ConKer: An algorithm for evaluating correlations of arbitrary order

Z. Brown, G. Mishtaku, and R. Demina

Department of Physics and Astronomy, University of Rochester, 500 Joseph C. Wilson Boulevard, Rochester, NY 14627, USA
E-mail: zbrown5@ur.rochester.edu

Received 30 July 2021; accepted 21 July 2022

ABSTRACT

Context. High order correlations in the cosmic matter density have become increasingly valuable in cosmological analyses. However, computing these correlation functions is computationally expensive.

Aims. We aim to circumvent these challenges by developing a new algorithm called ConKer for estimating correlation functions.

Methods. This algorithm performs convolutions of matter distributions with spherical kernels using FFT. Since matter distributions and kernels are defined on a grid, it results in some loss of accuracy in the distance and angle definitions. We study the algorithm setting at which these limitations become critical and suggest ways to minimize them.

Results. ConKer is applied to the CMASS sample of the SDSS DR12 galaxy survey and corresponding mock catalogs, and is used to compute the correlation functions up to correlation order \( n = 5 \). We compare the \( n = 2 \) and \( n = 3 \) cases to traditional algorithms to verify the accuracy of the new algorithm. We perform a timing study of the algorithm and find that three of the four distinct processes within the algorithm are nearly independent of the catalog size \( N \), while one subdominant component scales as \( O(N) \). The dominant portion of the calculation has complexity of \( O(N^{1/2} \log N) \), where \( N \) is the number of cells in a three-dimensional grid corresponding to the matter density.

Conclusions. We find ConKer to be a fast and accurate method of probing high order correlations in the cosmic matter density, then discuss its application to upcoming surveys of large-scale structure.

Key words. cosmology: observations – large-scale structure of Universe – dark energy – dark matter – inflation

1. Introduction

Understanding the dynamics of inflation in the early universe is linked with the study of primordial density fluctuations, in particular with their deviations from a Gaussian distribution (see, e.g., Maldacena 2003; Bartolo et al. 2004; Acquaviva et al. 2003). High order correlations have been shown to be sensitive to non-Gaussian density fluctuations (see Meerburg et al. 2019 and references within). However, a brute force approach leads to prohibitively expensive \( O(N^n) \) computations of correlations, where \( N \) is the number of tracer objects and \( n \) is the correlation order. This problem has been studied extensively (March 2013). Several approaches to mitigate it were suggested for the calculation of three-point correlations; for example, Yuan et al. (2018) used the small angle assumption, while Slepian & Eisenstein (2015) evaluated the Legendre expansion of the three-point correlation function (3pcf). The second approach was recently generalized for \( n \)-point correlation functions (npdfs) in Philcox et al. (2021) resulting in an \( O(N^2) \) algorithm.

Here we present an alternative and computationally efficient way of evaluating these correlations. Similarly to March (2013), here this algorithm exploits spatial proximity, and similarly to Zhang & Yu (2011) and Slepian & Eisenstein (2016) it uses a fast Fourier transform (FFT) to speed up the calculation. These characteristics combined with implementation facilities help achieve a notable reduction in computational time and complexity.

The developed algorithm convolves kernels with matter maps, hence it is named ConKer\(^1\). It is an extension of the CenterFinder algorithm (Brown et al. 2021), designed to find locations in space likely to be the centers of the baryon acoustic oscillations (BAOs). CenterFinder counts the number of galaxies removed from a particular location by a given distance by convolving spherical kernels with the matter distribution. ConKer uses the same functionality to evaluate npdfs. A similar approach was suggested in Slepian & Eisenstein (2016), and developed to evaluate the Legendre expansion of the 3pcf for continuous matter tracers in (Portillo et al. 2018). In addition to implementing this approach for higher order correlations, ConKer introduces spatial partitioning defined with respect to the light of sight (LOS). This partitioning minimizes memory usage, allows for parallel computing, and enables an easy calculation of npdfs in the \( \mu \)-slices, where the angle \( \theta \) in the definition of \( \mu = \cos \theta \) is measured with respect to the LOS.

ConKer is applicable to discrete matter tracers, such as galaxies, and to continuous tracers, such as Lyman-\( \alpha \) and 21 cm line intensity, or matter maps derived from weak lenses. The method can be applied to evaluate autocorrelations, and cross-correlations between different matter tracers.

2. Algorithm description

2.1. Strategy

The two-point correlation can be visualized as an excess (or deficit) of sticks of a given length over a random combination of two points distributed over space. The three-point correlation corresponds to an excess of triangles, the four-point correlation to an excess of pyramids (the four points do not necessarily lie...
in one plane), and so on (see Fig. 1). We refer to these figures as n-plets. We consider all possible n-plets with one vertex at point 0, characterized by a vector \( \mathbf{r} \) with the other vertices defined by vectors \( \mathbf{r}_i \), \( (i = 1, \ldots, n-1) \). For each point we define a vector connecting it with point 0: \( \mathbf{s}_i = \mathbf{r}_i - \mathbf{r} \). We refer to a unit vector corresponding to any vector \( \mathbf{r} \) as \( \hat{r} \).

We let \( \rho(\mathbf{r}) \) be the density of the matter tracer (e.g., galaxy count per unit volume) at a location \( \mathbf{r} \), with \( \bar{\rho}(\mathbf{r}) \) being the density of expected observations from tracers randomly distributed over the surveyed volume. We define the deviation from the expected density as

\[
\Delta(\mathbf{r}) = \rho(\mathbf{r}) - \bar{\rho}(\mathbf{r}) .
\]

### 2.1.2. Legendre expansion

We let \( \theta_i \) indicate the angle between a vector \( \mathbf{s}_i \) and \( \mathbf{r} \) (see Fig. 2). We define the basis as a product of the Legendre polynomials, \( \Pi_{\ell}(\cos \theta_i) = P_{\ell_1}(\cos \theta_1)P_{\ell_2}(\cos \theta_2)\ldots P_{\ell_{n-1}}(\cos \theta_{n-1}) \), where \( \ell = (\ell_1, \ell_2, \ldots, \ell_{n-1}) \) represents orders in the Legendre expansion. The angular dependence of the npcf can be characterized via a decomposition in this basis:

\[
\xi_n(\mathbf{s}_1, \ldots, \mathbf{s}_{n-1}) = \sum_{\ell} \xi_n^{\ell}(\mathbf{s}_1, \ldots, \mathbf{s}_{n-1}) \Pi_{\ell}(\cos \theta_i) .
\]

The coefficients \( \xi_n^{\ell} \) in this expansion are functions of the distances \( \mathbf{s}_i \), but not the angles \( \theta_i \).

Following the example of Slepian & Eisenstein (2015) and its generalization Philcox et al. (2021), we use the spherical harmonic addition theorem:

\[
P_\ell(\cos \theta_i) = \frac{4\pi}{2\ell + 1} \sum_{m=-\ell}^{\ell} Y_{\ell m}^* (\hat{r}) Y_{\ell m}(\hat{s}_i) .
\]

The evaluation of \( \xi_n^{\ell} \) is then reduced to

\[
\xi_n^{\ell}(\mathbf{s}_1, \ldots, \mathbf{s}_{n-1}) = \frac{1}{R_0^3} \int d\mathbf{r} \Delta(\mathbf{r}) \sum_{m_1} \ldots \sum_{m_{n-1}} C_{\ell m_1 m_2 \ldots m_{n-1}} a_{m_1 m_2 \ldots m_{n-1}}(\mathbf{s}_1, \ldots, \mathbf{s}_{n-1}),
\]

where coupling coefficients \( C_{\ell m} \) with \( M = (m_1, m_2, \ldots, m_{n-1}) \) are defined in terms of Wigner 3-j symbols (see Appendix A). The values of \( m_i \) are scanned from \(-l_i\) to \(+l_i\). The calculation of the coefficients, \( a_{m_1 m_2 \ldots m_{n-1}}(\mathbf{r}, \mathbf{s}_i) \), implies an integration over all possible orientations of vector \( \hat{s}_i \).

It is equivalent to convolving the matter density with a sphere of radius \( s_i \) centered on point 0 and populated with the values of \( Y_{\ell m} \).
2.1.3. Edge correction

Irregular survey boundaries and nonuniformities in the redshift selection function can introduce anisotropies in an otherwise isotropic distribution. The formalism to correct for the edge effects developed in Slepian & Eisenstein (2015) and Philcox et al. (2021) is ideal for implementation using the kernel convolution functionality. The procedure involves the evaluation of Legendre moments of the random distribution \( f_R^j \) according to Eqs. 6 and 7 with the \( \hat{p} \)-field used instead of \( \Delta \). In ConKer this is realized by convolving spherical kernels populated with the values of \( \Delta y_m \) with the random distribution of tracers.

A set of edge-corrected Legendre coefficients \( \hat{c}_{LM} \) are calculated based on uncorrected \( c^L_n \) evaluated using Eq. 6 and coefficients \( f^L_R \) by solving a system of linear equations

\[
\hat{c}^L_n = M_{LL}^{−1} c^L_n ,
\]

where the matrix \( M_{LL} \) is defined as

\[
M_{LL} = \sum_k (-1)^k G_{LKN} f_k .
\]

For the definition of the Gaunt integral \( G_{LKN} \), see Appendix A.

2.2. Input

The inputs to the algorithm are catalogs of the observed number count of tracers \( D \), with the total number being \( N_D \), and \( R \), which represents a number count of randomly distributed points within the same fiducial volume and corresponding number count \( N_R \). Most surveys provide the angular coordinates: right ascension \( \alpha \) and declination \( \delta \), and the redshift \( z \) of each tracer. The relationship between the redshift and the comoving radial distance is cosmology dependent:

\[
r(z) = \frac{c}{H_0} \int_0^z \frac{dz'}{\sqrt{\Omega_M (z' + 1)^3 + \Omega_k (z' + 1)^2 + \Omega_\Lambda}} .
\]

Here \( \Omega_M \), \( \Omega_k \), and \( \Omega_\Lambda \) are the relative present-day matter, curvature, and cosmological constant densities, respectively; \( H_0 \) is the present-day Hubble constant; and \( c \) is the speed of light. These user-defined parameters represent the fiducial cosmology. In this study we used the following values for the cosmological parameters: \( c = 300000 \) km/s, \( H_0 = 100h \) km/s/Mpc, \( \Omega_M = 0.29 \), \( \Omega_k = 0.71 \), and \( \Omega_\Lambda = 0 \). The integral in Eq. 10 is evaluated numerically in ConKer. Cartesian coordinates of a tracer labeled \((X,Y,Z)\) are evaluated based on \( r \), \( \alpha \), and \( \delta \).

The algorithm computes correlation functions over a given range of scales or separation distances \( s \). The range of distances from \( s_{\text{min}} \) to \( s_{\text{max}} \) is divided into \( N_s \) bins, which sets the bin width \( \Delta s \).

2.3. Partitioning and mapping

The initial step of the algorithm divides data and random catalogs into partitions based on the angular variables \((\alpha, \delta)\). Each partition spans the entire range of redshifts from \( z_{\text{min}} \) to \( z_{\text{max}} \) corresponding to the co-moving radii \( r_{\text{min}} \) and \( r_{\text{max}} \), evaluated according to Eq. 10. The angular size \( \theta_p \) is determined by the angle subtended by \( s_{\text{max}} \) at \( r_{\text{min}} \):

\[
\theta_p = \frac{s_{\text{max}}}{r_{\text{min}}} .
\]

Each \((jk)^{th}\) partition is populated by galaxies with angular coordinates within the following limits:

\[
\alpha_{j-1} \rightarrow \alpha_j = \alpha_{j-1} + \frac{2\theta_p}{\min (\cos \delta_j)} ,
\]

\[
\delta_{k-1} \rightarrow \delta_k = \delta_{k-1} + 2\theta_p .
\]

Here \( \min (\cos \delta_j) \) is the minimum value of \( \cos \delta_j \) in this partition. This factor is introduced for each region to have an approximately square span of \( 2\delta_{\text{max}} \) in the azimuthal and polar directions at the smallest comoving radius. The lowest boundaries are determined by the survey coverage.

The definition of the Cartesian \((\hat{x}, \hat{y}, \hat{z})\) system is unique to each partition with the \( \hat{x} \)-axis pointing to its center cell. The transformation from global sky to local Cartesian coordinates is given in Appendix B.

The LOS in each partition is defined as pointing along the \( \hat{x} \)-axis. Having the same definition of the LOS for all the objects in the partition introduces some inaccuracy in angles, especially for objects near the boundary. However, the maximum deviation in the LOS definition is on the order of \( \theta_p^2/2 \). Hence, this inaccuracy can be minimized by the proper choice of the partition size \( \theta_p \).

In each partition we define a grid with spacing \( g_z \), such that the volume of each cubic grid cell is \( g_z^3 \). The default value of \( g_z \) is set to be equal to the bin width \( \Delta s \). However, the user can select a finer resolution in the density field and kernels. In this case, \( g_z \) is set to a desired fraction of the radial bin size \( \Delta s \). During the final steps of the algorithm, correlation functions are resampled to the appropriate \( s \)-bins.

On the grid we define three-dimensional histograms, \( D(X,Y,Z) \) and \( R(X,Y,Z) \), which represent tracer counts in the cell \((X,Y,Z)\) from data and random catalogs, respectively. These histograms may be populated by the raw count or the weighted count of tracers from the input catalogs.

In every partition two additional grids are constructed, \( D_{\text{MP}}(X,Y,Z) \) and \( R_{\text{MP}}(X,Y,Z) \), which contain an extended map of objects within an additional \( \theta_p \) in the declination direction and an additional \( \theta_p/\min (\cos \delta_k) \) in the right ascension direction of the LOS. During the convolution step the center of the kernel is placed on each cell of \( D(X,Y,Z) \) and \( R(X,Y,Z) \), while the convolution is performed with the extended maps \( D_{\text{MP}}(X,Y,Z) \) and \( R_{\text{MP}}(X,Y,Z) \). This procedure ensures that the entire survey region is covered, but that double-counting is avoided.

A three-dimensional local density variation histogram \( N(X,Y,Z) \), which is a discretized representation of the \( \Delta \) field, is defined on the grid to represent the difference in counts between \( D \) and \( R \), normalized to \( N_D \):

\[
N(X,Y,Z) = D(X,Y,Z) - R(X,Y,Z) .
\]

A similar field \( N_{\text{MP}}(X,Y,Z) \) is defined using the extended grids.

The default mass assignment scheme used in ConKer is a three-dimensional histogram, or nearest grid point (NGP) method. However, Jing (2005) and Cui et al. (2008) showed that the galaxy power spectrum measured using \( FFT \) algorithms is sensitive to the choice of mass assignment scheme. The ConKer algorithm includes the option to use the cloud in cell (CIC) method when defining the density fields. Of the two methods, CIC is more computationally expensive since it maps each tracer to multiple grid cells. The stage of the algorithm that places matter tracers in grid cells is referred to as mapping.

A three-dimensional local density variation histogram \( N(X,Y,Z) \), which is a discretized representation of the \( \Delta \) field, is defined on the grid to represent the difference in counts between \( D \) and \( R \), normalized to \( N_D \):

\[
N(X,Y,Z) = D(X,Y,Z) - R(X,Y,Z) .
\]

A similar field \( N_{\text{MP}}(X,Y,Z) \) is defined using the extended grids.

The default mass assignment scheme used in ConKer is a three-dimensional histogram, or nearest grid point (NGP) method. However, Jing (2005) and Cui et al. (2008) showed that the galaxy power spectrum measured using \( FFT \) algorithms is sensitive to the choice of mass assignment scheme. The ConKer algorithm includes the option to use the cloud in cell (CIC) method when defining the density fields. Of the two methods, CIC is more computationally expensive since it maps each tracer to multiple grid cells. The stage of the algorithm that places matter tracers in grid cells is referred to as mapping.
average of positions within the each cell. We refer to this kernel as "flat."

If the user has elected to employ a grid resolution $g_S$ finer than the desired radial binning, an integer number of extra kernels are constructed -bin, where $s$ denotes a three-dimensional discrete convolution performed using FFT and $W^{lm}_{ij}$ is a discretized representation of coefficients $a_{lm}$ calculated according to Eq. 7. The procedure is performed in each partition, for each bin in $s$ and for all possible values of $l$ and $m$ according to a given $l_{\text{max}}$, resulting in $2(l_{\text{max}}+1)^2N_b$ convolutions. Once these maps of $W^{lm}_{ij}(X, Y, Z)$ are created, they can be used to calculate the correlation functions of arbitrary order $n$.

For normalization purposes we perform the same procedure on the field of random counts. The result of the convolution of random density field $R_{\text{MP}}$ with kernel $K^{lm}_i$ is referred to as $B^{lm}_i(X, Y, Z)$:

$$B^{lm}_i(X, Y, Z) \equiv R_{\text{MP}}(X, Y, Z) * \ast * K^{lm}_i(X, Y, Z) .$$  \hspace{1cm} (16)
3.2. Evaluation of \( n \)-point correlation functions

For each \((jk)\)th partition we concatenate every grid \( W_{lm}^{m} \) and \( B_{lm}^{m} \) such that they now represent a convolution of the kernel with the entire density field. A map of \( W_{lm}^{m} \) represents discretized coefficients \( a_{lm}^{m} \) in Eq. 6. According to this equation to evaluate coefficients \( \xi_{n}^{L} \) these maps must be convolved with the matter distribution \( \Delta(r) \), a discretized representation of which is \( N(X,Y,Z) \), and summed over the entire grid. Hence,

\[
\xi_{n}^{L}(s_{1}, \ldots s_{(n-1)}) = \frac{\sum_{grid} W_{0} \sum_{m_{1}} \ldots \sum_{m_{n-i}} C_{LML} W_{1}^{m_{1}} \ldots W_{n-1}^{m_{n-i}}}{\sum_{grid} B_{0}^{m_{1}} B_{0}^{m_{n-1}}}.
\]

where \( W_{0} = N(X,Y,Z) \) and \( B_{0} = R(X,Y,Z) \). This step in the algorithm is referred to as summation.

3.2.1. Cases

The specific case of \( l \)-th multipole of the 2pcf is calculated as

\[
\xi_{2}^{L}(s) = \frac{1}{\sqrt{2l + 1}} \frac{\sum_{grid} W_{0} W_{0}^{l}}{\sum_{grid} B_{0} B_{0}^{l}}.
\]

The \( L \)-th multipole of the 3pcf is then calculated as

\[
\xi_{3}^{L}(s_{1}, s_{2}) = \frac{4\pi}{\sqrt{2l + 1}} \frac{\sum_{grid} W_{0} W_{0}^{l}}{\sum_{grid} B_{0} B_{0}^{l}}.
\]

This procedure is repeated on the random field in order to construct the terms necessary for edge-correction (see Philcox et al. (2021)).

At the summation stage, computing the \( npcf \) requires that the already computed grids defined by \( s, l, \) and \( m \) be combined and summed over. The execution time of this step is about two orders of magnitude lower than the convolution step in the case of the 3pcf, and trivial in the case of the 2pcf. Only at very large correlation orders does the execution time of the summation process begin to approach that of convolution.

Of particular interest are correlations between objects with a defined scale, such as those arising from spherical sound waves in the primordial plasma, also known as BAOs. In this case one tracer taken as a starting point is displaced from the other \((n-1)\) points by the same distance \( s_{1} \) (e.g., the three-point correlation corresponds to isosceles triangles randomly distributed in space). This equidistant case corresponds to the diagonal of the \( n \)-point correlation function (hence referred to as the diagonal \( npcf \)) for \( l = 0 \), which is calculated as

\[
\xi_{0,n}^{\text{diag}}(s) = \frac{\sum_{grid} W_{0}(W_{0}^{0})^{n-1}}{\sum_{grid} B_{0}(B_{0}^{0})^{n-1}}.
\]

One of the advantages of the ConKer algorithm is that regardless of the desired correlation order, \( n \), no new convolution operations need take place. Thus, the time-consuming step is only performed once per catalog, which facilitates subsequent calculations of correlation functions to arbitrary order, \( n \).

3.8. Procedure

A schematic flowchart of the ConKer algorithm is shown in Fig. 5. Beginning with a definition of the cosmological parameters and binning, the partitioning is performed on the random catalog. If that catalog has already been used, the previous partitioning scheme is employed. During the convolution step, the user can choose whether or not to perform the convolution with the catalog of random tracers for edge correction. It is more time consuming, but only needs to be performed once per random catalog. During the summation step the user is able to compute an arbitrary number of correlation functions of a desired order \( n \) and \( l_{\text{max}} \). Since files corresponding to the convolved grids are often large, the user can choose to delete them upon completion of the summation step.

If the user wishes to employ the \( \mu \)-wedge kernels, the procedure is nearly identical; however, each step of the calculation must be repeated for each slice. This increases the relevant computational parameters such as run time and file sizes, but does not affect memory considerations as the calculations in each slice are performed independently.

3. Performance study

We evaluated the performance of ConKer using SDSS DR12 CMASS galaxies (Ross et al. 2017), their associated random catalogs, and an ensemble of MultiDark-Patchy mocks (Kitaura et al. 2016; Rodríguez-Torres et al. 2016). We applied ConKer to the SGC and NGC catalogs for data, randoms, and 20 mocks. For this study, which highlights the algorithm’s ability to probe correlations near the clustering and the BAO scales, we computed correlation functions for a distance range of 8–176 h\(^{-1}\)Mpc in 21 bins of width of 8 h\(^{-1}\)Mpc. In all cases, the standard systematic (Ross et al. 2017) as well as FKP weights (Feldman et al. 1993) were used to create the density field map. The default NGP mass assignment scheme was used and the grid spacing of \( g_{s} = 8 h^{-1}\)Mpc unless specified otherwise.
3.1. Timing study

We demonstrate the efficient nature of our algorithm with the following timing study, performed using a personal computer with a 10 CPU core Apple M1 Pro chip and 32 GB of memory. All execution times are in units of CPU seconds.

The primary advantage of ConKer is in the behavior of the execution time as a function of the total number of objects \( N = N_D + N_R \) shown in top plot of Fig. 6 for the four stages of the algorithm: mapping, convolution, file operations, and summation. The surveyed volume, \( V \) is kept fixed. As expected, the execution time of convolution, file operations, and summation are nearly independent of \( N \), and all three scale below \( O(N^{1/2}) \). Mapping is an \( O(N) \) calculation, and only starts to dominate for catalogs significantly larger than 100M objects.

The main parameter that determines the computation time of ConKer is the grid spacing \( g_s \), which is set by default to be equal to the bin width \( \Delta s \). The maximum distance probed \( s_{\text{max}} \) determines the number of steps in \( s : N_h = (s_{\text{max}} - s_{\text{min}})/\Delta s \). The total number of grid cells \( N_c \) depends on the surveyed volume: \( N_c = V/g_s^3 \). Mapping is independent of \( g_s \). For \( g_s \) below approximately 10 h^{-1} Mpc, the dominant process is convolution of cubic volumes, on which the kernels are defined, containing \( N_c \) cells: \( N_k = (s/g_s)^3 < (s_{\text{max}}/g_s)^3 \). It is repeated \( N_c \) times with each grid cell being the center of the kernel. Since the convolution is performed using FFT with a typical complexity of \( N \log N \), the complexity of each step in \( s \) is \( O(N_c \log N_c) \). Thus, the time complexity of the convolution is

\[
 t_{\text{conv}} \propto N_h N_c \log N_k < \frac{s_{\text{max}}}{g_s} \left( \frac{V}{g_s^3} \right) \log \left( \frac{V}{g_s^3} \right) \propto g_s^{-4/3} \log g_s \propto N_c^{4/3} \log N_c .
\]

(22)

The observed scaling of the convolution step as \( g_s^{-4/3} \) is in good agreement with this analytic prediction as depicted by the solid brown line in Fig. 6 (bottom).

The file operations step scales more favorably with \( g_s \); however, it dominates the execution time for \( g_s \) above approximately 5 h^{-1} Mpc.

3.2. Comparison to existing methods

We present comparisons of the 2pcf and 3pcf evaluated using ConKer to well-established methods. In Cuesta et al. (2016) the monopole and quadrupole terms of the Landy and Szalay estimator of the 2pcf (Landy & Szalay 1993; Hamilton 1993) are computed for the combined SGC and NGC catalogs of the SDSS DR12 CMASS survey galaxies. We performed the calculation using ConKer with the same binning, and compared it to the Cuesta et al. (2016) results in Fig. 7. For the monopole and the quadrupole terms, we note good agreement between the two methods. At low scales (\( s < 30 h^{-1} \text{Mpc} \)), we find the largest deviation between the two. The differences expectedly arise due to the discretization of the density field and kernel in ConKer. The kernel represents a spherical shell of width \( g_s \) mapped onto a three-dimensional Cartesian grid. Thus, once the kernel size becomes comparable with the grid spacing a resolution in the distance determination is degraded. This occurs, when the kernel size is less than approximately 5\( g_s \). This does not mean, however, that we are unable to probe correlations at small scales. Instead, this simply requires a finer grid, and resampling. By reducing the grid spacing (blue and red points in Fig. 7) we recovered the agreement down to lower scales. Based on this, we recommend setting the \( s_{\text{max}} \) parameter larger than 5\( \Delta s \) if using the default sampling. Any differences in the size of the errors results from the fact that we use two separate mock ensembles to estimate the covariance.
In addition to the Legendre expansion, we also compute the 2pcf of the NGC sample in two $\mu$-slices, corresponding to the transverse ($\xi_\perp$) and parallel ($\xi_\parallel$) cases. To compare, we repeated the calculation using nbodykit, an open source cosmology toolkit (Hand 2018). The results are shown in Fig. 8. We find the same behavior at small scales as in the case of the Legendre expansion, where the agreement is recovered by reducing the grid spacing.

In Fig. 7 and in Fig. 8 comparing the $l = 0, 2$ and $\xi_{\perp,\parallel}$ cases to existing clustering algorithms, the uncertainties on the survey data measurements were derived from the mock ensemble. The size of the error bar corresponding to point-i is $\sigma_i = \sqrt{C_{ii}}$, where $C$ is the mock covariance matrix.

The 3pcf algorithm implemented in nbodykit is based on the work of Slepian & Eisenstein (2015). Using the two methods, we computed the edge-corrected 3pcf up to $l = 3$ of the same subsample of NGC galaxies used in the timing study (see caption of Fig. 6 for details). For this comparison, we computed 3pcf in 11 bins of width of 10 $h^{-1}$Mpc for $s$ from 45 to 155 $h^{-1}$Mpc. Over this range of scales, we find a good agreement between the two methods. More importantly, the distribution over the residuals is centered at approximately zero, meaning our estimator is not biased compared to the nbodykit implementation. The largest deviations between the algorithms again arise at smaller scales where the kernel resolution is degraded. The 3pcf calculation using ConKer was faster by a factor of $\sim 3$, and scales more favorably with $N$, since nbodykit is an $O(N^2)$ algorithm.
A & A proofs: manuscript no. 41917corr_2column

Fig. 8. Points showing the transverse $\xi_\perp(s)$ and parallel $\xi_\parallel(s)$ components of the 2pcf computed using ConKer with three values of grid spacing $g_S = 8$ (red–yellow), 4 (purple–green), 2 (black–gray) $h^{-1}$ Mpc. Solid lines in the corresponding colors show the results from the nbodykit 2pcf algorithm. This figure corresponds only to NGC galaxies. The error bars and shaded regions are determined from the ensemble of Patchy mocks. The lower subpanels show the residual between the two methods.

3.3. ConKer npcf

We used ConKer to compute the diagonal elements of the npcf as defined in Eq. 21 for $n = 2, 3, 4, 5$ and $l = 0$, and off-diagonal elements of the 3pcf for an ensemble of MultiDark-Patchy mocks. For this calculation, the NGC/SGC catalogs were divided into 27/14 partitions, as shown in Fig. 10.

The diagonal elements of the npcf are shown in Fig. 11 for $n = 2, 3, 4, 5$. We observe the expected features of the npcf. These include an increase in magnitude at small scales present for $n=2$ and $n=3$, corresponding to galaxy clustering, and a well-defined “bump” at the BAO scale for all. The npcf based on the random catalog fluctuate about $\xi_n = 0$ at several orders of magnitude below the signal observed in mocks.

In each case, the covariance matrix of the diagonal npcf is estimated on the set of 20 Patchy mocks using the following procedure. For the $i$-th bin of the npcf, we let the average value across the mock ensemble be defined as $\langle \xi_i \rangle$. If the number of mocks in the ensemble is $N_q$, then the covariance matrix elements are

$$C_{ij} = \frac{1}{N_q - 1} \sum_q \left( \xi_{i,q}^{\text{nd}}(s_j) - \langle \xi_i \rangle \right) \left( \xi_{j,q}^{\text{nd}}(s_i) - \langle \xi_j \rangle \right).$$

We show an example of the reduced covariance matrix for the $n=2$, $l = 0$ diagonal case in Fig. 12.

Off-diagonal elements of the edge-corrected 3pcf are shown in Fig. 13 for the NGC sample. Emphasizing for large-scale features, we observe strong indicators of BAO in both the $l = 0$ and $l = 1$ cases.

4. Discussion

4.1. Algorithmic framework

The query for a pattern in matter distributions may prompt the employment of machine learning techniques. ConKer, being a spatial statistics algorithm, offers an alternative to such an approach that is fast and transparent. It exploits the fact that the full set of equidistant points from any given point makes a sphere, with its surface density being a direct measure of how spherically structured this subspace of points is. Aggregating and combining this measure over the whole space of $N$ objects allows us to calculate the space’s $n$-point correlation function.
The algorithm exploits an intrinsic spatial proximity characteristic in the objective of querying structures of negligible dimensions in a much bigger space. This spatial proximity factor leads to space partitioning algorithms targeting a nearest neighbor query approach (see, e.g., the tree-based npcf algorithms in March 2013). However, ConKer uses this factor as a heuristic in limiting its query space immediately to only the defined separation for each point in the space. We note how this is in contrast with the former technique family. In a nearest neighbor approach to the npcf problem, the query space is grown at each point in the greater embedding space, aggregating the n-point statistic until the greater space is fully queried, whereas ConKer aggregates the statistic over all embedding space, and then grows the query space before repeating.

This design choice realized in ConKer's core subroutine, a convolution of the query space with the embedding space performed by an FFT algorithm, distills the complexity from the brute force. This approach, combined with the heuristic above, lets the dominant components of ConKer achieve independence of the number of objects, as shown in Fig. 6. There is certainly a trade-off between the sparsity of the whole space and the bias toward linear complexity in number of objects, as expected from an FFT-based algorithm, but even for very dense catalogs, we expect the scaling in the number of objects to be capped by $O(N)$, where $N$ is the total number of objects. Ultimately, ConKer is a hybrid algorithm that draws from both computational geometry and signal processing to achieve linear complexity in the number of objects.
4.2. ConKer versus other methods

The idea of convolving spherically symmetric kernels with the density fields to evaluate the number of objects removed from a certain point by a given distance was originally proposed in Zhang & Yu (2011). In this work the Legendre expansion of \( npcf \) was not considered. In Sleipn & Eisenstein (2015) and Philcox et al. (2021) a KDTree algorithm was used and the spherical function decomposition was evaluated for each galaxy pair, resulting in an \( O(N^2) \) calculation. This method works well for smaller scales or sparse surveys. In the same papers using FFT-based convolution for the Legendre expansion was also suggested. This approach has an advantage for denser surveys or continuous tracers since the computational time depends on the volume but not on density. The idea was later realized in (Portillo et al. 2018) for the evaluation of the 3pcf of the continuous tracer.

ConKer extends the approach to all \( n \geq 1 \). ConKer computes the integral in Eq. 7 by convolving a spherical kernel \( K_i \) of radius \( s_i \) populated with the values of \( Y_{\text{lin}} \) with the matter density field. The definition of a kernel on the grid necessarily leads to some loss of precision in the distance definition. This is partially mitigated in ConKer by weighting the grid cells with the fraction of the cell’s volume contained within a given spherical shell and averaging \( Y_{\text{lin}} \)’s over this volume. Convolving the entire surveyed volume (as in Portillo et al. (2018)) leads to large arrays that need to be stored resulting in significant memory requirements. In light of anticipated large volume surveys such as DESI (DESI Collaboration et al. 2016), this limitation becomes particularly stringent. ConKer convolves a cubic volume just large enough to encompass a sphere of the specified radius, thus limiting the memory requirements for kernel storage. Additionally, ConKer introduces a partitioning scheme, as discussed in Sect. 2.3. As a result, the array size is limited by the partition’s volume. This scheme has an additional benefit; it allows for the evaluation of \( npcf \) in \( \mu \)-slices as, discussed in Sect. 2.4.2, which is particularly relevant for parallel versus transverse to the LOS analysis. Finally, partitioning naturally allows for parallelized computing processes.

4.3. Applications beyond correlation functions

Though traditionally the order of correlation \( n \) is viewed as an important parameter, for the diagonal \( npcf \), all the information is entirely encoded by the weights \( W_0 \) and \( W_1 \). It was pointed out in Carron & Neyrinck (2012) that the \( npcf \) is inadequate in capturing the tails of non-Gaussianities. The distributions over \( W_0 \) and \( W_1 \) (as opposed to their sum over the sample, as is used in the \( npcf \)) could recover that sensitivity, which is a subject for future studies. The distribution of the product of two weights, \( W_0 W_1 \) normalized by the average \( B_i B_j \), is presented in Fig. 14 for a kernel size of 108 \( h^{-1}\text{Mpc} \) for data, mock, and random catalogs.

Moreover, \( W_0 \) and \( W_1 \) as well as their product are maps. While in the \( npcf \) the location information is entirely lost, in the maps produced by ConKer it is preserved and can be used for cross-correlation studies between different tracers, such as weak lensing, \( \text{Ly} \alpha \), and CMB.

5. Conclusion

We presented ConKer, an algorithm that convolves spherical kernels with matter maps allowing for fast evaluation of the \( n \)-point correlation functions, its expansion in Legendre polynomials, and its \( \mu \)-slices. The algorithm can be broken into three stages: mapping, convolution, and summation. The execution time of convolution and summation are independent of the catalog size \( N \), while mapping is an \( O(N) \) calculation, which starts dominating for catalogs larger than 100M objects. The dominant part of the convolution is with complexity \( O(Nc^{4/3}) \log N_c \), where \( N_c \) is the number of grid cells.

A comparison to the standard techniques shows good agreement. We study the performance using SDSS DR12 CMASS galaxies, their associated random catalogs, and an ensemble of MultiDark-Patchy mocks. The results up to \( n = 5 \) are presented.
Further metrics that may offer additional sensitivity to primordial non-Gaussianities are also suggested such as the distribution over weights $W_i$ and their products.

Acknowledgements. The authors would like to thank Z. Slepian for his interest and insightful comments, F. Weisenhorn for assistance with the nbodykit 2pcf calculations, A Ross for providing numeric values of SDSS DR12 CMASS 2pcf, and S. BenZvi, K. Douglass and S. Gontcho A Gontcho for useful discussions. The authors acknowledge support from the U.S. Department of Energy under the grant DE-SC0008475. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University. The massive production of all MultiDark-Patchy mocks for the BOSS Final Data Release has been performed at the BSC Marenostrum supercomputer, the Hydra cluster at the Instituto de Física Teorica UAM/CSIC, and NERSC at the Lawrence Berkeley National Laboratory. We acknowledge support from the Spanish MICINNs Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064, MINECO Centro de Excelencia Severo Ochoa Programme under grant SEV-2012-0264, and grant AYA2014-60641-C2-1-P. The MultiDark-Patchy mocks was an effort led from the IFT UAM-CSIC by F. Prada’s group (C.-H. Chuang, S. Rodríguez-Torres and C. Scoccola) in collaboration with C. Zhao (Tsinghua U.), F.-S. Kitaura (AIP), A. Klypin (NMSU), G. Yepes (UAM), and the BOSS galaxy clustering working group.

References

Hand, N. e. a. 2018, Astron.J., 4, 156
Malda,ca, J. M. 2003, JHEP, 05, 013
March, W. B. 2013, PhD thesis, Georgia Institute of Technology
Appendix A: Useful formulae

For \( L = (l_1, l_2, \ldots, l_{n-1}) \) and \( M = (m_1, m_2, \ldots, m_{n-1}) \) with each \(-l_i \leq m_i \leq l_i\), we can define the coupling coefficient \( C_{LM} \) in terms of Wigner 3-j symbols (3x2 matrices) as

\[
C_{LM} = (-1)^{L+M} \sqrt{2l_1 + 1} \sqrt{2l_2 + 1} \sqrt{2l_3 + 1} \sum_{m_{L_1} \ldots m_{L_3}} (-1)^{L_1} \begin{pmatrix}
    l_1 & l_2 & l_{L_1} \\
    m_1 & m_2 & m_{L_1}
\end{pmatrix}
\begin{pmatrix}
    l_2 & l_3 & l_{L_2} \\
    m_2 & m_3 & m_{L_2}
\end{pmatrix}
\begin{pmatrix}
    l_3 & l_{L_3} \\
    m_3 & m_{L_3}
\end{pmatrix},
\]

where \( \kappa = l_1 + m_1 + l_2 + m_2 + \ldots, l_{n-1} + m_{n-1} \).

The Gaunt integral \( G_{L1L2L3} \), used in the edge correction procedure is defined as

\[
G_{L1L2L3} = \int d\Omega L_i(\hat{r}) \Pi_{L_1} (\hat{r}) \Pi_{L_2} (\hat{r}) \, (A.2)
\]

For \( n = 3 \) it is

\[
G_{l_0 l_1 l_2} = \frac{1}{4\pi} \sqrt{2l_1 + 1} \sqrt{2l_2 + 1} \sqrt{2l_3 + 1} \begin{pmatrix} l_1 & l_2 & l_3 \end{pmatrix}^2,
\]

for \( n = 4 \)

\[
G_{L'LL''} = \frac{1}{(4\pi)^{3/2}} \left[ \prod_{i=1}^{3} \sqrt{2l_{i1} + 1} \sqrt{2l_{i2} + 1} \right] \begin{pmatrix}
    l_1 & l_2 & l_3 \\
    0 & 0 & 0
\end{pmatrix} \begin{pmatrix}
    l_1' & l_2' & l_3' \\
    0 & 0 & 0
\end{pmatrix},
\]

for \( n = 5 \)

\[
G_{L'LL''} = \frac{1}{(4\pi)^2} \left[ \prod_{i=1}^{4} \sqrt{2l_{i1} + 1} \sqrt{2l_{i2} + 1} \right] \begin{pmatrix}
    l_1 & l_2 & l_3 & l_{L_1} \\
    0 & 0 & 0 & 0
\end{pmatrix} \begin{pmatrix}
    l_1' & l_2' & l_3' \\
    0 & 0 & 0
\end{pmatrix} \begin{pmatrix}
    l_1'' & l_2'' & l_3'' \\
    0 & 0 & 0
\end{pmatrix},
\]

Here three-dimensional matrices represent Wigner 9-j symbols.

Appendix B: Coordinate transformation

A global coordinate system is defined so that the \( \hat{z} \)-axis is pointing to the zenith, corresponding to declination angle \( \delta = 90^\circ = \pi/2 \). For each partition we define the local coordinate system

\[
n'_{\delta} = \cos \delta \cos (\alpha - \alpha_{LOS}),
\]

\[
n'_{\delta} = \cos \delta \sin (\alpha - \alpha_{LOS}),
\]

\[
n'_{\delta} = \sin \delta.
\]

Rotation by the polar angle \( \delta_{LOS} \) defines coordinate system \( F'' \), in which the galaxy’s unit vector is

\[
n''_{\delta} = n''_{\delta} \cos \delta_{LOS} + n''_{\delta} \sin \delta_{LOS},
\]

\[
n''_{\delta} = n''_{\delta},
\]

\[
n''_{\delta} = -n''_{\delta} \sin \delta_{LOS} + n''_{\delta} \cos \delta_{LOS}.
\]

In coordinate system \( F'' \), the \( \hat{x}'' \)-axis is directed along the LOS. The angular coordinates in \( F'' \) are

\[
\alpha'' = \alpha - \alpha_{LOS},
\]

\[
\delta'' = \sin^{-1}(n''_{\delta}).
\]

Thus, Cartesian coordinates in the local coordinate system are defined as

\[
X = r(z)n''_{\delta},
\]

\[
Y = r(z)n''_{\delta},
\]

\[
Z = r(z)n''_{\delta}.
\]