On the fast random sampling and other properties of the three point correlation function in galaxy surveys

Fidel Sosa Nuñez and Gustavo Niz

Departamento de Física, Universidad de Guanajuato, Loma del Bosque No. 103 Col. Lomas del Campestre, C.P 37150 León, Guanajuato, México
E-mail: f.sosanunez@ugto.mx, g.niz@ugto.mx

Received June 18, 2020
Accepted October 20, 2020
Published December 10, 2020

Abstract. In the forthcoming large volume galaxy surveys, the use of higher order statistics will prove important to obtain complementary information to the usual two point statistics. In particular, one of those higher order statistical tools is the Three Point Correlation Function (3PCF) over discrete data points, whose lowest variance estimators count triangle configurations with vertices mixing data and random catalogues. A popular choice is to use large density random catalogues, which reduces the shot noise but leads to a computational cost of one or two orders of magnitude more than the pure data histogram calculation. In this paper, we explore ideas to time reduce the random sampling without using random catalogues. We focus on the isotropic 3PCF case over periodic boxes. In a first approach, based on Hamilton’s construction of his famous two point estimator, we use an ad-hoc two point correlation piece for the mixed random-data histograms. A second approach relies on operators constructed from a geometrical viewpoint, using two sides and one angle to define the three dependencies of the isotropic 3PCF. We map the last result to the three triangle side basis either numerically or analytically, and show that the latter method performs best when applied to synthetic data. In addition, we elaborate on relaxing the no boundary condition, discuss other low variance n-point estimators and present useful 3PCF visualization schemes.

Keywords: cluster counts, redshift surveys, semi-analytic modeling

ArXiv ePrint: 2006.05434
1 Introduction and methodology

A detailed understanding of how matter distributes in the Universe is a key to untangle some of the greatest puzzles of modern cosmology, such as the nature of dark energy, dark matter or the physics from the Early Universe. The two point correlation function (2PCF) or its Fourier-space counterpart, the power spectrum, has proven very successful to reconstruct this matter distribution. However, as new experiments escalate the amount of cosmological data with an increasing precision, additional and useful information could be extracted from higher order point statistics. Moreover, primordial non-Gaussianities, the nature of the gravitational interactions, the scale and running of the DM-galaxy biases or any other physics of the standard cosmological model and its extensions would strongly benefit from these higher order statistical tools, in particular using the next level in the hierarchy: the three point correlation function (3PCF), or its Fourier counterpart, the bispectrum.

The 3PCF in galaxy surveys dates back to work of Peebles and Groth in the seventies [1, 2], and since then numerous studies and theoretical improvements have been developed (see for example the reviews of [3] and [4]). The interest on the higher-order statistics has substantially increased in the last two decades, as can be appreciated by the numerous
studies on the SDSS data or other experiments (see for example [5–15], or some of the more recent studies [16–21]).

Different estimators of the underlying true 3PCF have been studied, such as the Szapudi and Szalay (SS) [22] or Jing and Börner (JB) [14], but insufficient work has been done on exploring further possibilities as in the two point case. In particular, the SS estimator is the 3pt equivalent of the well known minimal-variance two point estimator of Landy-Szalay (LS) [23]. Although the three point structure is richer than its two point counterpart, the SS is also expected to be a good minimal variance estimator, and hence its popular adoption. We based our findings on the SS estimator but they can be easily obtained for other 3PCF estimators.

The use of \( n \)-point correlation functions comes with a price: the observational noise and the computational cost are bigger as \( n \) increases. New experiments, such as the stage IV galaxy surveys DESI [24], LSST [25], Euclid [26], or WFIRST [27], will achieve better resolution and larger volumes, increasing the signal to noise ratio for higher order statistics. For the 3PCF in particular, in-cell counts which benefit from organised distance counting using Kd-trees (see for example [28]) or other techniques, with the usual (two point) random sampling, are not fast enough to study the large scale 3PCF for these new volume experiments. Therefore, improvements to reduce the computational time of these naive algorithms are needed,\(^1\) such as the recent work using a multipole basis decomposition [34] or the efficient random counting for periodic boxes [35]. Our independently developed work extends on the latter using two different approaches. (See a similar discussion in Fourier space which also uses the multipole decomposition [36, 37]).

In the case of discrete galaxy catalogues, 3PCF estimators usually count triplets between data and random catalogues. The mixed counts take one or two vertices in the data catalogue and the remaining vertex or vertices of the triangle on the random sample. To minimise noise from the arbitrary random sampling, and in the same fashion as in the 2pt case, one either takes a large number of random points which increases computational triplet counting, or many random catalogues and average over them. However, these are both inefficient ways of sampling the random space. A first approach to efficiently sample the random counts is to understand how an estimator of a given \( n \)-point correlation function scales with lower point correlation functions and their uncertainties, in particular, with respect to one point statistics which do not need to be zero if the true averaged density is unknown. Hamilton [38] used this approach to construct an estimator and quantify its uncertainty in terms of 1-pt functions. We use this methodology to generalise Hamilton’s estimator to any order, reducing to the JB expression [14] for the 3pt case, and use the formalism to derive an expression for the 3PCF random sampling terms in terms of the 2PCF.

An alternative approach to fast sampling the random points is to understand how the points are geometrically distributed about the data points. For the isotropic 3PCF the picture becomes clearer in the basis where the triangle configurations are parametrised by two sides and one angle [39]. We construct analytic expressions using these ideas to count the random and random-data histograms, reducing enormously the computational cost of the low variance SS estimator for the isotropic 3PCF.

The paper is organised as follows: in the following section, 2, we review Hamilton’s formalism for the 2PCF and extend it to the three point case. Using this formalism we obtain the first random sampling method. In section 3, we describe the geometrical approaches to obtain analytical or semi-analytical expressions for the pure random triple histograms first.

\(^1\)We will not discuss here other approaches to non-Gaussian clustering such as conditional cumulants [29], percolation [30], minimum spanning trees [31] or Minkowski functionals [32, 33].
and then for the mixed data-random histograms. The latter histograms can be calculated by two different approaches, resulting in methods two and three. In the subsequent section, we apply our three methodologies to synthetic data and show which method performs best. In the last section, we include some discussions on the properties of the SS 3PCF estimator, extensions of our methods to include non periodic boundary conditions, and some remarks about our findings. Finally, we include three appendices where we describe Hamilton’s formalism applied to a general 3PCF estimator, visualization schemes for the 3PCF, and the CPU time comparison of our best method with calculations using random catalogues.

2 Approach 1: minimising one point uncertainties

2.1 Hamilton’s formalism

In practice, to calculate correlation functions in galaxy surveys one needs estimators that reduce to the underlying true correlation function in the infinite volume limit and when selection effects are mitigated. Consequently, for finite volume catalogues individual estimators may converge differently due to one point signals or edge effects. Let us explore this in more detail using Hamilton’s formalism [38]. Consider the physical local overdensity

\[ \delta(x) = \frac{n(x) - \bar{n}}{\bar{n}}, \quad (2.1) \]

where \( \bar{n} \) is the true mean galaxy density, together with a function \( W(x) \), which creates a finite subsample from the underlying distribution using the product of a local selection function \( \Phi(x) \) with the galaxy weights \( w \). If the variables are properly normalised, the number of galaxies is

\[ N = \langle n(x) \Phi(x) \rangle, \]

with the angular brackets denoting an average over all points \( x \) in space. Now, let us define the following \((n,k)\)-correlation functions (with 1 ≤ k ≤ n) as

\[ \Psi_k^{(n)}(r_1, r_2, \ldots, r_n) \equiv \frac{\langle W_1 W_2 \cdots W_n \delta_1 \delta_2 \cdots \delta_k \rangle_p}{\langle W_1 W_2 \cdots W_n \rangle_p} \]

\[ \equiv \int_P d^3x_1 d^3x_2 \cdots d^3x_n W(x_1) \cdots W(x_n) \delta(x_1) \cdots \delta(x_k), \]

where \( W_1 W_2 \cdots W_n = w_{12 \ldots n} \Phi_1 \Phi_2 \cdots \Phi_n \) is the weighted product of the selection functions, and the nested integrals are calculated over the region \( P \), which is a polygon shaped figure moved over all space. This polygon is defined by the vertices at \( r_i \) (i = 1 \ldots n) or equivalently by sides given by \( r_{ij} \) (such that \( r_{12} = r_1 - r_2 \), and so on). Notice that the weights do not need to be separable (i.e. \( w_{12} \neq w_1 w_2 \)), however, for later purposes it would be enough to consider only the finite volume effect on the sample and to assume the galaxies are equally weighted, so that \( w_{12 \ldots n} = 1 \).

Furthermore, in the case of statistical homogeneity and isotropy the degrees of freedom reduce\(^2\) to 3n – 6, instead of the 3n components of the vectors \( r_i \). Therefore, we only care about the side magnitudes (\( r_{ij} \equiv |r_{ij}| \)) and enough angles (\( \theta_i \)) between sides to uniquely defined the unorientable polygon. For example, for two-points correlations (pairs) we only have one relevant distance \( r_{12} \), while for three-point functions (triplets) the minimal number to defined the triangle is either three sides (\( r_{12}, r_{23}, r_{13} \)) or two sides and one angle (e.g. \( r_{12}, r_{23}, \) and the angle at the second vertex \( \theta_2 \)). In the case of tetragon, four sides is not

\(^2\)3n – 6 dof hold for n > 2, since for n = 2 one still gets one dof.
enough to fully characterised the shape and six parameters are needed. One option is to use $r_{12}$, $r_{23}$, $r_{34}$, the angles $\theta_2$ and $\theta_3$ (which up to here characterise a planar tetragon), and a further tilting angle between the vectors $r_{12}$ and $r_{34}$. Extrapolating this construction for a general polygon confined to a plane, one needs $2n - 3$ variables; one possibility is to use the $n - 1$ sides ($r_{12}, \ldots, r_{n-1}$) and the $n - 2$ internal angles ($\theta_2, \ldots, \theta_{n-1}$), since the rest of sides and angles will be fixed. The remaining $n - 3$ variables uniquely define how the polygon lives in 3d and not in the plane, and for this one may use the tilting angles $r_{i, i+1} \cdot r_{j, j+1}$ (with $i + 1 < j$).

There are two special limits of the $(n,k)$-correlation functions defined in (2.2). When $(n,k) = (1,1)$, it simply reduces to the measured average density given by

$$\tilde{\delta}(r_1) \equiv \Psi_1^{(1)} = \frac{\langle W_1 \delta_1 \rangle}{\langle W_1 \rangle},$$

and when $k = n$, one obtains the truth n-point correlation function, namely

$$\tilde{\xi}^{(n)}(r_1, \ldots, r_n) \equiv \Psi_n^{(n)} = \frac{\langle W_1 \cdots W_n \delta_1 \cdots \delta_n \rangle}{\langle W_1 \cdots W_n \rangle_P}.$$ (2.4)

Most estimators of n-point correlation functions of discrete samples use n-dimensional histograms which naturally define the oriented distances of the sides that form an “n-plet” (2-plets=pairst, 3-plets=triplets, etc.). In this language, an n-plet histogram $X_1 \cdots X_n(r_1, \ldots, r_n)$ is

$$X_1 \cdots X_n(r_1, \ldots, r_n) = \langle \epsilon_1 \cdots \epsilon_n W_1 \cdots W_n \rangle_P,$$ (2.5) where $X_i$ denotes that the vertex $r_i$ is either on the data ($D$) or the random ($R$) samples, integrals are over the $P$ region defined above, and $\epsilon_i$ is a function for the $i$-vertex which is equal to 1 for the random catalogue ($X_i = R$) or the density field $n_i \equiv n(r_i)$ for the data sample ($X_i = D$). Moreover, we add a subscript “s” when the histogram is being symmetrised over all its arguments. For example, for pairs we get the usual DD, DR and RR histograms. Moreover, we assume equal size bins for the histogram variables. To exemplify the notation, consider the 2-pt histograms

$$DD(r_1, r_2) = \langle n_1 n_2 W_1 W_2 \rangle = \bar{n}^2 \langle W_1 W_2 \rangle \left[ \xi^{(2)}(r_1, r_2) + \sum_i \Psi_1^{(2)}(r_1) + 1 \right]$$ (2.6)

$$DR(r_1, r_2) = \langle n_1 W_1 W_2 \rangle = \bar{n} \langle W_1 W_2 \rangle \left[ \Psi_1^{(2)}(r_1) + 1 \right],$$

$$DR(r_1, r_2) = \frac{1}{2} \langle (n_1 + n_2) W_1 W_2 \rangle = \bar{n} \langle W_1 W_2 \rangle \left[ \frac{1}{2} \sum_i \Psi_1^{(2)}(r_1) + 1 \right],$$

$$RR(r_1, r_2) = \langle W_1 W_2 \rangle,$$ (2.9)

where in the last equality of all expressions we have used the $(n,k)$-correlation function (2.2) and delta field (2.1). If the random and data samples have different number of objects, one could use the normalisation factor $n_{\text{norm}}$, which scales as

$$n_{\text{norm}} \equiv \frac{D}{R} = \bar{n}(1 + \tilde{\delta}),$$ (2.10)

where by $D$ and $R$ we just mean the number of objects in each sample. For the rest of the paper, we will often consider the same number of random points as data points to simplify the
The symmetrised histograms \( DR \) and \( DD \), as it is usually done, we would only consider a linear combination of products expressions. The simplest construction of the 2PCF, \( \xi^{(2)} \), is to use linearly the \( DD \) histogram, and then depending on different contributions of the other histograms with the adequate \( n_{\text{norm}} \) factor, one would obtain different corrections to the 2PCF from the 1pt correlation functions, which in the infinite volume limit will disappear. Table 1 summarises some of the most famous 2PCF estimators and their perturbative 1pt corrections, assuming \( \Psi^{(2)} \) and \( \bar{\delta} \) are smaller than one. Notice that Landy-Szalay (LS) and Hamilton (H) estimators are corrected at second order in \( \Psi^{(2)} \) and \( \bar{\delta} \), hence their low variance, particularly on large scales.

### 2.1.1 Three and higher order estimators

A three point correlation function would need the \( DDD \) histogram together with a combination of the other three possible options \( DDR, DRR \) and \( RRR \), whose structure could be tuned to obtain a low variance estimator such as Landy-Szalay or Hamilton expressions for the two point case. In the Hamilton’s notation introduced earlier, the scaling of these three-point histograms with lower point correlators is

\[
\begin{align*}
  DDD(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) &= \bar{n}^3 \langle W_1 W_2 W_3 \rangle \left[ \xi^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) + \sum_{i<j} \Psi^{(3)}_2(\mathbf{r}_i, \mathbf{r}_j) + \sum_i \Psi^{(3)}_1(\mathbf{r}_i) + 1 \right] \\
  DDR(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) &= \bar{n}^2 \langle W_1 W_2 W_3 \rangle \left[ \Psi^{(3)}_2(\mathbf{r}_1, \mathbf{r}_2) + \Psi^{(3)}_1(\mathbf{r}_1) + 1 \right] \\
  DRR(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) &= \bar{n} \langle W_1 W_2 W_3 \rangle \left[ \Psi^{(3)}_1(\mathbf{r}_1) + 1 \right] \\
  RRR(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) &= \langle W_1 W_2 W_3 \rangle .
\end{align*}
\]

The symmetrised histograms \( DRR_s \) and \( DDR_s \) can be easily obtained from the previous expressions. A general estimator may contain non-trivial functions of the symmetrised histograms but, as it is usually done, we would only consider a linear combination of products
up to a given order. We refer the reader to appendix A for details on the expressions, but bare in mind that there are 18 possible terms assuming up to third order in the numerator. Five of those 18 parameters are easily chosen to avoid corrections on the lowest orders, leaving a family of 2 effective parameters (that are linear functions of the remaining 13). It is important to stress that the simplest $DDD/RRR - 1$ operator would have a non-vanishing two-point correlation, $\Psi_2^{(3)}$, in the infinite sampling limit, or equivalently when $(\delta, \Psi_1^{(3)}) \to (0, 0)$. Of the two parameter family of low variance estimators, the most popular choices are the Szapudi-Szalay (SS) (2.12) and Jing-Börner (JB) [14] constructions, given by

\[
\xi^{(3)}_{SS} = \frac{DDD - 3n_{\text{norm}}DDR + 3n_{\text{norm}}^2DRR - n_{\text{norm}}^3RRR}{n_{\text{norm}}^3RRR},
\]

\[
\xi^{(3)}_{JB} = \frac{DDD[RRR]^2}{(DRR_s)^3} - 3\frac{DDDR_s RRR}{(DRR_s)^2} + 2,
\]

where we have dropped the vertex label to shorten the expressions. These are low biased estimators, whose corrections are given by

\[
\xi^{(3)}_{SS} = (1 + \delta)^{-3} \left[ (\xi^{(3)} - \delta) \sum_{i<j} \Psi^{(3)}_{2,ij} + \delta^2 \sum_i \Psi^{(3)}_{1,i} - \delta^3 \right]
\]

\[
\xi^{(3)}_{JB} = \left( 1 + \frac{1}{3} \sum_i \Psi^{(3)}_{1,i} \right)^{-3} \left[ \xi^{(3)} - \frac{2}{27} \sum_{i<j} \Psi^{(3)}_{1,ij} - \frac{1}{3} \sum_i \Psi^{(3)}_{1,i} \sum_{i<j} \Psi^{(3)}_{2,ij} \right]
\]

where $\Psi^{(3)}_{1,i} \equiv \Psi^{(3)}_{1}(r_i)$ and $\Psi^{(3)}_{1,ij} \equiv \Psi^{(3)}_{2}(r_{ij})$. This methodology may be generalised to higher orders, leading to the following low bias estimator for the n-point correlation function (nPCF)

\[
\xi^{(n)}_{\text{HAM}} = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{n-k} \frac{D^{(k)}R^{(n-k)}(r_1, \ldots, r_n)_s}{[DR^{(n-k)}(r_1, \ldots, r_n)_s]^{k-1}} \left[ R^{(n)}(r_1, \ldots, r_n) \right]^{k-1}
\]

where $D^{(k)}R^{(n-k)}(r_1, \ldots, r_n)_s$, $k = 0, 1, \ldots, n$ are the histograms with $k$ data vertices and $n - k$ random vertices. This expression reduces for $n = 2$ and $n = 3$ to the Hamilton (see 1) and (2.13) expressions respectively. This is to be contrasted with the n-point Szapudi-Szalay (Landy-Szalay for 2pts) estimator [22], given by $\xi^{(n)}_{LS} = (N/R)^n$, where $N \equiv D - n_{\text{norm}}R$.

Using the formalism explained here, one can show that the leading corrections to both 3PCF estimators (SS and JB) are never linear in the one-point functions $\delta, \Psi_1^{(3)}$. The SS estimator tries to push the $\Psi_1^{(3)}$ term to higher orders, whereas JB does it for the $\delta$ correction. Although, there are other 3PCF estimator options which could push these one point functions to order higher than linear, none can disappear the $\delta$ corrections (as JB) and push the $\Psi_1^{(3)}$ to higher order (as SS) at the same time, as shown in appendix A. Moreover, as one considers higher order correlators the choices increase and one could use Hamilton’s formalism to construct other low bias estimators.

2.2 Random sampling using 2pt statistics

After introducing the formalism developed by Hamilton, and a couple of popular three point and higher order point estimators, we proceed to show how one can sample the random-data
histograms of the 3PCF by only using two point estimators. We pay attention to the isotropic 3PCF case, in real space and over simply connected data domains.

Our starting point is the Szapudy-Szalay estimator (2.12), given its low variance and fast convergence. Towards the end of this work (section 5.1), we will justify this choice in more detail, but for the moment let us take this as a working assumption. In order to sample the random catalogues with the data ones to form the DDR or DRR histograms we need to count triplets with at least one leg in the random sample. However, to decrease the random’s noise contribution one usually takes a larger number of randoms (usually around 50 times the number of data points), or many random catalogues of similar size to the data field and average over them. In both approaches the scaling with the number of triplets is larger that those of the DDD histograms. At this point one may wonder if two point expressions can more efficiently sample the randoms.

A first approach based in the formalism we have introduced previously holds for uncorrelated weights only (and our choice of local selection functions). In this case, $w_1...n = \Pi_i w_i$, the correlation $\Psi_k^{(n)}$ reduces to $\Psi_k^{(k)}$, because the integrals $\int W_i (i = k + 1, \ldots, n)$ in the numerator and denominator of (2.2) cancel. Under this assumption the correction to the Szapudy-Szalay estimator (2.14) simplifies to

$$\hat{\xi}^{(3)}_{SS}(r_{12}, r_{13}, r_{23}) = \frac{1}{(1 + \bar{\delta})^3} \left[ \xi^{(3)} - \bar{\delta} \left( \sum_{i>j} \xi^{(2)}(r_{ij}) - \bar{\delta} \sum_i \Psi_1^{(1)}(r_i) + \bar{\delta}^2 \right) \right]. \quad (2.17)$$

To efficiently sample the random catalogues, we would like an estimator which only calculates the DDD/RRR piece (of the SS estimator (2.12)) and uses a two point estimator to approximate all further terms. In other words, we are looking for a 2PCF estimator $X$ to build up the 3PCF Szapudy-Szalay in the following way

$$\hat{\xi}_X(r_1, r_2, r_3) = \frac{\text{DDD}}{\text{RRR}} (r_{12}, r_{13}, r_{23}) - \sum_{i<j} \hat{\xi}_X^{(2)}(r_{ij}) - 1, \quad (2.18)$$

such that it gives the same corrections as in (2.17). By inspecting 2PCF estimators (see table 1), one could try to find the estimator that leads to the same one point expansion, in Hamilton’s language, as the term above in the round brackets. One should consider there are three 2PCF contributions, one for each triangle side, and the $(1 + \bar{\delta})^{-3}$ factor when looking for the correct 2PCF expression. It turns out to be the Hewett estimator [42]. In summary, our first method to simplify the random counting of the Szapudy-Szalay 3PCF estimator$^3$ is

$$\hat{\xi}^{(3)}_{SS}(r_{12}, r_{13}, r_{23}) = \frac{\text{DDD}}{n_\text{norm}^3 \text{RRR}} (r_{12}, r_{13}, r_{23}) - \sum_{i<j} \hat{\xi}_H^{(2)}(r_{ij}) - 1. \quad (2.19)$$

3 Approaches 2 and 3: geometrical schemes

In this section, we try to focus on periodic boundary conditions and isotropic correlation functions over smooth and continuous geometries, hence we only care about the triangle

$^3$If, instead, one chooses the JB estimator (2.13), Hamilton’s 2PCF estimator has the same one point corrections, resulting in the equivalent expression to (2.19) given by

$$\hat{\xi}^{(3)}_{JB}(r_{12}, r_{13}, r_{23}) = \frac{\text{DDD} (\text{RRR})^2}{(\text{DRR})^3} (r_{12}, r_{13}, r_{23}) - \sum_{i<j} \hat{\xi}_H^{(2)}(r_{ij}) - 1.$$

− 7 −
sides. The basic idea is to reduce the random counting with analytic expressions, which together with the DDD histogram, lead to the an accurate approximation of the Szapudi-Szalay 3PCF estimator. First, we explore the isotropic RRR histogram, then generalise it to the DRR case, and conclude with the non-trivial case of the DDR counting. By the end, we will come back to the possibility of non-periodic boxes and statistical anisotropies.

3.1 Analytical expression for the RRR estimator

To calculate the 2PCF in a box of volume $V$ with periodic boundary conditions, it is well known that the RR piece, over a smooth and simply connected geometry, can be obtained from an analytical expression (see, for example, the documentation of the popular code CUTE [43]). Let us review the construction argument of this analytic expression because it will be useful for the isotropic three point correlation case. Imagine we take a fixed pivot point in a random catalogue with $N$ points, and draw all possible pairs from it. If we assume homogeneous and isotropically distributed random points over a continuous region, the pairs are equally distributed along spherical shells around the pivot point, and the pair number at each shell is proportional to the density ($N/V$, with $N$ the number of points in the box) times the spherical shell volume. Moving from one pivot to all of them will imply an extra factor of $N$. Finally, for unweighted random catalogs this construction gets us to the histogram on an infinitesimal thick shell given by

$$dRR(r) = N \left( \frac{N}{V} \right) (4\pi r^2) \, dr,$$

which one needs to integrate over $r$ bins to get the final result, namely

$$RR(r) = \int_{r-\Delta r/2}^{r+\Delta r/2} dRR = \frac{N^2}{V} v(r, \Delta r),$$

where $v(r, \Delta r)$ is the spherical shell volume given by

$$v(r) = \frac{4\pi}{3} \left[ (r + \Delta r/2)^3 - (r - \Delta r/2)^3 \right].$$

The random counting converges to the previous formulae in the infinite data limit ($N_R \to \infty$). Inspired by this construction, it is straightforward to derive an analytical expression for the RRR histogram over the same random catalogue, which can be easily understood in the $\{r_1, r_2, \theta\}$ basis for the triplets ([39]), instead of using the three triangle’s sides.

As before, let us consider a fix pivot point in the R catalogue, that we call $r_1$, and from there draw all possible triangles with vertices $r_2$ and $r_3$ in the same sample. With our choice of basis, the resulting isotropic histograms of triplets will be functions of the following binned variables: two distances $r_{12} = |r_{12}|$ and $r_{13} = |r_{13}|$ and the opening angle around the pivot point $\mu_1 = \cos(\theta_1) = r_{12} \cdot r_{13}$. Because a random catalogue does not have a preferred orientation, $\mu_1$ is a random variable, hence the histogram only depends on its binning size, $\Delta \mu_1$. In other words, the monopole of a multipole decomposition in $\theta$ contains all the RRR information. Notice that using the three side triangle as our variables does not lead to the same argument because the closing triangle side, $r_{23}$, is a function of the other sides, as we will discuss in what follows. As a result of the angle independence, the RRR histogram only depends on the radial distributions, $N_{\text{piv}}(r)$, of the $r_2$ and $r_3$ points times the number of
pivot points \( (N) \) and the width of the angle bin \( (\Delta \mu_1) \). Under this construction, the \( RRR \) histogram gives

\[
RRR(r_{12}, r_{13}, \mu_1) = -\frac{N}{2} N_{pivot}(r_{12}) N_{pivot}(r_{13}) \Delta \mu_1
\]

\[
= -\frac{N^3}{2V^2} v(r_{12}) v(r_{13}) \Delta \mu_1. \tag{3.4}
\]

The factor of 2 and the minus sign are due to the fact that the Cosine function is decreasing as its arguments grows and the size of the interval is \( 2 = \cos(0) - \cos(\pi) \). One may use the number of bins, \( n_{\mu_1} \), instead of \( \Delta \mu_1 \) by substituting \( \Delta \mu_1 \) with \( 2/n_{\mu_1} \).

To make further progress we need to transform this analytic \( RRR \) expression to the \( r_{12}, r_{13} \) and \( r_{23} \) basis. A simple approach is to take the infinitesimal bin limit (take a linear order limit on the “Deltas” in (3.4) and assume they are of infinitesimal size) and write the \( \mu_1 \) in terms of \( r_{ij} \) using the law of Cosines, to substitute it back into (3.4). By following these steps, we obtain the infinitesimal histogram

\[
dRRR(r_{12}, r_{13}, r_{23}) = \frac{8\pi^2 N^3}{V^2} r_{12} r_{13} r_{23} \, dr_{1} dr_{2} dr_{3} \tag{3.5}
\]

Then, we just integrate this result over the variable’s bins, assuming the same bin size \( (\Delta r) \) for each side, leading to

\[
RRR(r_{12}, r_{13}, r_{23}) = \int_{r_{12} - \Delta r/2}^{r_{12} + \Delta r/2} \int_{r_{13} - \Delta r/2}^{r_{13} + \Delta r/2} \int_{r_{23} - \Delta r/2}^{r_{23} + \Delta r/2} dRRR(r_{12}, r_{13}, r_{23})
\]

\[
= \frac{8\pi^2 N^3}{V^2} r_{12} r_{13} r_{23} [\Delta r]^3. \tag{3.6}
\]

These integrals should be calculated over \( r \)-regions where the triangle inequality,

\[
r_{12} + r_{13} \geq r_{23}, \tag{3.7}
\]

is satisfied. A note of caution is that a coarse binning would introduce an important error in the result. Actually, this is only relevant where the equality sign holds in the previous inequality (3.7); see for example figure 1. Geometrically, this corresponds to collinear points where the area of the triangle vanishes. To avoid this error, we can include the triangular equality in the integration limits of (3.6), or consider a refined binning in those regions where the triangular equality is satisfied. We will exemplify this error and how to mitigate it in the following section, where our methods will be compared using numerical simulations.

### 3.2 Analytical expressions for the \( DRR \) and \( DDR \) crossed terms

In this subsection, we present two different methodologies to estimate the factors \( DDR_s \) and \( DRR_s \), which together with the previous analytic formula for \( RRR \) and the \( DDD \) histogram, will be needed to complete the isotropic 3PCF Szapudi-Szalay estimator (2.12). As a reminder, these factors must be included in any 3PCF estimator to remove the bare 2PCF contribution included in the \( DDD/RRR \) term.

Using periodic boxes, as well as local weights and the selection function, enormously constraints the statistical properties of some histograms in isotropic n-point correlation functions. As it is well known for the 2PCF, the \( DR \) tends to \( n_{\text{norm}}RR \) once the number of
random points is sufficiently large (this can be thought as a consequence of achieved translation invariance in this limit). For the same reason, all n-point histograms with only one leg of the polygon in the data catalogue and the rest on the random samples \((DR\ldots R)\) reduce to the purely random n-point histogram \((n_{\text{norm}}R\ldots R)\). Therefore, for our 3PCF estimator all \(DRR_s\) terms can be approximated by the \(RRR\) analytic expression (3.6). As previously explained, this result can be understood in Hamilton’s language as the fact that all one point contributions vanish. For the 2PCF and 3PCF cases, the details on how the one-leg data histograms converge to the pure random ones is better appreciate in figure 2. For the 2PCF this fact implies that all histograms of table 1 reduce to the Peeble-Hauser expression. In contrast, the vanishing one point correlations do not imply all 3PCF estimators are equal, instead, there are non-trivial two point pieces remaining from the \(DDD\). In the SS estimator, the removal of such tow point contribution is mostly encoded in the \(DDR_s\) term. Therefore, we will focus on this term for the rest of our analysis, since it is not trivial to obtain analytically.

Following the same construction as in section 3.1 for the \(RRR\) histogram, we can develop an equivalent expression for the \(DDR_s\) histogram. We start, as before, in \([r_{12}, r_{13}, \mu_1]\) basis where the intuition helps for the initial derivation. Afterwards we need to map our result to \([r_{12}, r_{13}, r_{23}]\) basis in order to obtain the final result. Consider a pivot point, \(r_1\), in the data catalogue with \(N\) points and draw all possible triangles with one further vertex in the random samples, \(r_2\), and another in the random sample, \(r_3\), which we also assume to contain \(N\) points.

As before, the angle \(\mu_1\) is randomly distributed so the \(DDR(r_{12}, r_{13}, \mu_1)\) only depends on the product of two independence pair distributions. The \(r_{12}\) pairs have both legs on the data points, hence the distribution cannot be approximated analytically and is proportional to \(DD(r_{12})\). In contrast, the \(r_{13}\) pairs are randomly distributed about the pivot point and given as before by \(N_{\text{piv}}(r)\). The resulting expression is

\[
DDR(r_{12}, r_{13}, \mu_1) = -\frac{N}{2V} DD(r_{12}) v(r_{13}) \Delta \mu_1,
\]

where, as before, the \(-2\) is due to the range and decreasing property of the Cosine function. Notice that this histogram is not necessarily symmetric with respect to interchanging the

**Figure 1.** Left: 2d slice of a spherically thin shell around a pivot point of radius \(r\) and thickness \(\Delta r\). The expected density of random points is \(Nv(r)/V\), with \(v(r)\) defined in (3.3). Right: grid slice for \(r_3\) fixed with forbidden region by the triangle inequality (3.7) in white. Most errors in our fast random sampling methods arise from points in the forbidden region that fit into bins where the triangle equality (red line).
Figure 2. As the number of random point increase for periodic boxes, the one-leg data n-point histograms ($DR_s$ in the left plot and $DRR_s$ in the right plot) converge to the purely random n-point histograms ($n_{norm}RR$ and $n_{norm}RRR$ respectively). We use the data points from the section 4 simulation, three sets of 10, 25 and 50 random catalogues and estimate the dispersion from repeating the 50 random catalogue calculation twenty five times. The random samples have the same points as the data ($n_{norm} = 1$).

first two variables, but that can be easily obtain by symmetrising it. However, this result can not be the full $DDR_s$ histogram, which should contain triangles where the pivot points, $r_1$, are placed in the random catalogue and the other vertices in the data points. In the case of a “R” pivot point, the opening angle is not a random variable anymore, and our line of thought cannot be followed. Nevertheless, our $DDR$ result (3.8) contains enough information to build up the reaming pieces, since it can be mapped to the $(r_{12}, r_{13}, r_{23})$ basis and symmetrise it in those variables. In order to do so we follow two possible methodologies, which lead to different precision results.

3.2.1 Approach 2: numerical completion of $DDR$

The first of these methodologies uses the law of cosines numerically to compute the closing triangle side $r_{23}$ and map the $DDR$ of equation (3.8) into the $(r_{12}, r_{13}, r_{23})$ basis. Then we just symmetrise over all $r_{ij}$ variables and obtain the final $DDR_s$ histogram. The main drawback of this approach is that uncertainty on each distance bin before the mapping may enlarge the error on the resulting $r_{23}$. To minimize this error, we can take smaller bins ($\Delta r^0 < \Delta r$ and $\Delta \mu_1^0 < \Delta \mu_1$) in $r_{12}$, $r_{13}$ and $\mu_1$ initially and the re-bin our final result to the desired size. This algorithm has a complexity of order $(2\pi/\Delta \mu_1^0)(r_{max}/\Delta r^0)^2$.

3.2.2 Approach 3: analytic completion of $DDR$

In this method, we follow the same procedure in section 3.1 to map the histogram to the $(r_{12}, r_{13}, r_{23})$ basis and symmetrise it in that space. For that, as we did for the $RRR$, it is easier to construct a differential form of equation (3.8), obtaining

$$
dDDR (r_{12}, r_{13}, r_{23}) = \frac{8\pi^2 N^3}{V^2} \left(1 + \xi^{(2)}(r_{12})\right) r_{13}^2 r_{13}^2 dr_{12} dr_{13} d\mu_1 \\
= \frac{8\pi^2 N^3}{V^2} \left[1 + \xi^{(2)}(r_{12})\right] r_{12} r_{13} r_{23} dr_{12} dr_{13} dr_{23},
$$

(3.9)

where we have used that $DD(r) \sim RR(r) \left(1 + \xi^{(2)}(r)\right)$ (remember all 2PCF estimators converge to the same form for periodic boxes, since $DR \rightarrow RR$), the analytic expression (3.2) for $RR$, taken only linear terms on each $dr_{ij}$, and the law of Cosines to obtain the last line.
in the three triangle sides’ basis. We integrate this differential expression over a bin of size \( \Delta r \) on each triangle side, resulting in

\[
\text{DDR}(r_{12}, r_{13}, r_{23}) = RRR(r_{12}, r_{13}, r_{23}) + \frac{8\pi^2 N^3}{V^2} r_{13} r_{23} [\Delta r]^2 \int_{r_{12} - \Delta r}^{r_{12} + \Delta r} r_{12} \xi^{(2)}(r_{12}) \, dr_{12}
\]  

(3.10)

To evaluate the integral, we can either interpolate the 2PCF and do the integral numerically, or approximate it with a sum over very small bins to avoid errors. We will use the latter, and even for a very fine-grained binning the calculation would not add many further computational resources compared to the DDD calculation. Notice that this histogram is not symmetric yet, thus after symmetrising the expression and adding the DDD piece, we obtain our final 3PCF estimator formula

\[
\xi^{(3)}_{SS}(r_{12}, r_{13}, r_{23}) = \frac{\text{DDD}}{n_{\text{norm}} RRR}(r_{12}, r_{13}, r_{23}) - \sum_{i<j} \frac{1}{r_{ij}} r_{ij} \xi^{(2)}(r_{ij}) - 1,
\]  

(3.11)

where the expression with the overbar is the average of \( r_{ij} \xi^{(2)} \) over the coarse-grained bin \([r_{ij} - \frac{\Delta r}{2}, r_{ij} + \frac{\Delta r}{2}]\) using the fine-grained bins. Moreover, where the triangle equality holds there is a further error, hence the need of further grid refinements.

4 Applying our fast sampling methodologies to mock catalogues

In this section we aim at applying the previous three methodologies to synthetic data and assess the degree of accuracy. For this purpose we use the L-PICOLA code [44] to generate a small simulation with \( N = 32^3 \) particles and a fixed volume of \( V = (250 \text{MPc}/h)^3 \). Since our aim is not to get a highly accurate signal but to find out which of the different methods performs best, it is enough to consider this box size and density.

In order to appreciate the difference between models we use different plotting techniques. On a first approach, we map all bins (a 3d array) into a one dimensional object by moving along one variable completely holding the other two variables fixed. We then jump to the next bin in the second variable and repeat the first variable bin-sweeping. We generalise this idea into the third variable, and call the 1d axis the triangle index. This technique is particularly useful to evaluate how different 3PCF contributions differ from each other while taking into account errors, as shown in figure 2. However, it does not perform well to understand particular features of the signal, since adjacent bins in the 3d bin array may not be adjacent in the triangle index, resulting in periodic jumps of the signal. A second, and popular, visualization scheme is obtained by fixing the third triangle side \( r_{23} \) to a set fixed values and do 2d density plots on the remaining triangle sides \( r_{12} \) and \( r_{13} \). It is hard to assess differences and errors compared to the signal with this technique. An alternative approach to this last one is to use the opening angle \( \mu_1 \) instead of \( r_{23} \) and do a multipole decomposition (based on Legendre polynomials) on that angle [45], which captures more information but is less intuitive. We refer the reader to appendix B for further details about 3PCF visualizations.

How do the methodologies presented before match the isotropic 3PCF Szapudy-Szalay estimator? Our results can be graphically summarised in figure 3. All methods do give approximate the same signal, however, there are small errors, mostly in the short scale regime, associated to the details of each method that we describe in what follows. In order to test our fast correlation function expressions in this small periodic box, we first calculate the 3PCF
Figure 3. 3PCF in the configurations space using the SS estimator. Each column shows 3PCF slices with $r_3$ fixed ($r_3 = 9, 24, 39, 54$). The allowed $r_1 - r_2$ region is determined by the triangle inequality (3.7). The first column corresponds to the SS estimator using 100 random catalogues while the remaining columns are the three fast random sampling methods developed here. Method 1 is based on Hamilton’s formalism (eq. (4.1)), while Methods 2 (section 3.2.1) and 3 (section 3.2.2) are the geometrical approaches. By eye, the overall agreement is good, particularly for the two geometrical methods. The largest failures are for the first method at the boundaries of the allowed region, where the triangle equality holds. Figure 4 shows in more detail how each method performs, and for larger simulation boxes see the similar plot of 7.

numerically using the SS estimator (2.12). For this, we use 1250 random catalogues with same mean density of particles as in the data ($n_{\text{norm}} = 1$), and average over 25 groups of 50 random samples each to obtain the mean of different histograms of the 3PCF and their dispersion. In figure 2 we show the convergence for the DRR histograms depending on a partial number of used random catalogues, together with the magnitude of the final dispersion using the largest number of randoms catalogues (50). Moreover, we consider a maximum distance of $r_{\text{max}} = 60\, \text{MPC}/\h$ (24% of the box side length) and bin sizes of $\Delta r = 3\, \text{MPC}/\h$. In order to achieve accurate results and avoid using interpolation schemes, we use smaller bins in the
geometrical approaches (methods 2 and 3) for different triangle configurations. We discuss the details of each method in what follows, but before doing so, keep in mind that we assume an integer number of smaller bins in each $\Delta r$, to avoid interpolating between the coarse and fine grained grids.

For the first method, one can notice that the two point estimator is the same as the basic one of PH ([40], table 1) because, as it has been discussed before, using periodic boxes implies that the $DR$ histograms reduce to the $RR$ one. Moreover, to avoid any random sampling in the two point contribution we can calculate the $DD$ piece and use the analytic formula for $RR$, given by (3.2). Therefore, the resulting formula based on our previous construction (2.19) for the first method is

$$
\hat{\xi}^{(3)}_{SS}(r_{12}, r_{13}, r_{23}) = \frac{DD}{RR} (r_{12}, r_{13}, r_{23}) - \frac{V}{N^2} \left[ \frac{DD(r_{12})}{v(r_{12}, \Delta r)} + \frac{DD(r_{13})}{v(r_{13}, \Delta r)} + \frac{DD(r_{23})}{v(r_{23}, \Delta r)} \right] - 1,
$$

(4.1)

where $v(r)$ is given by (3.3). In our study we calculate the $DD$ using the same binning $\Delta r$ as for the numerical 3PCF. This method approximates well at large scales but has larger departures at short and intermediate scales (see figures 4 and 3). Actually, it strongly fails where the triangle equality is satisfied, as it is appreciated by eye in figure 3, where one sees the deviation (shown as the opposite color to the numerical 3PCF) on the boundary allowed region of each plot (for the second column corresponding to method one). Those boundaries are precisely where the triangle equality holds. If we were to work with larger and more physical mock catalogues, the deviations would be very important, as we discuss in appendix B, particularly in figure 7. It is important to stress that in this method there is not any refinement of the grid used in the 2 point statistic part to improve the approximation, as opposed to the other two methodologies that we describe next. This is because we are not dealing with histograms but with a correction that is the 2PCF.

For the second method, we map the $DDR$ histogram (3.8) from $(r_{12}, r_{13}, \mu_1)$ to $(r_{12}, r_{13}, r_{23})$ using the law of Cosines, which by construction satisfies the triangle inequality (3.7). However, because of this nonlinear mapping we need a refined binning in the two point pieces to diminish the associated error in the isotropic 3PCF final expression. By using three choices for the fine-grained grid given by $\Delta r^0 = (1/100, 1/200, 1/400)\Delta r$, in figure 4, we show how the methodology converges towards the 3PCF signal within its numerical error. Moreover and for simplicity, we assume the same number of bins in the angular and radial directions, hence $\Delta \mu = \frac{2 \Delta r^0}{d_{\text{max}}}$. As a result of our trials, we recommend having at least two order of magnitude difference between coarse and fine grained bin sizes, $\Delta r$ and $\Delta r^0$. To discuss about the computational cost of this method, notice that the final calculation — without counting the $DDD$ piece — scales as the cube of the number of fine-grained bins from the mapping (remember that the $DD$ histograms could be included as a secondary product of the $DDD$ calculation), which is smaller than any algorithm that counts triplets between random and data catalogues.

In the third and final method, we take the formula (3.11), and consider two additional partitions or re-binnings in order to avoid precision errors of the final result. In the first refinement we take a finer grained grid to calculate the $r \xi^{(2)}$ average in expression (3.11). For this re-binning we use $\Delta r^0 = \Delta r/400$ and notice that these additional bins do not increase the total calculation time by much, because they only apply to the two point statistics. The result of this approach can be seen in the top plot of figure (3.7), where the only points that do not fit into the expected 3PCF numerical calculations are those where the triangle equality holds (the equal sign of (3.7)). Pictorially, we can see this error arising as those
Figure 4. The top plot shows the normalised differences between the numerical piece of the 3PCF with random-data mixing terms ($\Sigma = (3\text{DDR} - 3\text{DRR} + \text{RRR})/\text{RRR}$) and the 3 methods to fast sampling the random catalogues. The bottom left graph describes how the second method 3.2.1 converges as one increases the fine-grained binning in the two point statistics. The bottom right plot displays the convergence of third methodology 3.2.2 as one decreases the bin size on those $r$’s that satisfy the triangle equality (equal sign in (3.7)). Errors are estimated using 25 independent groups of 50 random catalogues.

triangular sections on the boundary region in figure 1. In order to improve the signal on those points we perform a second refinement, but on the corresponding bins only. In figure 4 we choose this additional rebinning to be $\Delta r_{\text{eq}} = (1/10, 1/100, 1/200)\Delta r$ and see how the result converges to the full numerical 3PCF if this second re-binning is sufficiently small.

Finally, an independent work [35] constructed a similar fast random sampling method to our geometrical approaches, but based on a different construction directly in the $\{r_{12}, r_{13}, r_{23}\}$ basis. Their formulae arise from the intersection of two spherical thin shells, about the centers $\mathbf{r}_1$ and $\mathbf{r}_2$, with radii $r_{13}$ and $r_{23}$ respectively. Details can be found in [35], but to summarise they find the formula

$$XXR_s(r_{12}, r_{13}, r_{23}) = 4\pi N^3 V \left\{ \int_{r_{12} - \Delta r/2}^{r_{12} + \Delta r/2} n_{XX}(r_{12}) n_{X}(r_{12}) V_{\text{inter}}(r_{12}) dr_{12} + \text{permutations} \right\}, \quad (4.2)$$

where $n_{\text{norm}} = 1$ and each $X$ could be either $D$ or $R$. The function $V_{\text{inter}}(r_{12}, r_{13}, r_{23})$ is the intersection volume of the two spherically thin shells, which also depends on $\Delta r$, whose analytic expression can be found in [35]. The density number is $n_{RR} = n_{RD} = 1$, but for $n_{DD}$, one uses an interpolation of the $DD$ histogram, to get $n_{DD}(r) = \text{Interpolate}(DD, r)/(3N v(r))$, where $v(r)$ is defined by (3.3). For the $RRR$ and $DRR_s$ histograms, the authors arrived to our same expressions however, we differ in the conceptually most complicated histogram, the $DDR_s$. Moreover, because of these constructions the authors conclude that their formalism does not perform well in bins there the triangle equality holds.
Figure 5. Comparison between our best performing method (number 3) and the one proposed in [35]. Differences between the numerical histograms and the fast-sampling random methodology are shown, weighted by the dispersion of the numerical 3PCF estimation using 100 random catalogues. If one chooses an appropriate normalization, the RRR and DRR histograms (top left and right respectively) are the same, except for bins where the method of [35] fails (i.e. triangle equality region). In contrast, for the same normalization, the DDR histogram by [35] departs on all scales (bottom left). A different normalization that adjusts the large scales, still shows deviations on shorter scales (bottom right).

We use their open-source code to assess the accuracy compared to our best method: number 3. Moreover, we use the same variable as in [35] to show the difference in performance, which is given by the simple difference of the analytic method with the numerical operator, divided by the dispersion of the latter. Results are shown in figure 5, where we can see that our method outperforms theirs. There is a normalization factor that makes the $\text{RRR}$ and $\text{DRR}_a$ results to agree. However, for this same normalization factor their $\text{DDR}$ signal strongly disagrees with the numerical expression, hence from our result as well. If one uses a different normalization factor only for their $\text{DDR}_a$ signal, then their result agrees on large scales, but again deviates considerably on the short scale regime, apart from the collinear points where their code does not give an answer. We believe our precision lies on using a basis with one angle instead of the third triangle side, which allow us to push the analytic expression further (equation (3.11)) than with the other methods.

5 Discussion

Before stating some conclusions on our findings, we would like to discuss two important extensions to our methodologies: how the low variance isotropic 3PCF estimators converge with respect to a numerical random sampling, and how some of our results can be extended to more general finite volume surveys.
Figure 6. Relative differences between the $3PCF_n$, estimated from the SS estimator (equation (2.12)) or the JB (equation (2.13)) estimator, and the analytic expression $3PCF_e$ of our third fast random sampling method (equation (3.11)). With 10 random catalogues, the SS estimator converges faster than the JB one.

5.1 Assessing estimators convergence with the number of random points

We have chosen to work with the Szapudi-Szalay (SS) estimator instead of other choices such as the Jing-Börner (JB). However, after the tools and results presented here, we are in a place to justify our choice. Among all low variance estimators (see appendix A for details), the structure of SS estimator has leading order corrections which depend on the biased overdensities $\bar{\delta}$ (see equation (2.14)), and not on the other one point function, $Ψ_1^{(3)}$, as it happens with the JB expression (equation (2.15)). The expectation that $\bar{\delta}$ has a lower shot noise contribution from the number of random points, suggests that the SS estimator converges faster than JB, or any other estimators whose leading correction has $Ψ_1^{(3)}$ terms. Notice, however, that the SS leading correction in Hamilton’s language may be obtained by other estimators (see appendix A for a discussion on the degeneracy of these corrections). Furthermore, we could use the third analytic method for random sampling to assess the fast convergence of the SS estimator over JB counterpart. For periodic boundary conditions, since the histogram $DRR$ converges to $RRR$, the SS (2.12) and JB (2.13) reduce to the same expression, namely

$$ξ^{(3)}(r_{12}, r_{13}, r_{23}) = \frac{DDD}{RRR} - 3\frac{DDR_s}{RRR} + 2. \quad (5.1)$$

However, provided a not so large number of random points, the convergence to the final result is different for the two estimators. If we assume the third methodology gives an accurate answer of the final asymptotic expression (as we have discussed in the previous section), then we can compare how the SS and JB estimators converge given the number of random points. The results are summarised in figure 6, where it is clear how the SS estimator converges faster than the JB one. Now that we have justified the SS estimator, we could think of using our results in arbitrary volumes.
5.2 Fast random sampling with non periodical boundary conditions

Can we apply our calculations for the random sampling pieces of isotropic 3PCF to a non-periodic finite volume with arbitrary continuous shape? In principle, there are a few extensions to our hypotheses that one can follow (see examples of methodologies to address the edge effects [22, 34, 46–48]). One idea corresponds to divide the arbitrary volume, \( V \), into a central region, \( W \), where all spheres drawn from each point in \( W \) of the maximal desired length, \( r_{\text{max}} \), fit entirely into the volume \( V \). In this region \( W \), our fast random sampling formulae would be valid as long as the weight on each galaxy remains local and there are no further weights on the random catalogues.\(^4\) In consequence, all boundary effects would come from the thick boundary region \((Q \equiv V - W)\) where the signal would be obtained in the usual way using random catalogues. Under this construction, the histogram \( DDD \) is calculated over the whole volume, but the \( XXR \) pieces (with \( X \) either \( D \) or \( R \)) are the addition of two parts: \( XXR_Q + XXR_W \). The pieces with sub-index \( Q \) are obtained numerically by counting enclosed triangles in the volume \( W \), whereas those in \( W \) come from our analytical expressions, with the following details. For the \( RRR_W \) histogram, one uses the analytical expression (3.6), but with \( r_{12} \) being within \( W \) and \( r_{13}, r_{23} \) running over the whole volume \( V \). In the \( DRR \) case, one gets \( DRR_W = RRR_W \), as we have discussed previously. Finally, if we consider the third methodology (section 3.2.2), the \( DDR_W \) histogram is obtained using the expression (3.11), with the restriction that \( r_{12} \) is confined to \( W \) but not the other distances. The final result of these \( W \)-histograms is obtained after symmetrising over the three distances. This 3PCF algorithm looses speed over a periodic box, but still has the benefit of the analytic expressions in the \( W \) region, which would introduce an important time reduction for large and simply connected survey volumes.

A further speed push may be achieved by using a adaptive boundary \( W^* \), whose size depends on the maximal scale involve in a given triangle configuration [46]. For a particular bin with \( r_{12}, r_{13} \) and \( r_{23} \) in the 3PCF, the size of \( W^* \) is such that all spheres of radius \( \text{Max}(r_{12}, r_{13}, r_{23}) \) about any point in \( W^* \) are fully contained within \( V \). As in the fixed volume case, our analytic expression can be used in this regime with the restriction of \( r_{12} \in W^* \), and symmetrising the final results over the three distances. The triangles in the complement space \((Q^* = V - W^*)\) are obtained using random catalogues. This algorithm becomes particularly effective when all scales of the triangle are small, leaving the full triplet random-data counting for triangles where at least one of the scales is large.

Although we have outlined a border correction strategy to shown how these fast sampling formulae can be applied to more general volumes, one still needs to address the inclusion of optimal and systematic weights, among other issues (such as redshift space), to really apply these techniques to more realistic galaxy surveys. However, the results presented here are useful at many stages of real data analysis, such as to validate the perturbation theory modeling of the signal and the assessment of some systematic errors.

6 Conclusions

The use of higher-order statistics in new large volume galaxy surveys, which provide complementary information to the two point clustering measurements, will be greatly pursued in the forthcoming years. However, efficient algorithms to overcome the calculation scaling

\(^4\)This would not be the case in surveys such as DESI where one could use pair weighting to mitigate the assignment of optical fibers to the targets [49].
with the number of objects of these correlation functions are needed to fully exploit their use. In this work, we focus on the isotropic Three Point Correlation Function (3PCF) over smooth closed volumes and propose three methodologies to count triplets of mixed random and data points without the need of random catalogues. These methodologies directly extend to higher order correlation functions by including not only the two point pieces but also the whole hierarchy of lower correlation functions up to one order less than the desired result.

In the isotropic 3PCF case, the pure data histogram $\text{DDD}$ contains the connected three point clustering (times the $\text{RRR}$ histogram), together with a non-trivial piece of two point correlations if the object weights are local. In order to single out the connected three point piece one needs to include data-random mixing histograms. Szapudi-Szalay (SS) and Jing-Börner choose particular combinations that ensure subleading corrections in one point operators at quadratic order. These corrections become important for finite volume surveys. We consider the SS estimator as the target expression for our random sampling techniques to converge to, because we believe it enjoys a similar mitigation of systematics as its two point cousin, the Landy-Szalay estimator (see for example [23, 38, 50–54]), though this has not been fully studied in detail. Therefore, we would expect our fast random approximations to inherit the same statistical properties as the SS estimator.

A first idea to simplify the random sampling is to substitute the random-data mixed terms in the SS estimator with a two point estimator that has the same one point corrections, which we show corresponds to the so-called Hewett estimator. We call this our method 1. At the practical level, we find that this prescription has a poor convergence to the full SS estimator result, especially in the spread limit (when the triangle vertices become collinear). However, this line of thought is useful to construct other low variance three point estimators (see appendix A), or to extend approximation beyond our working assumptions (e.g. in the anisotropic 3PCF case).

A different strategy to achieve fast random-samplings for periodic boxes is to understand how random points distribute about data points. When the random distribution is isotropic one could easily get convinced that the distribution does not depend on the opening angle between the triangle sides, which allow us to derive a pure random histogram ($\text{RRR}$) over homogeneous, smooth and continuous geometries. Moreover, in this case we are considering of periodic boxes, the same result we obtain for the $\text{RRR}$ holds true for the $\text{DRR}$ histogram, which corresponds to one triangle vertex laying in the data catalogue. This is true because of the vanishing one point corrections in this limit.

The difficult histogram to model is when the two triangle vertices are in the data catalogue ($\text{DDR}$), since it is this piece that contains the two point correlations, which are usually non-trivial for physical datasets. Using geometrical arguments, we arrive at finite analytic expressions in the basis of the triangles defined by two sides and their opening angle. Mapping these expressions to the three sides basis can be done numerically or analytically, leading to the methods that we call 2 and 3, respectively.

To assess the accuracy and speed of the three methods we use simulated data over periodic boxes. We find that method 3 outperforms the other two and reduces by slightly more than two orders of magnitude the CPU time on the full 3PCF calculation, as can be seen in figure 9 (see appendix C for details). This results are estimated by a code that computes the histograms by a nearest neighbour algorithm. A point to notice with respect to the full 3PCF calculation time is that, even though our fast random sampling prescriptions scale as $N^3$, with $N$ the number of objects. Nevertheless, clever
ideas such as the recent one by Sabiu et al. [55] using kd-trees and graph databases or the multipole decomposition by Slepian and Eisenstein [34] allow for a faster triplet counting, because of their, respectively, $N^{2.3}$ and $N^2$ scaling. Therefore, one may think of incorporating the ideas shown here to those algorithms for a faster estimation of the isotropic 3PCF.

In a complementary study we assume our method 3 is accurate enough to use it as the asymptotic signal that a three point estimator over periodic boxes converges to in the infinite random sampling limit. Under this consideration, we can explore the convergence properties of some of the most popular estimators. In particular, we show that the SS estimator converges faster than the Jing-Börner counterpart with the number of random points, as one may have expected from the two point analogy [51]. Finally, we want to point out that even though the results presented here focus on the SS estimator, our constructions can be easily generalised to approximate other 3PCF estimators, which may be of interest to study systematics in galaxy surveys or to other applications beyond cosmology.

Acknowledgments

We acknowledge the support of DAIP-UG, Instituto Avanzado de Cosmología A.C., and CONACyT (in particular, for FS graduate’s scholarship and research grant No. 286897). This research was partially supported through computational and human resources provided by the DCI-UG DataLab, and the LAMOD UNAM project (a collaborative effort between the IA, ICN and IQ institutes at UNAM), through the clusters Atocatl and Tochtli. We thank J. Clemente, C. Guzman, L. Díaz, E. Murrieta, M. Aguilar and O. Burgueño for computer support. Finally, we thank Baojiu Li for let us use his simulations, and Daniel Eisenstein and Zachary Slepian for sharing their 3PCF code.

A Construction of low variance 3PCF estimators

In this section we show in detail how to obtain the SS (2.12) and JB (2.13) estimators from a general 3PCF expression using Hamilton’s formalism (section 2.1). The most general 3PCF estimator combines terms with powers of the histograms $DDD$, $DDR$, $DRR$, and $RRR$. If we assume at most cubic order terms in the denominator and $n_{\text{norm}} = 1$, the most general estimator would be given by

$$
\xi^{(3)}_{\text{EST}} = a_0 \frac{DDD}{RRR^3} \frac{DRR^2}{RRR} + a_1 \frac{DDD}{DRR^3} \frac{RRR^2}{RRR} + a_2 \frac{DDD}{DDR^3} \frac{DRR^2}{RRR} + a_3 \frac{DDD}{DDR^3} \frac{RRR^2}{DRR} + a_4 \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_5 \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_6 \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_7 \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_8 \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_9 \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{10} \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{11} \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{12} \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{13} \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{14} \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{15} \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{16} \frac{DDD}{DDR^2} \frac{RRR}{RRR} + a_{17} \frac{DDD}{DDR^2} \frac{RRR}{RRR} - a_{18} .
$$

(A.1)

Different one or two point corrections to the 3PCF estimators depend on the tuning of the coefficients $a_i$. There are three natural conditions we can initially impose: that the $\xi^{(3)}_{\text{EST}}$ coefficient is one, a vanishing constant term and no isolated two point contributions. These
conditions resume, respectively, in
\[
\begin{align*}
  a_0 + a_1 + a_6 + a_7 + a_{12} + a_{13} &= 1, \\
  \sum_{1 \leq i \leq 17} a_i - a_{18} &= 0, \\
  3a_0 + 3a_1 + a_2 + a_3 + 3a_6 + 3a_7 + a_8 + a_9 + 3a_{12} + 3a_{13} + a_{14} + a_{15} &= 0.
\end{align*}
\]

By analyzing the remaining terms, we get a general three-point estimator:
\[
\xi_{EST}^{(3)} = \xi^{(3)} \left[ 1 + b_1 \sum_{1 \leq i \leq 3} \Psi_1^{(3)}(r_i) - 3(1 + b_1)\bar{\delta} \right] + b_3 \sum_{1 \leq i \leq 3} \Psi_2^{(3)}(r_i) - 3b_3\bar{\delta} + b_4 \bar{\delta} \sum_{1 \leq i \leq 3} \Psi_1^{(3)}(r_i)
\]
\[
+ \sum_{1 \leq i < j \leq 3} \Psi_2^{(3)} \left[ b_2 \sum_{1 \leq i \leq 3} \Psi_1^{(3)}(r_i) - (1 - 3b_2)\bar{\delta} \right]
\]
\[
+ \mathcal{O} \left[ \bar{\delta}^p \left( \sum_{1 \leq i \leq 3} \Psi_1^{(3)}(r_i) \right)^q \right]_{p,q \geq 2},
\]
where
\[
\begin{align*}
  b_1 &= \frac{1}{3} (2a_0 - 3a_1 - 2a_6 + a_7 - a_{13}) \\
  b_2 &= \frac{1}{9} (6a_0 - 9a_1 + 2a_2 - 3a_3 - 6a_6 + 3a_7 - 2a_8 + a_9 - 3a_{13} - a_{15}) \\
  b_3 &= \frac{1}{3} (5a_0 + 4a_2 - a_3 + 3a_4 - 3a_5 + a_6 + 4a_7 + 3a_9 + 2a_{10} - 2a_{11} + 3a_{12} + 2a_{13} + 2a_{14} + a_{15} + a_{16} - a_{17}) \\
  b_4 &= \frac{1}{3} (-25a_0 - 16a_2 - a_3 - 9a_4 - 9a_5 - a_6 - 16a_7 - 9a_9 - 4a_{10} - 4a_{11} - 9a_{12} - 4a_{13} - 4a_{14} - a_{15} - a_{16} - a_{17}).
\end{align*}
\]

The last term of the r.h.s. converges faster to zero than the other terms in the large volume limit, hence it can be neglected, resulting in a reliable 3PCF estimator with 15 \((18 - 3)\) free parameters. A natural further choice that lowers the variance is \(b_3 = b_4 = 0\), leaving a family of two effective parameters \(b_1\) and \(b_2\), whose values are degenerated among the remaining 13 free combinations of the \(a_i\)’s. At this level, all estimator are corrected at second order and the corrections cannot be made all zero. Therefore, further restrictions would allow us to choose among these low variance estimators for the 3PCF. There are two choices of parameters worth mentioning: 1) if only the linear term contributions in the numerator are considered \((a_k = 0\) for \(k \leq 11\)), and by imposing a leading second order correction in \(\sum \Psi_1^{(3)}\) and \(\bar{\delta}\) \((b_1 = b_2 = 0)\), we obtain the Szapudy and Szalay (2.12) estimator, meanwhile 2) using all terms up to cubic order in the numerator, and demanding a leading sixth order correction in \(\bar{\delta}\) \((b_1 = -1, b_2 = 1/3)\) results in the Jing and Börner (2.13) estimator. However, notice that there are other options that lead to the same quadratic corrections of JB or SS, which we do not intend to explore further in here. Furthermore, choosing to have \(\Psi_1^{(3)}\) corrections instead of \(\bar{\delta}\) would naturally lead to estimators that converge slower with the number of random points. We elaborate more on this in section 5.1.
B 3PCF visualization schemes

In the literature, we can find several visualization schemes for the 3PCF. Here, we focus on the isotropic correlation function case, which depends on three variables; thus the visualization is not a trivial task. Below we discuss a few schemes to visualize the 3PCF, its advantages and disadvantages.

In configuration space (with 3 sides of the triangle), the 3PCF defines a tetrahedral allowed region. A popular choice in Fourier Space is to sketch the Bispectrum as a transparent 3d density plot (see for example [56]). Although this scheme would allow a full 3PCF visualization, the superposition of information on the image can complicate the study of the information, specially far away from border of the allowed region where densities from different bins superpose. To avoid this visual saturation, a popular choice is to take one relation between variables (e.g. $r_{13}$ or $r_{12}/r_{13}$ fixed to a constant) to obtain a 2d density plot, or even better, a 2d scatter plot when a further relationship among the variables is considered. Actually, these particular triangle configurations, if chosen before constructing the histograms, may accelerate the calculation, since not all possible triangles would be scanned by the algorithm. Geometrically, the equilateral (all sides equal), squeezed (isosceles with one side tending to zero) and spread (towards the collinear points) configurations stand out, and due to their symmetry these shapes could also pick up strong signals from perturbation theory. However, to consider special triangles decreases the available information on the 3PCF, which may be relevant for some studies, such as testing gravity [57, 58], where the larger difference between models may not be where the largest 3PCF signal is. figure 7 shows an example of such visualization, where we have study the performance of our three random sampling methods in a halo mock constructed from a cosmological simulations with a larger volume and a maximum scale of 140 $Mpc/h$ using the ECOSMOG code [59] with Rockstar [60] to find halos (particular details on the mocks can be found in [57]). The result is similar to that of the smaller box in figure 3. One could appreciate for this larger box how to our three analytical methods perform when at least one triangle side is the BAO scale. We have a poor random sampling due to large computational cost for our computing resources, given the large box and maximal scale. However, under these conditions one can easily see how the method based on Hamilton’s formalism fails almost everywhere by eye, but specially for the boundary regions where the triangle equality holds, consistently with the discussion in section 4.

Extending on the idea of information compression from the previous paragraph, one may think of a cosmological model, which has a certain number of parameters, and search for triangular configurations that lead to the largest variance on those parameters. To our knowledge this approach has not been fully explored in the 3PCF, but there is some studies on this direction for the bispectrum (see for example [61, 62]).

Another visualization scheme that we can find in literature is the triangle index (see for example [35]). This scheme allows us to visualize the 3PCF as a 2d scatter plot. This is particularly relevant to compare 3PCF signals (e.g. between theories, to check convergence, etc.), and specially useful when there are errors for each bin involved. The construction is simple: one just needs to glide over the cubic matrix of the radial variables in any possible way to construct a simple 1d (longer) vector. Because of this mapping, many continuous bins may be separated at the end vector, hence the signal would show periodical jumps. For this reason, it is not a useful visualization to appreciate shapes in the signal, such as the BAO structure. Examples of such visualization are in figures 2, 4, 5 and 6. The third scheme
Figure 7. Same figure as 3 but with a larger mock catalogue of halos from a larger simulation with 342963 particles in a box of $(1\text{Gpc})^3$. We use 50 random catalogues for the numerical 3PCF and a binning of $\Delta r = 10$. For the method one we do not take smaller bins, for method two we use a $\Delta r^0 = \Delta r/400$ in all bins and for method 3 a $\Delta r_b = \Delta r/200$ where the triangle equality holds.

we want to exemplify was first discussed in [45]. As noted by author, the isotropic 3PCF can be characterized by two sides lengths and the angle between them, $\xi^{(3)}(r_1, r_2, \cos \theta)$, which allows for multipole decomposition in a similar way as for the anisotropic 2PCF. In detail, the 3PCF decomposes into

$$\xi^{(3)}(r_1, r_2, \cos \theta) = \sum_{\ell} \xi_{\ell}(r_1, r_2) P_\ell(\cos \theta), \quad (B.1)$$

where the coefficients $\xi_{\ell}(r_1, r_2) (\ell = 0, 1, 2, \ldots)$ can be easily seen as 2d density plots. These scheme can also be used to speed up the 3PCF calculation [34], since one may use the relation between spherical harmonics and Legendre polynomials to an $O(N^2)$ scaling of the code, as opposed to the $O(N^3)$ performance of naïve algorithms. In figure 8 we show how the result
Figure 8. Multipole coefficients $\xi_\ell(r_{12}, r_{23})$ for several values of $\ell$, using the fast 3PCF multipole calculation code of Slepian and Eisenstein [34] for the large mock catalogue of halos of figure 7. We use 50 random catalogs, and multiply the coefficients by $r_{12} r_{23}$ to increase the signal on large scales.

of calculating the multipole decomposition 3PCF signal directly using the code by Daniel J. Eisenstein and Zachary Slepian.  

C CPU time comparison

In this section we show the comparison of our fast method 3 (which has been verified to be the most accurate) against different random sampling methods using catalogues. We have considered a box of $(250 \text{ MPc} h^{-1})^3$, with three different: $\rho_0$, $\rho_0/2$ and $\rho_0/4$, with $\rho_0 = (32/250 \text{ MPc}^{-1} h)^3$. For the random catalogues, we have taken into account four different sets of files with different density, such that each set contains 100 times the number of data points, $N_D$. The random suits include: a set of 100 random catalogs with $N_R = N_D$ points; a set of 10 random catalogs with $N_R = 10N_D$ points, a set of 5 random catalogs with $N_R = 20N_D$ points and a set of 2 random catalogs with $N_R = 50N_D$ points. In figure 9, we show the cpu time (recorded using the clock function of the C library time.h) required for one processor to estimate the isotropic 3PCF with our fast method 3 and by using the four random catalogues describe above.

\footnote{A public version of the code [34] can be found in [63], https://nbodykit.readthedocs.io/.
Figure 9. CPU time as a function of data points, $N_D$, for different random samplings methods. Our third and most accurate method (with a $\Delta r = \Delta r/20$) reduces by more than two orders of magnitude the fastest method using random catalogues, which consists on averaging over 100 random catalogues with the same number of points as in $N_D$. For completeness, we include methods with increasing density in the random catalogues, whose running time we infer from calculating the DDR histogram and the scaling of the other histograms. We use 48 threads in a cluster with Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz processors.

References

[1] P.J.E. Peebles and E.J. Groth, Statistical analysis of catalogs of extragalactic objects. V. Three-point correlation function for the galaxy distribution in the Zwicky catalog, *Astrophys. J.* **196** (1975) 1.

[2] E.J. Groth and P.J.E. Peebles, Statistical analysis of catalogs of extragalactic objects. VII. Two- and three-point correlation functions for the high-resolution Shane-Wirtanen catalog of galaxies, *Astrophys. J.* **217** (1977) 385 [SPIRE].

[3] M. Takada and B. Jain, The three-point correlation function in cosmology, *Mon. Not. Roy. Astron. Soc.* **340** (2003) 580 [astro-ph/0209167] [SPIRE].

[4] F. Bernardeau, S. Colombi, E. Gaztanaga and R. Scoccimarro, Large scale structure of the universe and cosmological perturbation theory, *Phys. Rept.* **367** (2002) 1 [astro-ph/0112551] [SPIRE].

[5] SDSS collaboration, Three-point correlation functions of SDSS galaxies in redshift space: Morphology, color and luminosity dependence, *Publ. Astron. Soc. Jap.* **56** (2004) 415 [astro-ph/0403638] [SPIRE].
6. WiggieZ collaboration, The WiggleZ Dark Energy Survey: constraining galaxy bias and cosmic growth with 3-point correlation functions, *Mon. Not. Roy. Astron. Soc.* 432 (2013) 2654 [arXiv:1303.6644] [SPIRE].

7. C.K. McBride et al., Three-Point Correlation Functions of SDSS Galaxies: Luminosity and Color Dependence in Redshift and Projected Space, *Astrophys. J.* 726 (2011) 13 [arXiv:1007.2414] [SPIRE].

8. J. Pan and I. Szapudi, The Monopole moment of the three-point correlation function of the 2-degree Field Galaxy Redshift Survey, *Mon. Not. Roy. Astron. Soc.* 362 (2005) 1363 [astro-ph/0505422] [SPIRE].

9. R.C. Nichol et al., The effect of large-scale structure on the sdss galaxy three-point correlation function, *Mon. Not. Roy. Astron. Soc.* 368 (2006) 1507 [astro-ph/0602548] [SPIRE].

10. C.K. McBride et al., Three-Point Correlation Functions of SDSS Galaxies: Constraining Galaxy-Mass Bias, *Astrophys. J.* 739 (2011) 85 [arXiv:1012.3462] [SPIRE].

11. F. Marin, The Large-Scale 3-point correlation function of SDSS Luminous Red Galaxies, *Astrophys. J.* 737 (2011) 97 [arXiv:1011.4530] [SPIRE].

12. Y. Wang, X.-H. Yang, H.J. Mo, F.C. van den Bosch and Y.-Q. Chu, The Three-point correlation function of galaxies: Comparing halo occupation models with observations, *Mon. Not. Roy. Astron. Soc.* 353 (2004) 287 [astro-ph/0404143] [SPIRE].

13. G.V. Kulkarni, R.C. Nichol, R.K. Sheth, H.-J. Seo, D.J. Eisenstein and A. Gray, The Three-Point Correlation Function of Luminous Red Galaxies in the Sloan Digital Sky Survey, *Mon. Not. Roy. Astron. Soc.* 378 (2007) 1196 [astro-ph/0703340] [SPIRE].

14. Y.P. Jing and G. Borner, The three point correlation function of galaxies determined from the Las Campanas redshift survey, *Astrophys. J.* 503 (1998) 37 [astro-ph/9802011] [SPIRE].

15. E. Gaztañaga, P. Norberg, C.M. Baugh and D.J. Croton, Statistical analysis of galaxy surveys. II. The 3-point galaxy correlation function measured from the 2dFGRS, *Mon. Not. Roy. Astron. Soc.* 364 (2005) 620 [astro-ph/0506249] [SPIRE].

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

17. Z. Slepian et al., Detection of baryon acoustic oscillation features in the large-scale three-point correlation function of SDSS BOSS DR12 CMASS galaxies, *Mon. Not. Roy. Astron. Soc.* 469 (2017) 1738 [arXiv:1607.06097] [SPIRE].

18. Z. Slepian et al., Constraining the baryon-dark matter relative velocity with the large-scale three-point correlation function of the SDSS BOSS DR12 CMASS galaxies, *Mon. Not. Roy. Astron. Soc.* 474 (2018) 2109 [arXiv:1607.06098] [SPIRE].

19. M. Moresco et al., The VIMOS Public Extragalactic Redshift Survey (VIPERS). Exploring the dependence of the three-point correlation function on stellar mass and luminosity at $0.5 < z < 1.1$, *Astron. Astrophys.* 604 (2017) A133 [arXiv:1603.08924] [SPIRE].

20. E. de Carvalho, A. Bernui, H.S. Xavier and C.P. Novaes, Baryon acoustic oscillations signature in the three-point angular correlation function from the SDSS-DR12 quasar survey, *Mon. Not. Roy. Astron. Soc.* 492 (2020) 4469 [arXiv:2002.01109] [SPIRE].

21. S.S. Tie et al., UV background fluctuations and three-point correlations in the large-scale clustering of the Lyman α forest, *Mon. Not. Roy. Astron. Soc.* 487 (2019) 5346 [arXiv:1905.02208] [SPIRE].

22. I. Szapudi and A.S. Szalay, A New class of estimators for the N point correlations, *astro-ph/9704241* [SPIRE].
[23] S.D. Landy and A.S. Szalay, Bias and variance of angular correlation functions, *Astrophys. J.* **412** (1993) 64 [SPIRE].

[24] DESI collaboration, The DESI Experiment Part I: Science, Targeting, and Survey Design, *arXiv:1611.00036* [SPIRE].

[25] LSST collaboration, LSST: from Science Drivers to Reference Design and Anticipated Data Products, *Astrophys. J.* **873** (2019) 111 [arXiv:0805.2366] [SPIRE].

[26] EUCLID collaboration, Euclid Definition Study Report, *arXiv:1110.3193* [SPIRE].

[27] J. Green et al., Wide-Field InfraRed Survey Telescope (WFIRST) Final Report, *arXiv:1208.4012* [SPIRE].

[28] A.W. Moore et al., Fast algorithms and efficient statistics: N-point correlation functions, in MPA/ESO / MPE Joint Astronomy Conference: Mining the Sky, ESO Astrophysics Symposia European Southern Observatory, pp. 71–82 (2001) [DOI] [astro-ph/0012333] [SPIRE].

[29] J. Pan and I. Szapudi, Conditional cumulants in weakly non-linear regime, *Mon. Not. Roy. Astron. Soc.* **361** (2005) 357 [astro-ph/0405590] [SPIRE].

[30] S.F. Shandarin, Percolation Theory and the Cell/Lattice Structure of the Universe, *Sov. Astron. Lett.* **9** (1983) 104.

[31] J.D. Barrow, S.P. Bhavsar and D.H. Sonoda, Minimal spanning trees, filaments and galaxy clustering, *Mon. Not. Roy. Astron. Soc.* **216** (1985) 17.

[32] K.R. Mecke, T. Buchert and H. Wagner, Robust morphological measures for large scale structure in the universe, *Astron. Astrophys.* **288** (1994) 697 [astro-ph/9312028] [SPIRE].

[33] M. Kerscher, Statistical analysis of large scale structure in the universe, *Lect. Notes Phys.* **554** (2000) 37 [astro-ph/9912329] [SPIRE].

[34] Z. Slepian and D.J. Eisenstein, Computing the three-point correlation function of galaxies in $O(N^2)$ time, *Mon. Not. Roy. Astron. Soc.* **454** (2015) 4142 [arXiv:1506.02040] [SPIRE].

[35] D.W. Pearson and L. Samushia, Computing three-point correlation function randoms counts without the randoms catalogue, *Mon. Not. Roy. Astron. Soc.* **486** (2019) L105 [arXiv:1903.09715] [SPIRE].

[36] O.H.E. Philcox and D.J. Eisenstein, Computing the Small-Scale Galaxy Power Spectrum and Bispectrum in Configuration-Space, *Mon. Not. Roy. Astron. Soc.* **492** (2020) 1214 [arXiv:1912.01010] [SPIRE].

[37] O.H.E. Philcox, A Faster Fourier Transform? Computing Small-Scale Power Spectra and Bispectra for Cosmological Simulations in $O(N^2)$ Time, *arXiv:2005.01739* [SPIRE].

[38] A.J.S. Hamilton, Toward Better Ways to Measure the Galaxy Correlation Function, *Astrophys. J.* **417** (1993) 19 [SPIRE].

[39] J.N. Fry, Gravity, bias, and the galaxy three-point correlation function, *Phys. Rev. Lett.* **73** (1994) 215.

[40] P. Peebles and M. Hauser, Statistical analysis of catalogs of extragalactic objects. III. the shane-wirtanen and zwicky catalogs, *Astrophys. J. Suppl.* **28** (1974) 19.

[41] M. Davis and P.J.E. Peebles, A Survey of galaxy redshifts. V. The Two point position and velocity correlations, *Astrophys. J.* **267** (1982) 465 [SPIRE].

[42] P.C. Hewett, The estimation of galaxy angular correlation functions, *Mon. Not. Roy. Astron. Soc.* **201** (1982) 867.

[43] D. Alonso, CUTE solutions for two-point correlation functions from large cosmological datasets, *arXiv:1210.1833* [SPIRE].
C. Howlett, M. Manera and W.J. Percival, *L-PICOLA: A parallel code for fast dark matter simulation*, Astron. Comput. 12 (2015) 109 [arXiv:1506.03737] [inSPIRE].

I. Szapudi, *Three-point statistics from a new perspective*, Astrophys. J. Lett. 605 (2004) L89 [astro-ph/0404476] [inSPIRE].

R.T. Adrian Baddeley, Ege Rubak, *Spatial Point Patterns: Methodology and Applications with R*, Chapman & Hall CRC (2016).

B.D. Ripley, *Frontmatter*, pp. i–iv, Cambridge University Press (1988).

J. Pan and P. Coles, *Boundary corrections in fractal analysis of galaxy surveys*, Mon. Not. Roy. Astron. Soc. 330 (2002) 719.

M.-J. Pons-Borderia, V.J. Martinez, D. Stoyan, H. Stoyan and E. Saar, *Comparing estimators of the galaxy correlation function*, Astrophys. J. 523 (1999) 480 [astro-ph/9906344] [inSPIRE].

M. Kerscher, I. Szapudi and A.S. Szalay, *A Comparison of estimators for the two-point correlation function*, Astrophys. J. Lett. 535 (2000) L13 [astro-ph/9912088] [inSPIRE].

M. Vargas-Magana et al., *An optimized correlation function estimator for galaxy surveys*, Astron. Astrophys. 554 (2013) A131 [arXiv:1211.6211] [inSPIRE].