further refinement.

\(^{10}\) The Dark Energy Cluster (DEC) hosted in CPPM, Marseille.
Fig. 16. Difference in $\sigma$-deviation between the covariance matrices of the multipoles of the two-point correlation function estimated from 100 COV莫斯 realisations and 50 DEMUnI_cov realisations. This difference is normalised by the predicted Gaussian errors estimated from Eq. (28). The upper panel stands from the comparison of the monopole in real space, while the bottom panels stands for the covariance terms of the multipoles in the redshift space.

inspection to be carefully determined. However, the associated public code can already be useful for a broad variety of applications. Indeed, its ability to simulate a wide variety of cosmological models and to achieve almost no sampling noise in the produced covariance matrices makes this method particularly relevant for testing likelihood pipelines (see Gouyou Beauchamps...
et al. in prep., for a first application) aimed at constraining cosmological parameters.

Finally, we recall that the COVMOS code is public, fully written in Python, ready to be used and adapted to receive the needed inputs either from theory or simulations. This is a method specialised in producing two-point statistics and associated by-products (covariance matrices) that may produces unrealistic outputs when considering other statistics. In particular the method does not account for structure formation physics, and is not capable of producing a cosmic web with the richness of collapsed structures (e.g. clusters, filaments and voids). It is therefore not advisable to use statistics beyond those presented in this study.

A natural extension of this method, which would bring it closer to the results of observations, would be the reconstruction of the observed light-cone that is aimed at accounting, for instance, for survey window function effects. In addition, this would allow for the production of angular maps, suited for the study of photometric statistics in the Spherical Harmonics domain. Given it is a method specialised in two-point statistics, correlated maps of observables such as galaxy clustering and galaxy shear could be produced, as well as their associated cross-covariances. Future extensions could also include radio observations such as CMB temperature maps, CMB-lensing, or even 21-cm maps. Given the improvement in their performance that future balloon flights will certainly prove useful in modern cosmology.

Acknowledgements. PB and SGB were supported by CNES, focused on Euclid mission. The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University -A*MIDEX, a French “Investissements d’Avenir” programme (AMX-19-IEF-008 -IPhU). The DEMNUMi simulations were carried out in the framework of “The Dark Energy and Massive-Neutrino Universe” project, using the Tier-0 IBM BG/Q Fermi machine and the Tier-0 Intel OmniPath Cluster Marconi-AI of the Centro Interuniversitario del Nord-Est per il Calcolo Elettronico (CINECA). We acknowledge a generous CPU and storage allocation by the Italian Supercomputing Resource Allocation (ISCRRA) as well as from the coordination of the “Accordo Quadro MoU” for lo svolgimento di attività congiunta di ricerca Nuove frontiere in Astronomia: HPC e Data Exploration di nuova generazione”, together with storage from INFN-CNAF and INAF-I2A.

References
Agrawal, A., Makiya, R., Chiang, C.-T., et al. 2017, JCAP, 10, 003
Alonso, D., Ferreira, P. G., & Santos, M. G. 2014, MNRAS, 444, 3183
Alonso, D., Ferreira, P. G., & Santos, M. G. 2014, MNRAS, 444, 3183
Alonso, D., Ferreira, P. G., & Santos, M. G. 2014, MNRAS, 444, 3183
Agrawal, A., Makiya, R., Chiang, C.-T., et al. 2017, JCAP, 10, 003
Varimbelli, G., Carbone, C., & Renzi, A. 2022a, MNRAS, 503, 1149
Zennaro, M., Angulo, R. E., Aricò, G., Contreras, S., & Pellejero-Ibáñez, M. 2018, MNRAS, 477, 616
Appendix A: Tri-linear assignment

In this appendix, we derive the relations that allow us to populate particles in a cell following a tri-linear interpolation scheme. The cubic volume symbolised a given voxel in tri-linear interpolation is, by construction:

\[ \tilde{\rho}(x, y, z) = \rho_{000}(a - x)(a - y)(a - z) + \rho_{001}x(a - y)(a - z) + \rho_{010}y(a - z)(a - z) + \rho_{011}xy(a - z)z + \rho_{100}z(a - y)z + \rho_{101}xz(a - x)yz + \rho_{110}yz(a - y)z + \rho_{111}yz(a - y)z, \]  

(A.1)

where \( \rho_{ijk} \) represents the eight local density values defining the vertices of the cube, replacing \((i, j, k)\) by \(0\) or \(a = L/N_s\). The coordinates \(x, y\) and \(z\) are defined in the interval \([0, a]\). Converting it into normalised probability, the 3D PDF is defined as:

\[ P\rho(x, y, z) = \frac{8}{a^3} \tilde{\rho}(x, y, z). \]  

(A.2)

Equation A.2 shows that the three coordinates cannot be generated independently, leading to consider conditional probabilities.

More specifically, the method of sampling a set of points in the cube will use the so-called classical inverse-transform sampling method (Devroye 1986). It relies on the notion of cumulative distribution function (CDF) to generate random numbers from any 1D probability distribution function \( P(z) \) in the definition ensemble \([a_1, a_2]\), defined by:

\[ \text{CDF}[P](z) = \int_{a_1}^z dz' P(z'). \]  

(A.3)

Then we have to solve analytically for \( z \) using the relation \( g = \text{CDF}[P](z) \), calling this inversion \( z = F(g) \). Finally, given \( F \), a uniform number \( g \) in \([0, 1]\) follows \( P(z) \).

The different steps for the coordinates assignation are as follows:12

1. Integrate \( P\rho(x, y, z) \) over \( z \); then \( y \); then \( x \).
2. Generating \( N_{cell} \) pseudo-random numbers according to the inverse transform sampling method described above for \( P\rho(x, y) \) completes the coordinates \( x \).
3. Since \( P\rho(x, y, z) = P\rho(x, y)P\rho(z|x, y) \), where \( g \) means the conditional probability of \( y \) given \( x \), a number \( N_{cell} \) of \( y \)-pseudo-random-coordinates depending on the previous ones, \( x \) can be generated following \( P\rho(y|x) \).
4. In the same way, as \( P\rho(x, y, z) = P\rho(x, y)P\rho(z|x, y) \), \( N_{cell} \) new random numbers targeting \( P\rho(z|x, y) \) allows the \( z \)-coordinates to be simulated.

Each marginalised probability calculated in point 1 takes the following forms:

\[ P\rho^x(z) = \frac{2}{a^2} \sum_{y,z} C(x), \]  

(A.4)

\[ P\rho^y(z) = \frac{2}{a^2} \sum_{y,z} D(y), \]  

(A.5)

\[ P\rho^{y,z}(z) = \frac{2}{a^2} \sum_{y,z} E(z). \]  

(A.6)

12 Note: there is no uniqueness in the method.

Appendix B: Debiasing the Monte Carlo covariance matrices

The clipping procedure of the negatives elements of the power spectrum, \( P_s(k) \), (see Sect. 2.2) inevitably produces a bias in the estimated covariances if they are blindly obtained using Eq. 24. Indeed, in doing so without caution, we would have to conclude that the covariance matrix obtained from C0VMOS is systematically over-estimated. This is a spurious effect due to the fact that the \( k = 0 \) mode of the non-Gaussian density field obtained from the pipeline, related to \( \langle \delta \rangle \), is not rigorously null.

The reason for such a difference comes from the fact that the DC mode, \( P_s(k = 0) \), when applying the pipeline represented by Eq. 4, is likely to be negative while we impose it to be null in order to make sure that the Gaussian field has a zero mean (the whole process relies on this assumption). This produces an undesired non-null \( k = 0 \) mode in the power spectrum of the
density field, $\delta$, which introduces a correlation between the volume averaged density, $\bar{\delta}$, with the power spectrum at all $k$ modes. Thus, the power spectrum at different wave mode being correlated with a hidden variable ($\bar{\delta}$), the covariance appears larger than expected. In this appendix, we describe this effect in detail and propose a correction.

A Monte Carlo realisation following the presented pipeline in Sect. 2 will be characterised by a finite total number of object $X$, which will vary from a realisation to another. In principle, the variance should be given by the Poisson distribution as $\sigma_{X}^{2} = N_{p} \rho_{0}$, where $\rho_{0} \equiv \langle X \rangle$ is the expectation value of the number of objects in the catalogues. Thus, the expectation value of the number of objects can be obtained by multiplying the volume $L^{3}$ by the expected number density $\rho_{0}$:

$$N_{p} = \rho_{0} L^{3}.$$  

When computing the cross-correlation, $C_{XP} \equiv \langle (X - N_{p}) (P(k) - \langle P(k) \rangle) \rangle$, between the number of object and the amplitude of the power spectrum at a given wave mode, $k$, we can show that it is given by:

$$C_{XP} = N_{p} k_{P} \bar{B}(0,k).$$  \hspace{1cm} \text{(B.1)}

In Eq. B.1, $\bar{B}(k_{1}, k_{2})$ stands for the shell average bi-spectrum, a quantity that is, in principle, expected to be null when taken at any $k_{1}$ or $k_{2}$ being 0. This is given by the fact that the mean number density is supposed to be known; thus, the mean density contrast is by definition zero and no power appears at any wave-modes $k = 0$.

As anticipated, if a negative value appears at the $k = 0$ mode, then it is enforced to be zero. It results in the $k = 0$ mode of the density power spectrum being expected to be larger than 0. The first effect of this is to provide an extra variance to the Poisson one, such that $\sigma_{X}^{2} = N_{p}^{2} (1 + k_{P}^{2} P_{\delta}(k = 0))$.

The second effect is to induce a non-negligible shell average bi-spectrum $\bar{B}(0,k)$. Indeed, at leading order, we can express $\bar{B}(0,k) \approx 2 c_{1} c_{2} P^{2}(k)$ with respect to the Hermite coefficients, $c_{1}$ and $c_{2}$, that are formally defined in Baratta et al. (2020). This shows that an extra correlation is introduced between the total number of object (or the mean density) and the power spectrum. After having tried to predict precisely the correlation coefficient (using the Mehler expansion of the bi-spectrum) in order to remove it from the computed covariance, we found that the following empirical formula works well and allows us to efficiently recover the correct covariance:

$$P_{\text{deb}}(k) = \left[ \frac{N_{p}}{X} \right]^{3} \left( \hat{P}(k) - P_{\text{SN}} \right).$$  \hspace{1cm} \text{(B.2)}

where $P_{\text{SN}} = (1/(2\pi))/\rho_{0}$ is the shot noise contribution to the power spectrum. The overall correction is just taking into account that the global effect of having more or less particles is to reshape the amplitude of the power spectrum. At the end of the day, the shot noise corrected power spectrum that will be used for covariance matrix estimation will therefore be $P_{\text{deb}}(k)$.