Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **504,** 2911–2923 (2021) doi:10.1093/mnras/stab961
Advance Access publication 2021 April 22

Cosmological cross-correlations and nearest neighbour distributions

Arka Banerjee [1,2,3,4]★ and Tom Abel[2,3,4]

[1]*Fermi National Accelerator Laboratory, Cosmic Physics Center, Batavia, IL 60510, USA*
[2]*Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA*
[3]*Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[4]*SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*

Accepted 2021 March 29. Received 2021 March 23; in original form 2021 February 15

ABSTRACT

Cross-correlations between data sets are used in many different contexts in cosmological analyses. Recently, *k*-nearest neighbour cumulative distribution functions (*k*NN-CDF) were shown to be sensitive probes of cosmological (auto) clustering. In this paper, we extend the framework of NN measurements to describe joint distributions of, and correlations between, two data sets. We describe the measurement of *joint k*NN-CDFs, and show that these measurements are sensitive to all possible connected *N*-point functions that can be defined in terms of the two data sets. We describe how the cross-correlations can be isolated by combining measurements of the joint *k*NN-CDFs and those measured from individual data sets. We demonstrate the application of these measurements in the context of Gaussian density fields, as well as for fully non-linear cosmological data sets. Using a Fisher analysis, we show that measurements of the halo-matter cross-correlations, as measured through NN measurements are more sensitive to the underlying cosmological parameters, compared to traditional two-point cross-correlation measurements over the same range of scales. Finally, we demonstrate how the NN cross-correlations can robustly detect cross-correlations between sparse samples – the same regime where the two-point cross-correlation measurements are dominated by noise.

Key words: cosmology – cosmology: large-scale structure of Universe.

1 INTRODUCTION

Measurements of statistical cross-correlations between data sets are widely used in cosmology and astrophysics, often in the context of spatial clustering of the data. These measurements, which characterize the spatial correlations in the fluctuations in the density field or number counts of the two data sets, have a number of uses, depending on the context (see e.g. Rhodes et al. 2013). For example, cross-correlations can help break various degeneracies and allow for stronger constraints on parameters of interest. The measurement of galaxy–galaxy lensing – the cross-correlation of galaxy counts and weak lensing maps is widely used, in conjunction with measurements of galaxy clustering and cosmic shear, to break degeneracy between halo bias and cosmological parameters (e.g. Seljak et al. 2005; Mandelbaum et al. 2013; Miyatake et al. 2015; More et al. 2015; Abbott et al. 2018; Joudaki et al. 2018; Heymans et al. 2020; Wibking et al. 2020). Similarly, cross-correlations between galaxy populations and galaxy clusters can also break individual bias degeneracies, and better constrain cosmological parameters (e.g. Croft, Dalton & Efstathiou 1999; Zu & Weinberg 2013; Paech et al. 2017; Salcedo et al. 2020; To et al. 2020). In other contexts, cross-correlations can be used to mitigate the effects of survey systematics on the cosmology analysis. This is especially relevant for data sets that are measured in different surveys – while the cosmological signal is correlated, the systematics between the two surveys are usually uncorrelated. Examples of this aspect is the use of cross-correlations between cosmic microwave background lensing maps, and galaxy–galaxy lensing and galaxy clustering (e.g. Baxter et al. 2016; Kirk et al. 2016; Schaan et al. 2017; Singh, Mandelbaum & Brownstein 2017; Abbott et al. 2019; Singh et al. 2020).

Cross-correlations and their applications in cosmology can be roughly divided into two categories, based on the statistical significance of the relevant signal. In the high signal-to-noise regime, the variations in the cross-correlation signal as a response to a change in the underlying cosmological parameters are large compared to the errors in the measurement process. Therefore, the cross-correlation measurements can be used to infer the values of the cosmological parameters. This is precisely how the cross-correlation measurements are used in most of the references cited above. On the other hand, when the signal-to-noise ratio is low, the focus is primarily on the detection of a cross-correlation signal, rather than its use directly in parameter inference (e.g. Blake et al. 2006; Granett, Neyrinck & Szapudi 2008a,b; Bianchini et al. 2015; Li, Yalinewich & Breysse 2019; Namikawa et al. 2019; Ammazzalorso et al. 2020; Fang et al. 2020). A common theme in this regime is to look for cross-correlations between rare but interesting astrophysical signals and a set of relatively dense, well-calibrated tracers of large-scale structure. The rare events usually have such low number densities that their autoclustering is completely dominated by noise, but the clustering signal can be recovered through the cross-correlations, which is not affected by shot noise. This technique has been explored in the context of cross-correlating ultrahigh-energy neutrino sources detected by IceCube with galaxies (Fang et al. 2020), gamma-ray sources with weak lensing measurements (Ammazzalorso et al. 2020), and fast radio bursts with galaxies (Li et al. 2019).

★ E-mail: arka@fnal.gov

It is worth noting that the term 'cross-correlations' in the context of cosmology generally refers to the two-point cross-correlations of the data sets. The two-point cross-correlations capture the full information between two Gaussian fields. However, at late times, and on small scales, cosmological density fields can be highly non-linear, and depart strongly from a Gaussian distribution. Therefore, correlations between two of these fields can exist beyond the two-point cross-correlation. These higher order cross-correlations can, in principle, be used to better characterize clustering of the two fields (e.g. Schneider & Watts 2005; Munshi et al. 2014; Rizzo, Mota & Valageas 2017).

Recently, Banerjee & Abel (2021) introduced a new approach to studying clustering in the cosmological data – through the use of k-nearest neighbour cumulative distribution functions (kNN-CDF). This is the empirical cumulative distribution function of distances from a set of volume-filling, Poisson distributed random points to the k-nearest data points, and has various attractive properties. It is computationally inexpensive to measure, and is formally sensitive to all connected N-point functions of the continuous field from which the data points are drawn. Banerjee & Abel (2021) demonstrated that these summary statistics are more sensitive to the underlying cosmological parameters than the two-point autocorrelation function (over the same range of scales), and therefore promises to be a useful tool to optimally extract information from small scales in cosmological surveys.

Although Banerjee & Abel (2021) focused on the clustering of only one set of tracers, in this paper, we extend the theoretical and measurement framework of kNN distributions to describe the joint clustering of two sets of tracers, enabling the same formalism to describe both autocorrelations and cross-correlations. We demonstrate how the *joint* NN distributions are formally sensitive to all N-point functions that can be defined by the two fields, and identify the parts that are sensitive to the cross-correlations. In this context, the term 'cross-correlations' refer to any statistical dependence of the two fields with each other, not just the traditional two-point correlations. We outline how to measure these joint distributions efficiently using distances to NN data points of each set from a set of dense volume-filling randoms, as well as the method to measure only the cross-correlation piece from the same set of measurements. We apply these measurements in the context of two Gaussian tracers, where the distributions can be predicted analytically, and then to tracers of fully non-linear fields. For the latter, we demonstrate the statistical power of the kNN cross-correlations compared to the two-point cross-correlations through two examples – one in the high signal-to-noise regime and one in the low signal-to-noise regime. Given the fact that cross-correlations are used so widely in cosmological analyses, this extension of the kNN formalism and measurements vastly increases the range of possible applications.

The paper is arranged as follows: in Section 2, we introduce the framework of joint NN distributions for two correlated fields and outline how these are measured on actual data. Then, in Section 3, we consider the case of joint kNN distributions and cross-correlations for tracers of correlated Gaussian fields. In Section 4, we apply the same measurements to sets of tracers of fully non-linear fields. We compare the sensitivity of kNN cross-correlation measurements of simulation haloes and matter field to underlying cosmological parameters, and compare it to that of two-point cross-correlations. We also demonstrate the detection of a cross-correlation signal for sparse samples of haloes using kNN measurements when two-point measurements fail to detect a signal. Finally, we conclude, and discuss some aspects of the presentation in Section 5

2 FORMALISM AND MEASUREMENT OF CROSS-CORRELATION IN THE NEAREST NEIGHBOUR FRAMEWORK

2.1 Formalism

In this section, we lay out the formalism for describing spatial cross-correlations between two data sets in terms of joint data counts in a given volume. We consider two sets of tracers – each set tracing an underlying continuous density field. Throughout this paper, we will assume that the tracers are distributed according to a local Poisson point process on the underlying fields, and that the Poisson parameter is proportional to the enclosed 'mass' of the density field over the volume of interest. Therefore, fluctuations in the underlying fields are imprinted on to the fluctuations in the number counts of the tracers. The two sets of tracers will have non-zero cross-correlations when the overdensities and underdensities of the underlying fields coincide with each other on average – the cross-correlation is the highest when the fluctuations in the two fields coincide exactly.

These cross-correlations will be imprinted on the *joint* count distribution of the tracers of the two fields, i.e. the probability of finding k_1 tracers from set 1 *and* k_2 tracers from set 2, in a volume V centred around random points in the volume of interest. The relationship between the joint count distribution and the underlying correlation functions are best summarized through the generating function of the counts. In this particular example, the generating function is a power series in the dummy variables z_1 and z_2. The coefficients of various powers of z_1 and z_2 correspond to the probabilities above. The generating function is therefore a useful way to summarize the infinite set of joint probabilities that can be computed from the data. As derived in Appendix A, the generating function, $P(z_1, z_2|V)$, for the joint counts in volume V is given by

$$P\left(z_1, z_2|V\right) = \exp\Bigg[\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\frac{\bar{n}_1^{k_1}(z_1-1)^{k_1}}{k_1!}\frac{\bar{n}_2^{k_2}(z_2-1)^{k_2}}{k_2!} \times \int_V d^3\boldsymbol{r}_1 ... d^3\boldsymbol{r}_{k_1} d^2\boldsymbol{r}'_1 ... d^3\boldsymbol{r}'_{k_2}\xi^{(k_1,k_2)}\Bigg], \quad (1)$$

where $\bar{n}_i$ represent the mean number density of each set of tracers, and $\xi^{(k_1,k_2)}$ represents the connected correlation function (in terms of the continuous fields) defined with k_1 factors of field 1 and k_2 factors of field 2. For example $\xi^{(2,0)}$ represents the two-point autocorrelation function of field 1, $\xi^{(0,2)}$ represents the two-point autocorrelation function of field 2, and $\xi^{(1,1)}$ represents the two-point cross-correlation between the two fields. Note that if the two fields are uncorrelated, or statistically independent, i.e. $\xi^{(k_1,k_2)}$ is non-zero only when either $k_1 = 0$ or $k_2 = 0$, the generating function factorizes into two independent generating functions – one for each set of tracers: $P(z_1, z_2|V) = P_1(z_1|V)P_2(z_2|V)$. Also note that while the volume V in Eq, 1 can have any arbitrary shape in general, we will focus on the scenario where it corresponds to a sphere of radius r.

The probability $\mathcal{P}(k_1, k_2|V)$ of finding exactly k_1 counts of data from set 1, and k_2 counts of data from set 2 in volume V are given by various derivatives of the generating function with respect to the

dummy variables z_1 and z_2:

$$\mathcal{P}(k_1, k_2|V) = \frac{1}{k_1!}\frac{1}{k_2!}\left[\left(\frac{\mathrm{d}}{\mathrm{d}z_1}\right)^{k_1}\left(\frac{\mathrm{d}}{\mathrm{d}z_2}\right)^{k_2} P(z_1, z_2|V)\right]_{z_1,z_2=0}. \quad (2)$$

From the forms of equation (1) and 2, it is evident that, for all values of k_1, k_2, the probabilities $\mathcal{P}(k_1, k_2|V)$ are related to all possible connected N-point functions that can be defined from fields 1 and 2. This includes all possible cross-terms – $\xi^{(k_1,k_2)}$ for non-zero values of both k_1 and k_2.

Using the same formalism, it is also possible to write down the generating function $C(z_1, z_2|V)$ for the joint *cumulative* counts, i.e. the probability of finding more than k_1 tracers from set 1 *and* more than k_2 tracers in volume V:

$$C(z_1, z_2|V) = \frac{1 - P_1(z_1|V) - P_2(z_2|V) + P(z_1, z_2|V)}{(1 - z_1)(1 - z_2)}, \quad (3)$$

where $P_i(z_i|V)$ represent the generating function for the counts of each individual set of tracers. Note that in the absence of cross-correlations, i.e. when $P(z_1, z_2|V) = P_1(z_1|V)P_2(z_2|V)$, this generating function also factorizes into a product of the generating functions for the cumulative counts for each distribution individually:

$$\begin{aligned} C(z_1, z_2|V) &= \frac{1 - P_1(z_1|V) - P_2(z_2|V) + P_1(z_1|V)P_2(z_2|V)}{(1 - z_1)(1 - z_2)} \\ &= \left(\frac{1 - P_1(z_1|V)}{1 - z_1}\right)\left(\frac{1 - P_2(z_2|V)}{1 - z_2}\right) \\ &= C_1(z_1|V)C_2(z_2|V). \end{aligned} \quad (4)$$

Just as in equation (2), one can compute the individual terms $\mathcal{P}(> k_1, > k_2|V)$ for any value of k_1, k_2 from the derivatives of $C(z_1, z_2|V)$:

$$\mathcal{P}(> k_1, > k_2|V) = \frac{1}{k_1!}\frac{1}{k_2!}\left[\left(\frac{\mathrm{d}}{\mathrm{d}z_1}\right)^{k_1}\left(\frac{\mathrm{d}}{\mathrm{d}z_2}\right)^{k_2} C(z_1, z_2|V)\right]_{z_1,z_2=0}. \quad (5)$$

There are two issues to note here that will be relevant throughout the paper: first, the joint cumulative probabilities $\mathcal{P}(> k_1, > k_2|V)$ are sensitive to terms that capture the cross-correlation between fields 1 and 2, *i.e.* $\xi^{(k_1,k_2)}$ for $k_1 \neq 0$, $k_2 \neq 0$, as well as terms which capture the clustering of the two fields individually, i.e. terms such as $\xi^{(k_1,0)}$ and $\xi^{(0,k_2)}$. Secondly, in the case where the underlying fields are uncorrelated,

$$\mathcal{P}(> k_1, > k_2|V) = \mathcal{P}_1(> k_1|V)\mathcal{P}_2(> k_2|V), \quad (6)$$

and therefore any deviation from this condition can be treated as a measure of the degree of correlation between the two fields under consideration. We will return to the implications of this factorization in Section 2.2, but it is worth noting here that the term 'uncorrelated' as used here is a more stringent criterion than just the absence of two-point cross-correlation, as often used in the literature. Since the *k*NN measurements are sensitive to not only the two-point cross-correlation, but all higher order terms, 'uncorrelated' in this context implies complete statistical independence of the two distributions.

Having set-up the formalism to show how these joint cumulative probabilities of the tracers capture the correlations between two different underlying fields, in the next section we explore how these probabilities can be measured efficiently from given data sets, through the *joint* *k*NN-CDF.

2.2 Measurement of joint *k*NN-CDF

To set-up the measurements for the joint *k*NN-CDF, we populate the volume of interest, $V_{\rm tot}$, typically representing the simulation volume throughout this paper, with $N_{\rm R}$ random points. Using higher numbers of random points to sample the volume leads to lower measurement noise in the cumulative distribution function (CDF). Let the two sets of tracers for which we want to compute the joint clustering have N_{D_1} and N_{D_2} points each, distributed over the volume under consideration. We build two separate *k*-d trees (see e.g. Wald & Havran 2006) from each set of particle positions, and then use these trees to find and store the distance to the k NN data points from each random point. There are publicly available tree codes (e.g. SCIPY's CKDTREE implementation, and JULIA's NEARESTNEIGHBOR.JL[1] library) that can be used to efficiently carry out this calculation in $N\log N$ operations.

First, consider the case of $k = 1$ for both data sets – the distance from the randoms to the *first* NN data point in each set. For each random point therefore we can associate two distances – one to the NN data point from the first set, and the other to the NN data point from the second set. Now, for every random point, we choose the larger of the two distances. These distances are then sorted to get the empirical CDF of the distances chosen in this manner. We will refer to this distribution as the *joint* NN-CDF, $\mathrm{CDF}_{1,1}$. This joint NN-CDF can be interpreted in the following way: at a fixed radius r (or the corresponding spherical volume $V = 4/3\pi r^3$), the value of the CDF represents the fraction of spheres for which the distance to the NN data point in *both* sets are smaller than r. It is, therefore, equivalent to the fraction of spheres of radius r which contains at least one data point of the first data set *and* the second data set, i.e.

$$\mathrm{CDF}_{1,1}(r) = P(> 0, > 0|V). \quad (7)$$

It is easy to generalize this argument to show that

$$\mathrm{CDF}_{k_1,k_2}(r) = P\left(> k_1 - 1, > k_2 - 1|V\right), \quad (8)$$

where the CDF_{k_1,k_2} is the CDF computed as outlined above, but by considering the distance to the k_1th NN data point from set 1 and the distance kth NN data point from set 2 from every random point. Through the formalism in Section 2.1 therefore the *joint* *k*NN-CDF as defined here are sensitive to all the connected N-point correlation functions that can be formed from the two underlying fields from which the data points are sampled.

It is also possible, through the measurements of the NN distributions, to isolate those parts of $P(> k_1, > k_2|V)$ that depend *only* on the cross-correlation of the two fields, and not on the clustering of the two fields individually. To do this, we use equation (6) to obtain the prediction for the measurement of $P(> k_1, > k_2|V)$ in the uncorrelated scenario – in this case, it is just a product of the individual probabilities, $\mathcal{P}_1(> k_1|V)$ and $\mathcal{P}_2(> k_2|V)$. As shown in Banerjee & Abel (2021), $\mathcal{P}(> k|V) = \mathrm{CDF}_{k+1}(r)$, where $\mathrm{CDF}_{k_i}(r)$ represents the Empirical CDFs for the distances to the kth NN data points computed separately for each data set (labelled by i). Since these individual distances from the random points in the box to the kth NN of each data set are measured anyway as one of the steps toward building up the joint CDF, there is no significant additional resources needed to sort the distances and compute the individual CDFs. Therefore, by subtracting the product of the relevant k_ith NN distributions of each set of tracers from the *joint* NN measurements,

[1] https://github.com/KristofferC/NearestNeighbors.jl

we are left with the piece $\psi^{(k_1,k_2)}(r)$ that parametrizes the cross-correlation of the two fields:

$$\psi^{(k_1,k_2)}(r) = \mathrm{CDF}_{k_1,k_2}(r) - \mathrm{CDF}^{(1)}_{k_1}(r)\mathrm{CDF}^{(2)}_{k_2}(r), \tag{9}$$

where the superscripts on the RHS indicate the set of tracers for which the CDF has been computed. Higher absolute values of $\psi^{(k_1,k_2)}(r)$, for fixed amplitudes of clustering in the individual fields, indicate higher levels of spatial correlation between fields 1 and 2 at scale r – positive values indicate positive correlations, while negative values indicate anticorrelations in the two fields. If $\psi^{(k_1,k_2)}(r) = 0$, the two sets of tracers are uncorrelated. We reiterate that this implies not only an absence of the two-point cross-correlations, but a complete statistical independence of the two fields, as captured by all possible combinations of N-point functions of the two fields.

It is worth noting here the computational expense associated with these measurements. We focus on a typical example where we use 2×10^5 tracers in each set, distributed over a $(1h^{-1}\mathrm{Gpc})^3$ volume. The number of randoms, from the NN distances are measured, was 4×10^6 distributed uniformly over the same volume. We measure $\psi^{(k_1,k_2)}(r)$ for $k_1 = k_2 = k \in \{1, 2, 3, 4\}$. The entire measurement takes ∼45 s on a single core. The tree construction for each sample only takes less than a second, while each tree query takes ∼18 s. Since the largest cost for typical parameter choices is associated with the tree search for the NNs from the randoms, parallelizing this part can further reduce runtime – SCIPY's CKDTREE implementation, for example, already allows for this through the `njobs` flag. We also note that the computational expense does not scale strongly with the value of k, especially for the range of values of k we use in this study – the runtime with $k_1 = k_2 = 1$ is roughly similar to that in the example above.

In the rest of the paper, we will generally consider data sets with roughly equal number densities. Even if the data sets have different number densities originally, such as dark matter particles and haloes in a simulation, we will usually downsample the denser one to match the number density of the sparser sample. Having matched the number densities, we will generally consider $\mathrm{CDF}_{k,k}(r)$ in our analysis, where $k_1 = k_2 = k$. However, it should be noted that these are merely choices driven by considerations of the scales of interest, as well as computational time. One could also consider the joint NN distributions when $k_1 \neq k_2$. This may be especially relevant and appropriate when the two data sets have very different number densities.

3 CORRELATED GAUSSIAN FIELDS

In this section, we demonstrate the application of the measurement method outlined in Section 2 to the tracers of two correlated Gaussian fields. For completely Gaussian fields, the expression for the generating function in equation (1) can be truncated by only retaining terms up to the 2-point correlation functions – all higher order terms can be set to 0. Writing this out explicitly, we have

$$\begin{aligned} P(z_1, z_2|V) = \exp\Big[&\bar{n}_1(z_1-1)V + \bar{n}_2(z_2-1)V \\ &+ \frac{1}{2}\bar{n}_1^2(z_1-1)^2\bar{\xi}_V^{(2,0)} + \frac{1}{2}\bar{n}_2^2(z_2-1)^2\bar{\xi}_V^{(0,2)} \\ &+ \bar{n}_1\bar{n}_2(z_1-1)(z_2-1)\bar{\xi}_V^{(1,1)}\Big], \end{aligned} \tag{10}$$

where

$$\bar{\xi}_V^{(k_1,k_2)} = \int_V \mathrm{d}^3\boldsymbol{r}_1 ... \mathrm{d}^3\boldsymbol{r}_{k_1} \mathrm{d}^2\boldsymbol{r}'_1 ... \mathrm{d}^3\boldsymbol{r}'_{k_2} \xi^{(k_1,k_2)}, \tag{11}$$

and $\bar{n}_i$ represent the mean number density of each set of tracers. In this notation $\bar{\xi}_V^{(2,0)}$ and $\bar{\xi}_V^{(0,2)}$ represent the two point autocorrelation functions of fields 1 and 2, respectively, integrated over volume V. $\bar{\xi}_V^{(1,1)}$ represent the 2-point cross-correlation of fields 1 and 2 integrated over volume V. Note that the integrated cross-correlation can be negative if the two fields are anti-correlated. Using the fact that for a single Gaussian field (see Banerjee & Abel 2021 for details)

$$P(z|V) = \exp\left[\bar{n}(z-1)V + \frac{1}{2}\bar{n}^2(z-1)^2\bar{\xi}_V^{(2)}\right], \tag{12}$$

the full expression for generating function of the joint cumulative counts in equation (3), $C(z_1, z_2|V)$, can be written down, and the individual $\mathcal{P}(> k_1, > k_2|V)$ can be evaluated from the derivatives. The functional forms of the first few terms, are shown in Appendix B, and we will use these expressions to compare with the measurements outlined below.

For the measurements, we consider two $(1h^{-1}\mathrm{Gpc})^3$ simulations with 512^3 CDM particles run from $z = 99$ to $z = 4$ at the *Planck* best-fitting cosmology (Planck Collaboration VI 2020). The two simulations have different realizations of the initial power spectrum, and therefore the final density fields should not be spatially correlated. At the redshift under consideration, the matter field is still sufficiently close to Gaussian for the purposes of this exercise. We randomly downsample the set of simulation particles from the first realization down to 2×10^5 tracer particles. We then choose a different set of 2×10^5 particles from the same realization, ensuring that the same particle does not end up in both data sets. We first measure the kNN-CDF for only the first set of 2×10^5 particles for $k = 1$ and $k = 2$. These measurements are represented by the dashed lines on the upper left-hand and right-hand panels of Fig. 1, respectively. We then perform the *joint* NN CDF measurements for $k_1 = k_2 = 1$ and $k_1 = k_2 = 2$ using both sets of particles. These measurements are represented by the purple solid lines in the upper left-hand and right-hand panels of Fig. 1. Since both sets have been sampled from the same realization, we expect them to be spatially correlated. Next, we choose a random set of 2×10^5 particles from the second realization, and measure the joint NN measurements for $k_1 = k_2 = 1$ and $k_1 = k_2 = 2$ between this set of particles and one of the set of downsampled particles from the first realization. These measurements are represented by the orange solid line in the top panels (left and right) of Fig. 1. Since the two sets of particles in this latter case are from two different realizations, they are expected to be spatially uncorrelated.

Since the fields at $z = 4$ are still close to Gaussian, we use the COLOSSUS software[2] to compute the variance of the matter field fluctuations as a function of scale $\sigma(r)$, as expected from linear perturbation theory. We note that COLOSSUS by default uses the Eisenstein-Hu approximation (Eisenstein & Hu 1998) for the matter power spectrum, and is accurate at the level of $\lesssim 5$ per cent. However, this is sufficient for the purposes of this exercise: to demonstrate differences in the joint NN distributions of correlated and uncorrelated samples. The output from COLOSSUS is used to evaluate the analytic expectations for the joint NN-CDFs for both

[2] http://www.benediktdiemer.com/code/colossus/

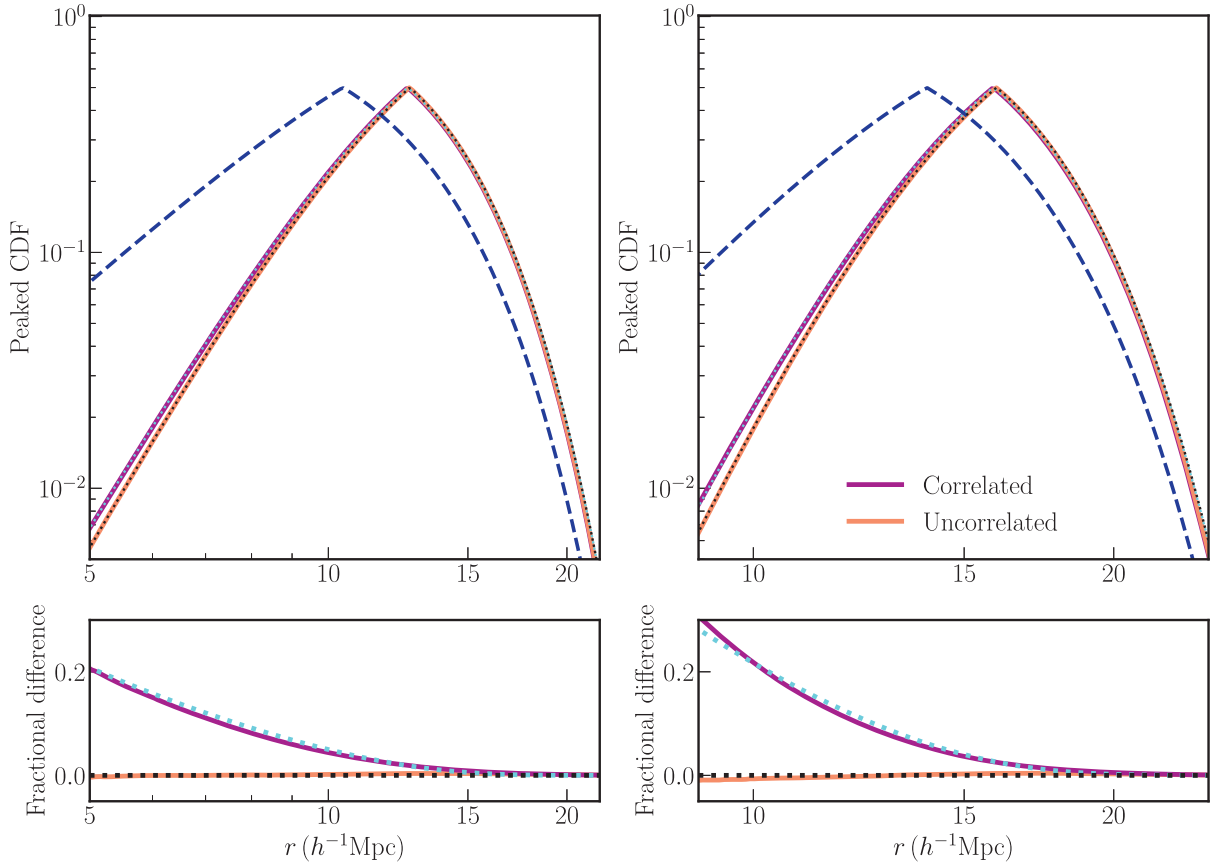

Figure 1. Top panels: The solid lines represent the peaked CDF of the *joint* 1NN (top left-hand panel) and *joint* 2NN (top right-hand panel) distributions for two correlated (the purple lines) and two uncorrelated (the orange lines) sets, each composed 2×10^5 tracers of a Gaussian random field over a $(1h^{-1}\mathrm{Gpc})^3$ volume (see Section 3 for details). The dotted lines indicate the theoretical expectations for these measurements. The dashed lines represent the 1NN (left-hand panel) and 2NN (right-hand panel) measurements for only one of the tracer sets, shown as a reference. Bottom panels: We plot the fractional differences of the predictions and measurements from the upper panels with respect to the analytic predictions for the uncorrelated sets for the joint 1NN (bottom left-hand panel) and joint 2NN (bottom right-hand panel) distributions. The differences in the joint CDFs between the correlated and uncorrelated data sets are especially clear on small scales, and match well with the analytic expectations. The different scales plotted on the left-hand and right-hand panels indicate the range of scales over which the distributions are well measured with the choice of measurement parameters mentioned in Section 3.

the correlated and uncorrelated sets, using equations (B2) and (B4). These predictions are plotted using the dotted lines in the two upper panels of Fig. 1. In the bottom panel of Fig. 1, we plot the fractional difference in the various predicted and measured CDF, using the analytic prediction for the uncorrelated case as the baseline. The bottom left-hand panel is for 1NN measurements, and the bottom right-hand panel is for the 2NN measurements. The colour scheme is the same as in the top panel. It is clear, especially at the smaller scales that there is a difference in the NN distributions between the two scenarios, and that these differences are captured correctly by the analytic predictions. Note that in the uncorrelated case, the prediction is that the joint NN-CDFs are simply a product of the individual NN-CDFs of each data set.

The cosmological matter density field, and therefore the distribution of tracer particles, is even closer to a true Gaussian random field at higher redshifts. However, since the amplitude of the fluctuations are also lower at these higher redshifts, it is difficult to visualize the differences between the correlated and uncorrelated scenarios in Fig. 1. As we will see in the next section, the differences between correlated and uncorrelated data sets become clearer when the joint NN measurements are applied to fully non-linear fields, like the matter density field at $z = 0$.

4 CORRELATIONS IN NON-LINEAR FIELDS

We now focus on measuring joint kNN-CDF and cross-correlations, $\psi^{(k_1,k_2)}(r)$, in cosmological fields with non-linear clustering, i.e. clustering of matter and dark matter halo at low redshifts. Unlike in the Gaussian example presented in Section 3, the number of terms that are important in the generating functions, both for the individual counts, and for the joint counts, are not known beforehand. This implies that it is not always possible to write down an analytic expression for the counts in terms of the N-point correlation functions. However, the measurements of the individual and joint cumulative counts using the NN measurements can be be performed exactly as outlined in Section 2.2.

To demonstrate the measurement of cross-correlations in these non-linear fields, we follow the general outline of Section 3 and consider the following example: we take a $(1h^{-1}\mathrm{Gpc})^3$ simulation at $z = 0$ with 512^3 particles and choose a random subset of 10^5 particles. We measure the NN distributions for this single set of particles and plot the peaked CDF using the blue-dashed lines in the Fig. 2 – the left-hand panel shows the $k_1 = k_2 = 1$ NN distribution while the right-hand panel shows the $k_1 = k_2 = 2$ NN distribution. Next, we choose another random subset of 10^5 particles from the same simulation –

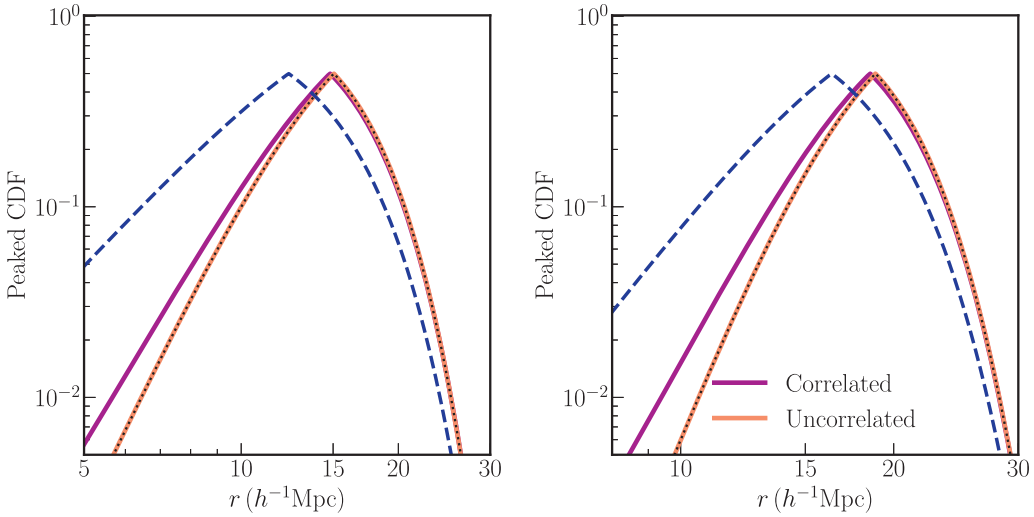

Figure 2. The solid lines represent the peaked CDF of the *joint* 1NN (left-hand panel) and *joint* 2NN (right-hand panel) distributions for two correlated (the purple lines) and two uncorrelated (the orange lines) sets of simulation particles at $z = 0$, when the matter field is highly non-linear on small scales. Each set has 10^5 particles downsampled from a $(1h^{-1}\mathrm{Gpc})^3$ simulation with 512^3 particles. The dashed line on each panel represents the first and second NN peaked CDF for a single set of particles for reference. The dotted line in each panel represents the expectation for the joint 1NN and 2NN CDFs of two uncorrelated sets of particles given the measurements of their individual 1NN and 2NN distributions. Deviations from this dotted line in each panel represents the degree of cross-correlation between the data sets. The range of scales on each panel represents the range over which the distributions are well measured for the specific choice of parameters. See text for more details.

we do not use any of the particles that were part of the first sample. Since both sets of particles are drawn from the same simulation, and therefore trace the same underlying density field, they should be fully correlated, modulo sampling noise. We now measure the *joint* NN distribution for these two sets of tracers. The results of the $k_1 = k_2 = 1$ and $k_1 = k_2 = 2$ joint NN peaked CDF are represented by the purple solid lines in the left-hand and right-hand panels of Fig. 2, respectively.

We then take another simulation with the same resolution and at the same cosmology, but with a different random realization of the initial density field. We randomly select 10^5 particles from this simulation. Since the initial modes for this realization and that of the first simulation are different, the final density fields do not align with each other, and therefore, the tracers are also not expected to have any statistical correlations in their clustering. We measure the joint NN distributions for the two sets of 10^5 particles from the two realizations. The results of the $k_1 = k_2 = 1$ and $k_1 = k_2 = 2$ joint NN peaked CDF are represented by the orange solid lines in the left and right-hand panels of Fig. 2, respectively.

Unlike in the Gaussian case, there is no simple analytic expectation for the joint kNN-CDF in the most general correlated case. However, it is still possible to make a prediction for the joint kNN-CDF for uncorrelated samples based on the measurements of the *individual* kNN-CDF using equation (6). This expectation is plotted using the dotted black lines in both panels of Fig. 2, and agrees with the direct measurements for the uncorrelated samples on all scales displayed in the plot. The difference between the correlated and uncorrelated samples are especially pronounced on smaller scales where the clustering is stronger, but the difference persists out to the largest scales that we measure, and will become evident when considering $\psi^{(k_1,k_2)}(r)$ in the following subsections.

Having demonstrated the measurement of joint kNN-CDF measurements for non-linear cosmological fields, we explore two applications of these measurements below – one in the high signal-to-noise regime, where we can use cross-correlation measurements to infer cosmology, and one in the low signal-to-noise regime, where we focus simply on the detection of a signal. Given the scope of this paper, we will focus on the parts of the measurements, $\psi^{(k_1,k_2)}$, that capture cross-correlations in both applications.

4.1 Parameter constraints using cross-correlations

In this section, we use the Fisher matrix formalism to compare the information content, and sensitivity to the underlying cosmological parameters, of the two-point cross-correlation with that of the NN method of computing cross-correlations. For this exercise, we use data from the QUIJOTE suite of simulations[3] (Villaescusa-Navarro et al. 2020). These simulations, run over different cosmologies, have a volume of $(1h^{-1}\mathrm{Gpc})^3$ and use 512^3 CDM particles for cosmologies without massive neutrinos, and 512^3 CDM and 512^3 neutrino particles for cosmologies with massive neutrinos. We consider the cross-correlations of the 10^5 most massive haloes in the simulations with the underlying matter field at $z = 0$. Banerjee & Abel (2021) has demonstrated that NN measurements of the auto clustering of these data sets are also more sensitive to cosmological parameters than two-point measurements – here we demonstrate the same for only the cross-correlation piece. For the two-point cross-correlation, ξ^{hm}, we use CORRFUNC[4] (Sinha & Garrison 2020; Sinha & Garrison 2019) to compute the results over 30 bins between 8 and $30h^{-1}$Mpc. For the NN calculation, we downsample the simulation particles to 10^5 randomly selected particles, and compute the joint NN distributions for $k_1 = k_2 = k \in \{1, 2, 3, 4\}$ for these two data sets (haloes and downsampled particles). We use 8×10^6 random points over the simulation volume for the joint NN calculations, using 16 bins for each joint kNN-CDF, between 8 and $30h^{-1}$Mpc – the same range used for the two-point cross-correlations. We use measurements of $\psi^{(k,k)}(r)$ in our analysis, which isolates the cross-correlations, rather than the direct measurements of the joint kNN-CDF that are also

[3] https://github.com/franciscovillaescusa/Quijote-simulations
[4] https://github.com/manodeep/Corrfunc

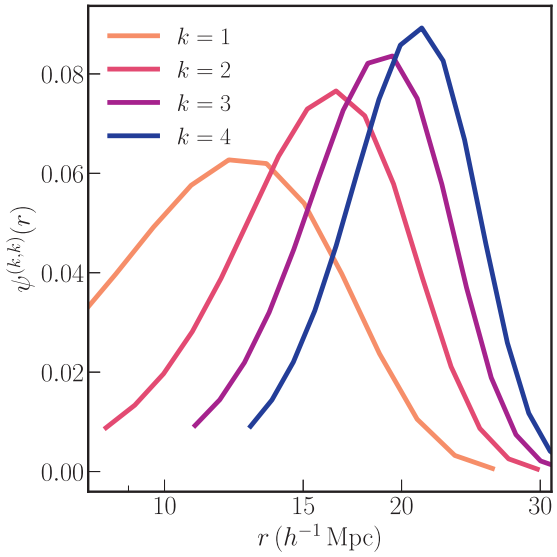

Figure 3. $\psi^{(k,k)}(r)$; see equation 9), which measures the spatial correlation between two samples, for various k measured from the 10^5 most massive haloes and 10^5 randomly chosen particles from a $(1h^{-1}\mathrm{Gpc})^3$ simulation at $z = 0$. These measurements are used in the Fisher matrix calculations in Section 4.1. For each k, we plot the measurements over the range of scales that are used in the analysis for that particular k.

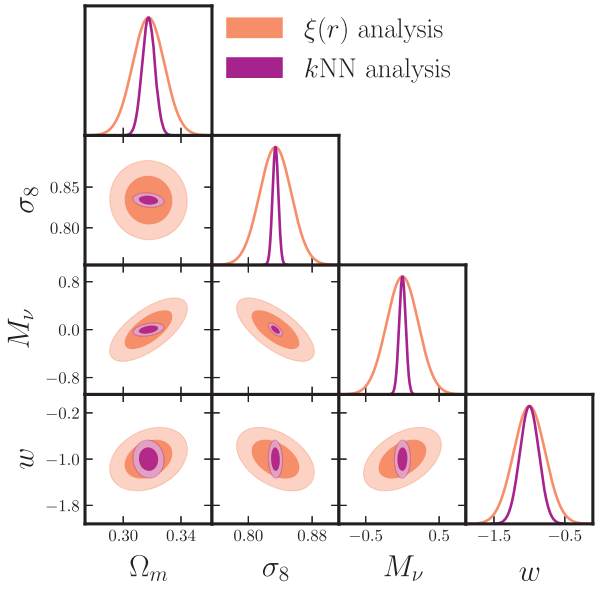

Figure 4. Constraints on various cosmological parameters from the Fisher analysis in Section 4.1. These are obtained from the cross-correlations of the 10^5 most massive haloes in the simulation volume ($(1h^{-1}\mathrm{Gpc})^3$) with the matter distribution. There is a marked improvement in the constraints when the cross-correlations are measured through the nearest neighbour distributions (kNN), compared to when measured through the two-point cross-correlation $\xi(r)$, over the same range of scales. The improvement is especially pronounced in some of the parameters, such as σ_8, and M_ν. The individual constraints are listed in Table 1.

sensitive to the clustering of each data set individually. We further restrict the range of scales for each k to ensure that the distributions are well measured. This is done by considering, for each k, only those range of scales for which the measurement of $\psi^{(k,k)}(r)$ is >0.005 at the fiducial cosmology of the QUIJOTE suite. This choice ensures that the measurements are not affected by the noise associated with the number of randoms used in the calculation, nor by the statistical fluctuations in the tails of the distribution. We plot $\psi^{(k,k)}(r)$ for different k, as measured from one of the simulations in the QUIJOTE suite in Fig. 3. The scales for which measurements from an individual value of k are used in the analysis can be clearly seen from the different ranges over which they are plotted in the figure.

We now briefly summarize the Fisher matrix formalism. We denote the data vector, either as measured through $\xi^{hm}(r)$ or through $\psi^{(k,k)}(r)$ as $\boldsymbol{D}$. The elements of the Fisher matrix are defined as

$$\boldsymbol{F}_{\alpha\beta} = \sum_{i,j} \frac{\partial D_i}{\partial p_\alpha} \left[\mathbf{C}^{-1}\right]_{ij} \frac{\partial D_j}{\partial p_\beta}, \tag{13}$$

where $\boldsymbol{p}$ represents the vector of cosmological parameters, and $\mathbf{C}$ represents the covariance matrix for the data vector. The Fisher matrix can be inverted to estimate how well various cosmological parameters are constrained by the particular data vector:

$$\sigma_\alpha = \sqrt{\left(\boldsymbol{F}^{-1}\right)_{\alpha\alpha}}. \tag{14}$$

In this paper, the cosmological parameters considered are $\{\Omega_m, \sigma_8, M_\nu, w\}$. The covariance matrix is computed from the data vector computed over 1000 realizations at the fiducial cosmology of the QUIJOTE suite. The raw covariance matrix is computed as

$$\mathbf{C}'_{ij} = \langle (D_i - \langle D_i \rangle)(D_i - \langle D_i \rangle) \rangle\,, \tag{15}$$

where $\langle \boldsymbol{D} \rangle$ is the mean data vector averaged over the 1000 realizations. We test for the stability of the Fisher analysis by checking that the condition number of the matrix is reasonable before inversion, and by checking that the distribution of values in each bins is roughly

Table 1. 1σ constraints on cosmological parameters, from the kNN and $\xi(r)$ analysis of the cross-correlation of the matter field with the 10^5 most massive haloes in the box. The same range of scales are used in both analyses.

Parameter	$\sigma_{k\mathrm{NN}}$	$\sigma_{\xi(r)}$
Ω_m	0.0061	0.0109
σ_8	0.0036	0.0191
ΔM_ν	0.0451	0.2161
w	0.1344	0.2239

Gaussian around the mean. For each data vector considered, we use the Hartlap factor to correct for the fact that a finite number of realizations are used to estimate the covariance matrix (Hartlap, Simon & Schneider 2007):

$$\mathbf{C}^{-1} = \frac{n - p - 2}{n - 1} \left(\mathbf{C}'\right)^{-1}, \tag{16}$$

where n represents the number of realizations, while p represents the length of the data vector. This corrected covariance matrix is used in equation (13). The QUIJOTE suite is designed to allow for the derivatives of the data vectors with respect to the cosmological parameters to be easily evaluated, since the cosmologies are changed by one parameter at a time. We average over 100 realizations at each of the derivative cosmologies to compute the data vector derivatives required in equation (13).

The results of the Fisher analysis, in terms of constraints on various cosmological parameters, and their covariance are presented in Fig. 4. The constraints on individual parameters are also summarized in Table 1. The orange contours and standard Fisher errors in individual

parameters in Fig. 4 represent the results from the analysis of the two-point cross-correlation, $\xi(r)$ while the purple contours and standard Fisher errors represent the results from the NN (kNN) analysis. We find that the NN cross-correlations are more sensitive to cosmological parameters compared to the two-point cross-correlations. This is true for all the cosmological parameters considered here, but the improvement in constraints is especially pronounced for σ_8 and M_ν. The degeneracy directions are also somewhat different in the two cases, which suggests further improvements in parameter constraints if the two measurements are combined. The overall improvements seen in Fig. 4 and Table 1 are not surprising, since we have shown that the two-point cross-correlations are a subset of all the terms that contribute to the measurements of $\psi^{(k,k)}$. Of course, a full analysis would consider both the cross-correlations of the matter field and haloes, as well as the autoclustering of each sample. Since the NN measurements are considerably more sensitive to each of these (see Banerjee & Abel 2021 for discussions on the autoclustering), we can conclude that NN measurements is a promising tool for all clustering measurements in cosmological analyses.

We now discuss a few caveats about the analysis presented above. First, it worth noting that the purpose of this exercise is to show the relevant improvement in cosmological parameter constraints when using kNN measurements of cross-correlations over two-point cross-correlations in an idealized scenario. Therefore, the absolute numbers presented in Fig. 4 and Table 1 should be interpreted accordingly. We have used the simulation volume of $(1h^{-1}\mathrm{Gpc})^3$ throughout our analysis – current cosmological surveys have orders of magnitude larger volumes, and the absolute constraints are expected to be even tighter, modulo survey systematics. Secondly, we have only used measurements for the first four NNs in this analysis. As pointed out in Section 2.2, extending to higher k does not add any significant computational cost. Including these higher k measurements can further improve the constraints from the kNN cross-correlations, as shown in Banerjee & Abel (2021). However, it should be kept in mind that increasing the length of the data vector, at fixed number of realizations, reduces the Hartlap factor, and the analysis might be unstable if the data vector is too large. Thirdly, we have only used measurements of joint kNN-CDFs when $k_1 = k_2$. It is also possible to consider cases where $k_1 \neq k_2$, again, without significant additional costs. Since each combination has a unique expression in terms of various correlation functions (see Section 2.1), adding them to the analysis can also help improve sensitivity to various cosmological parameters. The limitation, once again, would be ensuring that the data vector size does not become comparable to the number of realizations used to estimate the covariance matrix.

4.2 Detecting cross-correlations in sparse samples

As a final example of the use of joint kNN-CDF for describing spatial correlations of different data sets, we consider the problem of detecting spatial correlations in sparse samples. Under these conditions, measurement noise from a finite number of tracers can dominate the signal even if a true correlation exists in the underlying continuous fields.

As a concrete example, we choose 1000 haloes at random from the most massive 10^5 haloes in one of the realizations from the fiducial cosmology of the QUIJOTE simulations. Next, we choose another 1000 haloes at random (without replacement from the ones chosen in the first set) from the same box. These two sets of haloes should be correlated spatially since they both trace the same underlying field. We then compute the two-point correlation function between these two sets, $\xi_d(r)$ using CORRFUNC, as well as by computing the joint first NN distribution $\mathrm{CDF}_{1,1}(r)$ over the same set of scales – 10–$100h^{-1}$Mpc, using 40 measurement bins. Once again, since we are looking to quantify the cross-correlation between the sets, we remove the signal from the clustering of the individual data sets by using equation (9), and using only $\psi_d^{(1,1)}(r)$ in the rest of our analysis. For the NN measurements, we use 4×10^6 random points distributed over the simulation volume.

Next, keeping the first sample of 1000 haloes fixed, we iterate over 1000 other realizations of the fiducial cosmology of the QUIJOTE simulations. For each realization, we choose 1000 haloes at random from the 10^5 most massive haloes in the box. Note that since the cosmology is held fixed here, the amplitude of clustering in all of these halo samples should roughly be the same. We then compute the spatial correlations – in terms of both the two-point cross-correlation, and the joint kNN distributions – of these 1000 samples from different realizations with the original sample of haloes. For halo samples drawn from different realizations of the same cosmology, no correlation is expected, since they sample different modes, and therefore the spatial fluctuations of counts are not correlated with each other.

Using the measurements of cross-correlations from the 1000 uncorrelated samples, we find the mean signal, along with the full covariance matrix. First, we consider the two-point correlation function measurements. Let us denote the mean data vector of the measurements of uncorrelated samples by $\xi_0(r)$, and the covariance matrix by $\mathbf{C}$. For sufficiently large number of realizations, we should have $\xi_0(r) \to 0$. The values of $\xi(r) - \xi_0(r)$ for 50 of the the realizations are plotted using the (orange) solid lines on the left-hand panel of Fig. 5. Notice that the measured signal is not 0, even though the samples should be uncorrelated – this is due to noise introduced by the sparsity of the samples. Next, the χ^2 value can be found for each of the 1000 realizations by computing

$$\chi_i^2 = (\xi_i(r) - \xi_0(r))^T \, \mathbf{C}^{-1} \, (\xi_i(r) - \xi_0(r)) \,, \tag{17}$$

where the label i denotes any one of the 1000 realizations. The distribution of χ^2 over all the realizations is represented by the orange coloured histogram in Fig. 6. We use measurements from 40 bins, and so nominally have $\sim$40 degrees of freedom, as some of the measurements will be correlated. This is consistent with the fact that the χ^2 distribution peaks near 40.

For the case where there is an underlying correlation between the halo samples, we plot $\xi_d(r) - \xi_0(r)$ using the dashed (purple) line in the left-hand panel of Fig. 5. We also compute the value of χ^2 in this instance by replacing $\xi_i(r)$ by $\xi_d(r)$ in equation (17). This value is represented by the dotted vertical line in Fig. 6. Notice that this value lies well within the range of χ^2 obtained for samples drawn from the null hypothesis – i.e. uncorrelated samples. This is mainly due to the fact that the samples are so sparse, and therefore the measurement noise dominates over the true signal. This latter fact can also be seen by comparing the dashed and solid lines in the left-hand panel of Fig. 5.

We repeat this procedure outlined above for the joint NN measurements. It should be noted that the χ^2 statistic is reliable when the distribution of of the statistics under consideration is Gaussian around the mean for the null hypothesis. While there is no *a priori* reason for this to be true of any general statistic under consideration, such as the joint NN measurements presented here, we have checked explicitly that their distribution is roughly Gaussian around the mean in every bin. Further, as we demonstrate below, the joint NN measurements produce an extremely close χ^2 distribution to that obtained from

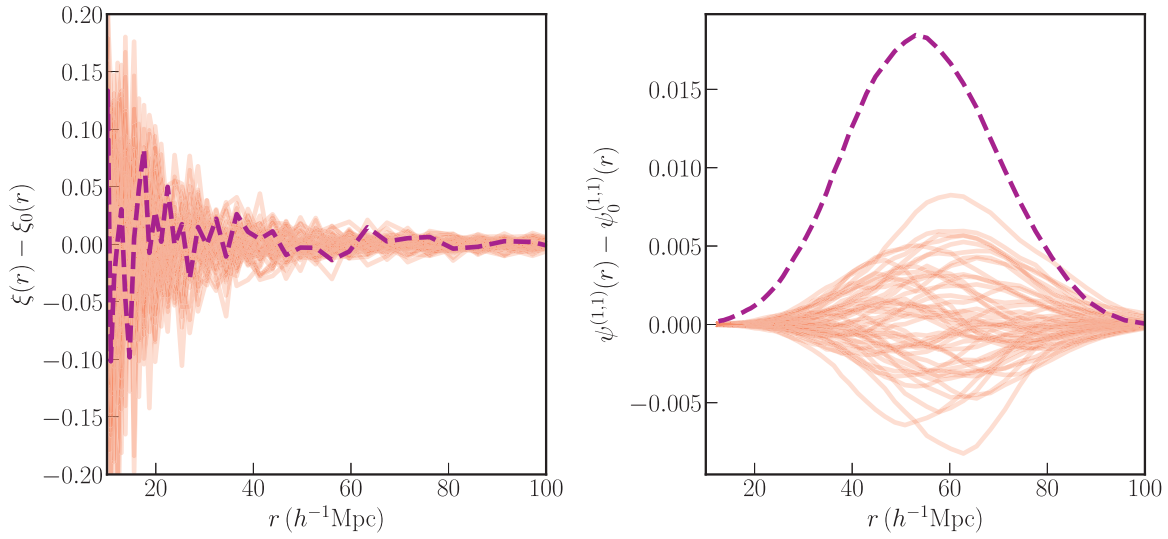

Figure 5. Left-hand panel: Difference of the two-point cross-correlation measurements of two sets of dark matter haloes (1000 haloes each) from the mean of 1000 such measurements where the two sets are spatially uncorrelated (drawn from different realizations). The orange solid lines represent the difference for 50 of these uncorrelated samples, meant to serve as a visual measure of the spread in the measurements when there are no true correlations. The purple-dashed line represents the measurement of the same quantity in the case when the two sets of haloes are from the simulation, and therefore correlated. Right-hand panel: Same measurements as in the left-hand panel, but using $\psi^{(1,1)}(r$; see equation 9) to measure cross-correlations instead of the two-point cross-correlation. Using $\psi^{(1,1)}(r)$, the correlated measurement is clearly separated from the uncorrelated ones. See Section 4.2 for more details.

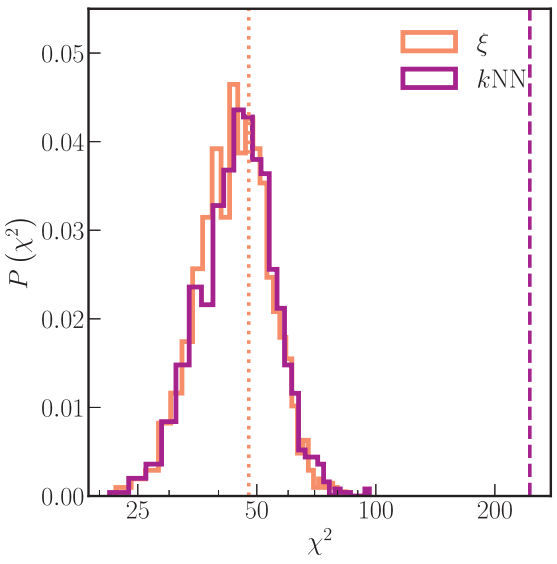

Figure 6. The solid lines represent the binned distribution of χ^2 values for 1000 measurements of cross-correlations between two samples of haloes (1000 haloes each from a $(1h^{-1}\mathrm{Gpc})^3$ volume) that are spatially uncorrelated. The orange line represents the distribution when cross-correlations are measured through the two-point function (ξ), while the purple line represents the distribution when kNN measurements ($\psi^{(1,1)}$) are used to measure the cross-correlation. The dotted line represents the value of χ^2 in the case when the two halo samples are spatially correlated, and when the cross-correlation is measured through ξ. The dashed line represents the χ^2 value when the cross-correlation of these samples is measured via the nearest neighbour distribution. The cross-correlation is clearly detected in the latter measurement, as seen by the χ^2 value being far to the right of the distribution for the uncorrelated samples (p-value $< 10^{-3}$). The two-point measurement, on the other hand, fails to detect a statistically significant correlation.

considering $\xi(r)$ measurements with the same number of bins, that is, the same nominal degrees of freedom. This is also consistent with the assumption that the NN measurements are close to being Gaussian distributed, and that the χ^2 statistic is meaningful to use in this context. We denote the mean of the measurement for uncorrelated samples by $\psi_0^{(1,1)}(r)$ and the covariance matrix by $\tilde{\mathbf{C}}$. The values of $\psi^{(1,1)}(r) - \psi_0^{(1,1)}(r)$ for 50 of the uncorrelated samples are plotted using the (orange) solid lines on the right-hand panel of Fig. 5. The χ^2 values are obtained in this case by simply replacing $\xi(r)$ by $\psi^{(1,1)}(r)$, and $\mathbf{C}$ by $\tilde{\mathbf{C}}$ in equation (17). The distribution of χ^2 for uncorrelated samples as measured through the joint NNs is represented by the purple histogram in Fig. 6. As with the two-point correlation function, 40 bins were used, and the distribution peaks near 40, roughly 1 per degree of freedom. Comparing the two histograms, we see that the two-point correlation function and the NN measurements perform the same for uncorrelated samples – i.e. in the absence of a true clustering signal. Next, we consider the NN measurements for the correlated sample. We plot $\psi_d^{(1,1)}(r) - \psi_0^{(1,1)}(r)$ with the dashed (purple) line on the right-hand panel of Fig. 5. For the NN measurements, it is clear that the measurement for the correlated sample is clearly separated from the measurements on the uncorrelated samples. To quantify this separation, we compute the χ^2 value for the correlated samples, using the measurement of $\psi_d^{(1,1)}(r)$. This value is represented by the dashed vertical line in Fig. 6, and, unlike in the two-point correlation case, is far to the right of the distribution of χ^2 values for the uncorrelated samples. Since we used 1000 different uncorrelated halo samples to characterize the χ^2 distribution, and none of these produce as large a χ^2 value, we can summarize this difference in terms of the p-value: $p < 10^{-3}$. Note that the shape of the distribution of the χ^2 values, and the actual value of χ^2 obtained for the correlated sample, suggest that actual p-value is likely even smaller. Therefore, the hypothesis that the two halo samples from the same underling distribution are uncorrelated is highly disfavored when the correlation is measured through the NN method.

From this simple example, we can conclude that spatial correlations between sparse data sets can be detected more robustly when using the joint NN distributions compared to the two-point correlation functions. The fact that the χ^2 distributions for the uncorrelated samples, as measured through either $\xi(r)$ or $\psi^{(1,1)}(r)$, match closely indicates that the joint NN distribution is truly detecting the underlying cross-correlation, rather than being a numerical artefact. We note once again that the same input data, i.e. the same sets of halo positions, is used in both calculations – the difference lies only in the way the data is summarized. It is also worth noting that we have only used $\mathrm{CDF}_{1,1}$ to capture the correlations. In principle, we could have used other joint *k*NN measurements as well, adding to the statistical significance of the detection. Since our purpose here is only to demonstrate the usefulness of the joint NN correlation measure, and given that the $\mathrm{CDF}_{1,1}$ is already sufficient to demonstrate this, we do not delve further into combining the higher *k*NN measurements in this section. For real data sets, the choice of which joint *k*NN distributions to use will depend on the goals of the analysis, as well as the features of the data sets under consideration.

5 SUMMARY AND DISCUSSION

In this paper, we have extended the NN framework for measuring clustering developed in Banerjee & Abel (2021) to include the joint NN distributions of two different data sets and use these to measure cross-correlations in the spatial clustering of the two. We have demonstrated that this way of measuring cross-correlations is generally more powerful than through two-point cross-correlations. Since cross-correlations between data sets are used widely and in various different contexts in cosmology, this extension greatly increases the range of analyses that the NN measurements can be applied to. We now summarize the main points of this work.

We have shown that *joint* *k*NN-CDFs are sensitive to all *N*-point correlation functions that can be formed from the two fields whose tracers we consider. Therefore, these *k*NN-CDF can capture correlations beyond just the linear correlations between data sets – the latter is what is generally measured through the two-point cross-correlations. We have outlined how the relevant distributions can be computed from data. This is done by considering the distances from random points in the volume to the k_1th NN from the first set, and the k_2th NN from the second set, and then considering the distribution of the larger of these two distances. The cross-correlation piece can be isolated by subtracting the product of the individual *k*NN (for each data set) distributions from the joint *k*NN distribution. We reiterate that this entire calculation is really fast and can be performed in less than a minute on a single core for typical applications explored in this paper.

We have demonstrated the applications of these measurements, first in the context of Gaussian fields, where it is possible to analytically predict these distributions for both correlated and uncorrelated tracers, and then in the context of fully non-linear cosmological fields. In the latter context, one can still predict the joint *k*NN distribution for uncorrelated samples in terms of the measurement of the *k*NN measurements on each sample separately. Using a Fisher matrix formalism, we have shown that cross-correlations of massive haloes with the underlying matter field as measured through *k*NN measurements are more sensitive to the underlying cosmological parameters, compared to measurements of the two-point cross-correlation. Finally, we have demonstrated that the NN measurements can robustly detect cross-correlations in low number density samples where the two-point cross-correlation measurements are dominated by measurement noise. For both of these applications, we use only those parts of the *k*NN measurements that come only from the cross-correlations of the data sets, to enable a direct comparison to the two-point cross-correlation function. Unlike the two-point cross-correlation, a lack of detection of cross-correlations using the *k*NN distributions imply a complete statistical independence of the two distributions from which the tracers are drawn.

It is worth noting that the specific method outlined here is not the only NN measurement approach that is sensitive to cross-correlations of two data sets. For example, combining the two sets of tracers, and performing *k*NN measurements on this combined set should also, in principle, be sensitive to the cross-correlations of the two sets. Another possible way to measure cross-correlations in this framework is to compute distance to a specific *k*th NN data point from the first data set, and then computing the distribution of how many data points from the second set are found within that volume. However, there are certain attractive features of the method presented here: first, the expression for the joint *k*NN-CDFs can be conveniently expressed in terms of the various *N*-point correlation functions formed from the two fields. Secondly, it is both conceptually and computationally easy to isolate those parts of the measurement which depends on just the cross-correlations. This factorization is useful for various common applications in cosmology, and is not guaranteed for other ways of measuring cross-correlations with NNs.

Another aspect worth noting is that the joint *k*NN measurements outlined here are applied to cross-correlations between two samples. It is possible to extend this formalism to include more data sets. the simplest way to consider various pairs of data sets and consider their cross-correlations – this is routinely done using the two-point cross-correlations. However, it is also possible to consider multiple sets of tracers at the same time, and consider whether they are all drawn from the same distribution.

Lastly, we have focused on cross-correlations between discrete data sets in this paper. However, as illustrated in Section 3, the *k*NN measurements on discrete tracers are able to capture the cross-correlations of the underlying continuous fields, even at the level of downsampling we have used – the *k*NN measurements are performed on only 2×10^5 particles out of the 512^3 particles that describe the density field in the simulations. This is similar to the findings for the *k*NN autocorrelation measurements in Banerjee & Abel (2021). Taken together, this implies that the formalism presented here can be extended to measuring cross-correlations between continuous maps, by simply sampling the maps correctly with a fixed number of tracers. We will explore this aspect in detail in future, to widen the range of applications of *k*NN-CDF cross-correlations to data sets which are inherently continuous.

ACKNOWLEDGEMENTS

This work was supported by the Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, the U.S. Department of Energy (DOE) Office of Science Distinguished Scientist Fellow Program, and the U.S. Department of Energy SLAC Contract No. DE-AC02-76SF00515. The authors thank the referee for comments that helped improve the paper. The authors also thank Alvaro Zamora for comments on an earlier version of the paper. Some of the computing for this project was performed on the Sherlock cluster. The authors would like to thank Stanford University and the Stanford Research Computing Center

for providing computational resources and support that contributed to these research results. The PYLIANS3[5] analysis library was used extensively in this paper. We also acknowledge the use of the GETDIST[6] (Lewis 2019) software for plotting.

DATA AVAILABILITY

The simulation data used in this paper is publicly available at https://github.com/franciscovillaescusa/Quijote-simulations. Additional data are available on reasonable request.

REFERENCES

Abbott T. M. C. et al., 2018, Phys. Rev. D, 98, 043526
Abbott T. M. C. et al., 2019, Phys. Rev. D, 100, 023541
Ammazzalorso S. et al., 2020, Phys. Rev. Lett., 124, 101102
Banerjee A., Abel T., 2021, MNRAS, 500, 5479
Baxter E. et al., 2016, MNRAS, 461, 4099
Bianchini F. et al., 2015, ApJ, 802, 64
Blake C., Pope A., Scott D., Mobasher B., 2006, MNRAS, 368, 732
Croft R. A. C., Dalton G. B., Efstathiou G., 1999, MNRAS, 305, 547
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Fang K., Banerjee A., Charles E., Omori Y., 2020, ApJ, 894, 112
Granett B. R., Neyrinck M. C., Szapudi I., 2008a, preprint (arXiv:0805.2974)
Granett B. R., Neyrinck M. C., Szapudi I., 2008b, ApJ, 683, L99
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Heymans C. et al., 2021, A&A, 646, A140
Joudaki S. et al., 2018, MNRAS, 474, 4894
Kirk D. et al., 2016, MNRAS, 459, 21
Lewis A., 2019, preprint (arXiv:1910.13970)
Li D., Yalinewich A., Breysse P. C., 2019, preprint (arXiv:1902.10120)
Mandelbaum R., Slosar A., Baldauf T., Seljak U., Hirata C. M., Nakajima R., Reyes R., Smith R. E., 2013, MNRAS, 432, 1544
Miyatake H. et al., 2015, ApJ, 806, 1
More S., Miyatake H., Mandelbaum R., Takada M., Spergel D. N., Brownstein J. R., Schneider D. P., 2015, ApJ, 806, 2
Munshi D., Joudaki S., Coles P., Smidt J., Kay S. T., 2014, MNRAS, 442, 69
Namikawa T. et al., 2019, ApJ, 882, 62
Paech K., Hamaus N., Hoyle B., Costanzi M., Giannantonio T., Hagstotz S., Sauerwein G., Weller J., 2017, MNRAS, 470, 2566
Planck Collaboration VI, 2020, A&A, 641, A6
Rhodes J. et al., 2013, preprint (arXiv:1309.5388)
Rizzo L. A., Mota D. F., Valageas P., 2017, A&A, 606, A128
Salcedo A. N., Wibking B. D., Weinberg D. H., Wu H.-Y., Ferrer D., Eisenstein D., Pinto P., 2020, MNRAS, 491, 3061
Schaan E., Krause E., Eifler T., Doré O., Miyatake H., Rhodes J., Spergel D. N., 2017, Phys. Rev. D, 95, 123512
Schneider P., Watts P., 2005, A&A, 432, 783
Seljak U. et al., 2005, Phys. Rev. D, 71, 043511
Singh S., Mandelbaum R., Brownstein J. R., 2017, MNRAS, 464, 2120
Singh S., Mandelbaum R., Seljak U., Rodríguez-Torres S., Slosar A., 2020, MNRAS, 491, 51
Sinha M., Garrison L., 2019, in Majumdar A., Arora R., eds, Software Challenges to Exascale Computing. Springer, Singapore, p. 3
Sinha M., Garrison L. H., 2020, MNRAS, 491, 3022
Szapudi I., Szalay A. S., 1993, ApJ, 408, 43
To C., et al., 2021, Phys. Rev. Lett., 126, 141301
Villaescusa-Navarro F. et al., 2020, ApJS, 250, 2
Wald I., Havran V., 2006, in 2006 IEEE Symposium on Interactive Ray Tracing, IEEE, Salt Lake City, UT, USA, p. 61
Wibking B. D., Weinberg D. H., Salcedo A. N., Wu H.-Y., Singh S., Rodríguez-Torres S., Garrison L. H., Eisenstein D. J., 2020, MNRAS, 492, 2872
Zu Y., Weinberg D. H., 2013, MNRAS, 431, 3319

[5]https://github.com/franciscovillaescusa/Pylians3
[6]https://getdist.readthedocs.io/en/latest/

APPENDIX A: DERIVATION OF THE GENERATING FUNCTION FOR JOINT COUNTS

Here, we extend the formalism from Szapudi & Szalay (1993) and appendix A in Banerjee & Abel (2021) to two continuous fields, instead of one. For two continuous fields $\rho_1(\boldsymbol{r})$ and $\rho_2(\boldsymbol{r})$, the generating functional for all correlation functions can be written as an integral over all possible joint configurations of the two fields:

$$\mathcal{Z}[J_1, J_2] = \int [D\rho_1][D\rho_2]\mathcal{P}(\rho_1, \rho_2) \times \exp\left[i\int \mathrm{d}^3\boldsymbol{r}\,(\rho_1 J_1 + \rho_2 J_2)\right], \tag{A1}$$

where $\mathcal{P}(\rho_1, \rho_2)$ represent the joint distribution function of the two fields. Note that it is this joint distribution function that encodes any correlation between the two fields – in the absence of correlations, the joint distribution can be factorized into the individual distribution functions of the two fields. The connected N-point correlations, $\xi^{(k_1,k_2)}$, with k_1 factors of ρ_1 and k_2 factors of ρ_2, are given by the functional derivatives of the generating functional with respect to the sources J_1 and J_2:

$$\bar{\rho}_1^{k_1}\bar{\rho}_2^{k_2}\xi^{(k_1,k_2)}\left(\boldsymbol{r}_1, ..., \boldsymbol{r}_{k_1}; \boldsymbol{r}'_1, ..., \boldsymbol{r}'_{k_2}\right) = \frac{(-i)^{(k_1+k_2)}\delta^{(k_1+k_2)}\left(\log\mathcal{Z}[J_1, J_2]\right)}{\delta J_1(\boldsymbol{r}_1)...\delta J_1\left(\boldsymbol{r}_{k_1}\right)\delta J_2\left(\boldsymbol{r}'_1\right)...\delta J_2\left(\boldsymbol{r}'_{k_2}\right)}\bigg|_{J_1=0, J_2=0}, \tag{A2}$$

where $\bar{\rho}_i$ are the mean densities of the two fields. Conversely, the generating functional can be expressed in terms of the N-point correlation functions as

$$\mathcal{Z}[J_1, J_2] = \exp\left[\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\frac{(i\bar{\rho}_1)^{k_1}(i\bar{\rho}_2)^{k_2}}{k_1!k_2!}\int \mathrm{d}^3\boldsymbol{r}_1...\mathrm{d}^3\boldsymbol{r}_{k_1} \times \mathrm{d}^3\boldsymbol{r}'_1...\mathrm{d}^3\boldsymbol{r}'_{k_2}\xi^{(k_1,k_2)}\left(\boldsymbol{r}_1, ..., \boldsymbol{r}_{k_1}; \boldsymbol{r}'_1, ..., \boldsymbol{r}'_{k_2}\right) \times J_1(\boldsymbol{r}_1)...J_1\left(\boldsymbol{r}_{k_1}\right)J_2\left(\boldsymbol{r}'_1\right)...J_2\left(\boldsymbol{r}'_{k_2}\right)\right]. \tag{A3}$$

For a set of tracers of these underlying fields generated by a local poisson process, the number of tracers of each type contained in volume V around a point $\boldsymbol{r}$ depends on the integral of the field over the same volume

$$\mathcal{M}_V^{(i)}(\boldsymbol{r}) = \int \mathrm{d}^3\boldsymbol{r}'\rho_i(\boldsymbol{r}')W(\boldsymbol{r}, \boldsymbol{r}'), \tag{A4}$$

where W represents the window function for smoothing the fields.

The probability of finding k_i tracers of type i, in a sphere of volume V, is given by

$$P\left(k_i|\mathcal{M}_V^{(i)}\right) = \frac{\left(\mathcal{M}_V^{(i)}/m_i\right)^{k_i}}{k_i!}\exp\left[-\frac{\mathcal{M}_V^{(i)}}{m_i}\right], \tag{A5}$$

where m_i is the 'mass' associated with tracer of type i. The *joint* probability of finding k_1 tracers of field $\rho_1(\boldsymbol{r})$ and k_2 tracers of field

$\rho_2(\boldsymbol{r})$ in a volume V can be written in terms of the integral $\mathcal{M}_V^{(1)}$ and $\mathcal{M}_V^{(2)}$ as

$$P\left(k_1, k_2|\mathcal{M}_V^{(1)}, \mathcal{M}_V^{(2)}\right) = \frac{\left(\mathcal{M}_V^{(1)}/m_1\right)^{k_1}}{k_1!}\exp\left[-\frac{\mathcal{M}_V^{(1)}}{m_1}\right] \times \frac{\left(\mathcal{M}_V^{(2)}/m_2\right)^{k_2}}{k_2!}\exp\left[-\frac{\mathcal{M}_V^{(2)}}{m_2}\right]. \quad \text{(A6)}$$

The probability of k_1 and k_2 at a fixed volume V, or equivalently, radius R, can be computed by averaging over all $\mathcal{M}_V^{(i)}(\boldsymbol{r})$:

$$P\left(k_1, k_2|V\right) = \left\langle \frac{\left(\mathcal{M}_V^{(1)}/m_1\right)^{k_1}}{k_1!}\exp\left[-\frac{\mathcal{M}_V^{(1)}}{m_1}\right] \times \frac{\left(\mathcal{M}_V^{(2)}/m_2\right)^{k_2}}{k_2!}\exp\left[-\frac{\mathcal{M}_V^{(2)}}{m_2}\right]\right\rangle. \quad \text{(A7)}$$

Note that if the two fields ρ_1 and ρ_2 are correlated, then fluctuations in $\mathcal{M}_V^{(i)}$ are correlated, and the above average does not factor into the individual averages over $\mathcal{M}_V^{(1)}$ and $\mathcal{M}_V^{(2)}$.

The generating function for the discrete counts can be written as

$$P\left(z_1, z_2|V\right) = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty} P\left(k_1, k_2|V\right) z_1^{k_1} z_2^{k_2} = \left\langle \exp\left[\frac{\mathcal{M}_V^{(1)}}{m_1}\left(z_1-1\right) + \frac{\mathcal{M}_V^{(2)}}{m_2}\left(z_2-1\right)\right]\right\rangle. \quad \text{(A8)}$$

While equation (A8) is the generating function for the joint discrete counts, the average over $\mathcal{M}_V^{(i)}$ can be evaluated in terms of the underlying continuous fields:

$$P\left(z_1, z_2|V\right) = \int [D\rho_1][D\rho_2]\mathcal{P}\left(\rho_1, \rho_2\right) \times \exp\left[\frac{(z_1-1)}{m_1}\int_V \mathrm{d}^3\boldsymbol{r}'\rho_1(\boldsymbol{r}')W(\boldsymbol{r}, \boldsymbol{r}') + \frac{(z_2-1)}{m_2}\int_V \mathrm{d}^3\boldsymbol{r}'\rho_2(\boldsymbol{r}')W(\boldsymbol{r}, \boldsymbol{r}')\right]. \quad \text{(A9)}$$

With the following identification,

$$J_i(\boldsymbol{r}') = W(\boldsymbol{r}, \boldsymbol{r}')\frac{(z_i-1)}{im_i}, \quad \text{(A10)}$$

Equations (A9) and (A1) are equivalent. Then, noting that the RHS of both equations (A1) and (A3) define the same quantity, and focusing on the spherical top hat smoothing window for $W(\boldsymbol{r}, \boldsymbol{r}')$, equation (A9) can be written as

$$P\left(z_1, z_2|V\right) = \exp\left[\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\frac{\bar{n}_1^{k_1}(z_1-1)^{k_1}}{k_1!}\frac{\bar{n}_2^{k_2}(z_2-1)^{k_2}}{k_2!} \times \int_V \mathrm{d}^3\boldsymbol{r}_1 ... \mathrm{d}^3\boldsymbol{r}_{k_1}\mathrm{d}^2\boldsymbol{r}'_1 ... \mathrm{d}^3\boldsymbol{r}'_{k_2}\xi^{(k_1,k_2)}\right], \quad \text{(A11)}$$

where $\bar{n}_i = \bar{\rho}_i/m_i$ represents the mean number density of each sample of tracers. Equation (A11) makes explicit the fact that the generating function of the joint counts is sensitive to all possible correlations functions that can be constructed from the two underlying fields. In the absence of statistical correlations, the generating function for the discrete counts factorizes into two independent generating functions, one for each set of tracers.

APPENDIX B: JOINT CUMULATIVE COUNTS FOR CORRELATED GAUSSIAN FIELDS

In this section, we present the analytic expressions for some of the joint kNN-CDFs for a Gaussian random field. In particular, we look at the cases $k_1 = k_2 = 1$ and $k_1 = k_2 = 2$, *i.e.* the first and second joint NN distributions. For $k_1 = k_2 = 1$,

$$\mathrm{CDF}_{1,1}(r) = \mathcal{P}(>0, >0|V) = C(z_1, z_2|V)\Big|_{z_1,z_2=0}, \quad \text{(B1)}$$

where $C(z_1, z_2|V)$ is given by equation (3). For the Gaussian case, we use equations (10) and (12), and therefore

$$\mathcal{P}(>0, >0|V) = 1 - \exp\left[-\bar{n}_1 V + \frac{1}{2}\bar{n}_1^2\bar{\xi}_V^{(2,0)}\right] - \exp\left[-\bar{n}_2 V + \frac{1}{2}\bar{n}_2^2\bar{\xi}_V^{(0,2)}\right] + \exp\left[-\bar{n}_1 V - \bar{n}_2 V + \frac{1}{2}\bar{n}_1^2\bar{\xi}_V^{(2,0)} + \frac{1}{2}\bar{n}_2^2\bar{\xi}_V^{(0,2)} + \bar{n}_1\bar{n}_2\xi_V^{(1,1)}\right]. \quad \text{(B2)}$$

For the joint second NN distribution, we use the fact that

$$\mathrm{CDF}_{2,2}(r) = \mathcal{P}(>1, >1|V) = \frac{\mathrm{d}^2C(z_1, z_2|V)}{\mathrm{d}z_1\mathrm{d}z_2}\Big|_{z_1,z_2=0}. \quad \text{(B3)}$$

After performing the derivatives with respect to z_1 and z_2, the joint CDF can be expressed as

$$\mathcal{P}(>1, >1|V) = 1 - \mathcal{P}_1(0|V) - \mathcal{P}_2(0|V) - \mathcal{P}_1(1|V) - \mathcal{P}_2(1|V) + \mathcal{P}(0,0|V) + \mathcal{P}(1,0|V) + \mathcal{P}(0,1|V) + \mathcal{P}(1,1|V), \quad \text{(B4)}$$

with

$$\mathcal{P}_1(0|V) = \exp\left[-\bar{n}_1 V + \frac{1}{2}\bar{n}_1^2\bar{\xi}_V^{(2,0)}\right], \quad \text{(B5)}$$

$$\mathcal{P}_2(0|V) = \exp\left[-\bar{n}_2 V + \frac{1}{2}\bar{n}_2^2\bar{\xi}_V^{(0,2)}\right], \quad \text{(B6)}$$

$$\mathcal{P}_1(1|V) = \left(\bar{n}_1 V + \bar{n}_1^2\bar{\xi}_V^{(2,0)}\right)\exp\left[-\bar{n}_1 V + \frac{1}{2}\bar{n}_1^2\bar{\xi}_V^{(2,0)}\right], \quad \text{(B7)}$$

$$\mathcal{P}_2(1|V) = \left(\bar{n}_2 V + \bar{n}_2^2\bar{\xi}_V^{(0,2)}\right)\exp\left[-\bar{n}_2 V + \frac{1}{2}\bar{n}_2^2\bar{\xi}_V^{(0,2)}\right], \quad \text{(B8)}$$

$$\mathcal{P}(0,0|V) = \exp\left[-\bar{n}_1 V - \bar{n}_2 V + \frac{1}{2}\bar{n}_1^2\bar{\xi}_V^{(2,0)} + \frac{1}{2}\bar{n}_2^2\bar{\xi}_V^{(0,2)} + \bar{n}_1\bar{n}_2\bar{\xi}_V^{(1,1)}\right], \quad \text{(B9)}$$

$$\mathcal{P}(1,0|V) = \mathcal{P}(0,0|V) \times \left(\bar{n}_1 V - \bar{n}_1^2\bar{\xi}_V^{(2,0)} - \bar{n}_1\bar{n}_2\bar{\xi}_V^{(1,1)}\right), \quad \text{(B10)}$$

$$\mathcal{P}(0,1|V) = \mathcal{P}(0,0|V) \times \left(\bar{n}_2 V - \bar{n}_2^2\bar{\xi}_V^{(0,2)} - \bar{n}_1\bar{n}_2\bar{\xi}_V^{(1,1)}\right), \quad \text{(B11)}$$

$$\mathcal{P}(1,1|V) = \mathcal{P}(0,0|V) \times \Bigg(\bar{n}_1 \bar{n}_2 \bar{\xi}^{(1,1)} + \left(\bar{n}_1 V - \bar{n}_1^2 \bar{\xi}_V^{(2,0)} - \bar{n}_1 \bar{n}_2 \bar{\xi}_V^{(1,1)} \right) \times \left(\bar{n}_2 V - \bar{n}_2^2 \bar{\xi}_V^{(0,2)} - \bar{n}_1 \bar{n}_2 \bar{\xi}_V^{(1,1)} \right) \Bigg). \quad \text{(B12)}$$

Since $\bar{\xi}_V^{(2,0)} = V^2 \sigma_1^2(r)$, and $\bar{\xi}^{(0,2)} = V^2 \sigma_2^2(r)$, where $\sigma^2(r)$ is the variance of fluctuations on scale r, these expressions above can be directly evaluated if the variance for each field, along with the degree of cross-correlation is known. For the specific case when both sets of tracers under consideration trace the same matter field, $\bar{\xi}_V^{(1,1)} = V^2 \sigma^2(r)$, and so all the expressions can be evaluated with the help of, for example, the COLOSSUS software, which returns the linear theory value for the variance as a function of scale and redshift. In the case where there are no cross-correlations, we set $\bar{\xi}_V^{(1,1)} = 0$ for all V, and evaluate the above expressions accordingly.

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