MNRAS 000, 1–21 (2020) Preprint 28 July 2020 Compiled using MNRAS LATEX style file v3.0

Nearest Neighbor distributions: new statistical measures for cosmological clustering

Arka Banerjee⋆ and Tom Abel†
Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA
Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA
SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The use of summary statistics beyond the two-point correlation function to analyze the non-Gaussian clustering on small scales, and thereby, increasing the sensitivity to the underlying cosmological parameters, is an active field of research in cosmology. In this paper, we explore a set of new summary statistics — the k-Nearest Neighbor 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 is sensitive to all connected N-point correlations in the data. The kNN-CDF can be used to measure counts in cell, void probability distributions and higher N–point correlation functions, all using the same formalism exploiting fast searches with spatial tree data structures. We demonstrate how it can be computed efficiently from various data sets - both discrete points, and the generalization for continuous fields. We use data from a large suite of N-body simulations to explore the sensitivity of this new statistic to various cosmological parameters, compared to the two-point correlation function, while using the same range of scales. We demonstrate that the use of kNN-CDF improves the constraints on the cosmological parameters by more than a factor of 2 when applied to the clustering of dark matter in the range of scales between $10h^{-1}$Mpc and $40h^{-1}$Mpc. We also show that relative improvement is even greater when applied on the same scales to the clustering of halos in the simulations at a fixed number density, both in real space, as well as in redshift space. Since the kNN-CDF are sensitive to all higher order connected correlation functions in the data, the gains over traditional two-point analyses are expected to grow as progressively smaller scales are included in the analysis of cosmological data.

Key words: cosmology – cosmological parameter constraints

1 INTRODUCTION

Over the past three decades, a large part of the progress in cosmology — especially in the pursuit of increasingly precise and accurate constraints on the standard cosmological parameemeters — has been predicated on the use of two-point statistics in the analysis of cosmological datasets, either in real space, or in Fourier space. This includes analysis of the Cosmic Microwave Background (Planck Collaboration et al. 2018; Hinshaw et al. 2013) as well as Large Scale Structure analyses at low redshifts (Alam et al. 2017; Ivanov et al. 2020; d'Amico et al. 2020; Hildebrandt et al. 2017; Abbott et al. 2018). The latter include the analysis of both galaxy clustering — discrete data points — and weak lensing measurements — in the form of continuous maps. These studies employ a wide range of theoretical approaches in their modeling, including linear perturbation theory, higher order perturbation theory, and N-body simulations, but the two-point correlation function, or the power spectrum is usually the summary statistic of choice to compare theoretical predictions and data. Given its widespread use, a number of tools have been developed over the years for fast and efficient calculation of two-point statistics from cosmological data, or from simulations.

The two-point statistics provides a complete statistical description of a Gaussian random field. Perturbations in the early Universe are believed to closely follow the statistics of a Gaussian random field (*e.g.* Peacock 1998), though attempts have also been made to quantify any departures from Gaussianity in the early Universe (Planck Collaboration et al. 2019). As long as the evolution of the early Universe perturbations remain well described by linear perturbation theory, the Gaussian nature of the field is not affected. High redshift CMB analyses ($z \sim 1100$) and the analyses of clustering on extremely large scales at low redshift fall in this regime. Therefore, current analyses, utilizing two-point statistics, are able to extract maximal information about cosmological parameters from these redshifts and scales. On the other hand, at low redshifts, the field develops non-Gaussian features sourced by continued gravita-

⋆ E-mail: arkab@stanford.edu
† E-mail: tabel@stanford.edu

tional collapse — smaller the scale, more non-Gaussian the underlying field. Higher connected N-point correlation functions, which encode the more complicated nature of the field, start to be statistically important. The two-point statistics themselves can still be accurately modeled on relatively small scales, either through higher order perturbation theories (Carrasco et al. 2012; Taruya et al. 2012; Vlah et al. 2015; d'Amico et al. 2020; Ivanov et al. 2020), or using N-body simulations (Mead et al. 2015; Euclid Collaboration et al. 2019; Lawrence et al. 2017; Nishimichi et al. 2017). Nonetheless, an analysis using only the two-point statistic is insufficient to probe all the effects of different cosmological parameters on the evolution of the overall cosmological field. This motivates the need for considering other summary statistics in the analysis of clustering on smaller scales. Since the total information available in a survey scales with the number of independent modes (Tegmark 1997), and given that there are many more independent modes on small scales, it is important to develop these new statistical methods, that better extract information from small scales, to make optimal use of data from ongoing and future cosmological surveys.

Various approaches toward harnessing information beyond that contained in the two-point statistics have been explored in the literature. One method is consider the higher N-point functions in the analysis — the three-point function and its Fourier transform, the bispectrum (*e.g.* Scoccimarro et al. 1998; Takada & Jain 2004; Sefusatti et al. 2006), or the four-point function and its Fourier transform, the trispectrum (*e.g.* Verde & Heavens 2001). While these additional statistics can be highly informative about small scale clustering, and promise to yield significantly tighter constraints on some cosmological parameters (Hahn et al. 2020; Coulton et al. 2019), one drawback is that they are generally entail much higher computational costs to measure either from simulations, or from data. The computational complexity rises with N - the order of the highest connected correlation function considered, while keeping the data size fixed. Further, the noise properties for these higher order estimators often make it difficult to obtain good signal-to-noise ratio (SNR) over a wide range of scales. In spite of these issues, analyses including higher order correlations, especially the bispectrum, have successfully been applied to certain cosmological datasets (Gil-Marín et al. 2015; Gualdi et al. 2019; Slepian et al. 2017).

In this context, there also exist approaches which attempt to undo the nonlinear effects of gravitational clustering on various cosmological fields. Once this procedure is applied, the analysis proceeds with the measurement of the two-point function of the linearized field. This approach has already been applied specifically to the reconstruction of the Baryon Acoustic Oscillation (BAO) signal (Eisenstein et al. 2007; Padmanabhan et al. 2009; Padmanabhan et al. 2012). Schmittfull et al. (2015) showed that the reconstruction process can be interpreted as the transfer of information from the higher connected N-point functions in the observed field into the reconstructed two-point function. More recently, methods have been proposed for the full reconstruction of the initial linear modes from nonlinear observables at low redshift (Baldauf et al. 2010; Seljak et al. 2017; Horowitz et al. 2019; Modi et al. 2018).

There is also a large body of literature investigating the one point Probability Distribution Function (PDF) of matter density in the Universe, and the related counts-in-cell (CIC) statistics for discrete tracers like dark matter halos or galaxies (Coles & Jones 1991; Colombi 1994; Kofman et al. 1994; Gaztañaga et al. 2000; Lam & Sheth 2008; Bernardeau et al. 2014; Uhlemann et al. 2016; Klypin et al. 2018). While the PDF or CIC statistics are close to Gaussian when evaluated on large scales, and therefore contain the same information as the two-point function, on small scales the PDF captures information about all higher moments of the distribution, and therefore can be used to place stronger constraints on various cosmological parameters (Uhlemann et al. 2019). Variations of this type of analysis have also been applied already to different cosmological datasets (Petri et al. 2015; Gruen et al. 2015, 2018; Friedrich et al. 2018; Repp & Szapudi 2020). While this approach is extremely attractive in terms of its sensitivity to all higher order correlations, calculating the PDF from data involves multiple steps of smoothing and averaging, and the calculations have to be done separately for every radius bin used in the analysis.

Other statistical measures that have been employed to extract non-Gaussian information from cosmological fields, especially in the context of weak lensing, include peak counts (Peel et al. 2017; Fluri et al. 2018) and Minkowski functionals (Matsubara 2010; Munshi et al. 2012; Petri et al. 2013). Yet another set of studies have attempted to use special properties of nonlinear regions, such as halos and voids, in specific cosmologies, to enhance the constraints on some of the cosmological parameters. These include searches for scale dependent bias on large scales in the context of massive neutrinos (Villaescusa-Navarro et al. 2014; LoVerde 2016; Banerjee & Dalal 2016; Chiang et al. 2019; Banerjee et al. 2019) and primordial non-Gaussianity (Dalal et al. 2008; Desjacques et al. 2009; Castorina et al. 2018), and the use of marks, or density dependent weights, for correlation functions in the context of modified gravity (White 2016; Hernández-Aguayo et al. 2018; Armijo et al. 2018) and massive neutrinos (Massara et al. 2020). While these methods attempt to use additional information from the nonlinear density field, the statistic considered is usually the two-point function.

Finally, there are studies exploring the clustering of halos or galaxies in terms of the Void Probability Function (VPF), which was also shown to be the generating function for the full distribution of the clustering, and sensitive to all connected N-point functions White (1979); Fry (1986); Fry & Colombi (2013). A related approach is explored in Paranjape & Alam (2020). VPF and related measurements have already been applied to data (Sharp 1981; Way et al. 2011; Walsh & Tinker 2019). Over the years, the concept of the generating function, beyond just probabilities of finding completely empty volumes, as captured in the VPF, was further developed in the context of cosmological clustering (Balian & Schaeffer 1989; Szapudi & Szalay 1993; Bernardeau 1994), and provides an overarching theoretical framework to connect the parallel approaches of using higher N-point functions, and the use of one point PDF analysis, toward the extraction of non-Gaussian information on small scales.

In this paper, we introduce the k-nearest neighbor Cumulative Distribution Functions (kNN-CDF), *i.e.*, the empirical cumulative distribution function of distances from a set of volume-filling Poisson distributed random points to the k–nearest data points. The kNN-CDF are a set of new summary statistics that can be applied to the clustering analysis of cosmological datasets - both discrete tracers, and continuous fields, where the latter can be sampled by a set of tracers. We set out the connections between these new statistics and the generating function formalism. Through the latter, we describe the relationship between the kNN-CDF statistics and the statistics of higher N-point functions, as well as the density PDF over a range of scales. Importantly, we demonstrate how kNN-CDF statistics can be computed efficiently on a given dataset - and how a single measurement step is sufficient to gain information about all N-point correlation functions present in the data over a relatively broad range of scales. We apply these statistics in the context of familiar distributions, and finally, quantify the improvements in

cosmological parameter constraints, compared to two-point function analyses over the same range of scales.

The layout of the paper is as follows: in Sec. 2, we introduce the mathematical framework relevant for the kNN-CDF statistics, and outline how they can be computed for a given dataset. In Sec. 3, we apply the kNN-CDF statistics to various underlying fields to illustrate its novel features. In Sec. 4, the constraints on cosmological parameters using the kNN-CDF statistics is explored. Finally, in Sec. 5, we summarize our main findings, and discuss possible directions in which this study can be extended.

2 INTRODUCTION TO NEAREST NEIGHBOR CUMULATIVE DISTRIBUTION FUNCTIONS

In this section, we introduce the concept of the Nearest Neighbor Cumulative Distribution Function, and explore its connections to other statistical measures for clustering used in the literature. We also outline the method by which NN-CDF can be computed quickly for a given dataset.

2.1 Formalism

We consider a set of tracers of a underlying continuous field, with mean number density $\bar{n}$ and connected N-point correlation functions denoted by $\xi^{(N)}$. $\xi^{(0)} = 0$ by definition, and $\xi^{(1)} = 1$ to correctly normalize the distribution. The generating function, $P(z|V)$, of the distribution of the counts of data points enclosed in volume V can be written as (White 1979; Balian & Schaeffer 1989; Szapudi & Szalay 1993):

$$\begin{aligned} P(z|V) &= \sum_{k=0}^{\infty} P_{k|V} z^k \\ &= \exp\left[\sum_{k=1}^{\infty} \frac{\bar{n}^k (z-1)^k}{k!} \times \right. \\ &\qquad \left. \int_V \dots \int_V d^3\mathbf{r}_1 \dots d^3\mathbf{r}_k \xi^{(k)}(\mathbf{r}_1, \dots, \mathbf{r}_k) \right] . \end{aligned} \tag{1}$$

The derivation of Eq. 1 starting from the statistics of an underlying continuous field is sketched out in more detail in Appendix A. The shape of the volume V can, in general, be arbitrary. In this paper, we will only consider the volumes to be associated with spheres of radius r. As shown in Appendix A, this is a natural choice for describing statistics of a top-hat smoothed field. In terms of notation, we will switch between r and V throughout the paper, under the implicit assumption $V = 4/3\pi r^3$.

The probability of finding a count of $k \in \{0, 1, 2, \dots\}$ data points in a volume V can be computed from the generating function by computing various derivatives,

$$P_{k|V} = \frac{1}{k!} \left[\left(\frac{d}{dz} \right)^k P(z|V) \right]_{z=0} . \tag{2}$$

The quantity $P(z|V)|_{z=0}$ or, its mathematical equivalent, $P_{0|V}$ is referred to as the Void Probability Function (VPF) (White 1979), and represents the probability of finding no data points within a volume V. Note that the expression for the VPF still contains all N-point correlation functions, and White (1979) showed, using a slightly different formalism from that used here, the VPF itself can be considered as the generating function of the full distribution. In the literature, $P_{k|V}$ corresponds directly to the CIC statistics for tracer particles. For scales much larger than the mean inter-particle separation of the tracers, where the mean number of data points per volume, $\langle k_V \rangle \gg 1$, $P_{k|V}$ corresponds to the density PDF of the underlying continuous field (*e.g.* Klypin et al. 2018; Uhlemann et al. 2019). Using a very similar formalism, it is possible to write down the generating function for various cumulants of the distribution, which are directly related to the N-point connected correlation functions. This is demonstrated in Appendix B.

Next, we consider the statistics of volumes which have more than k data points, $P_{>k|V}$, where, once again, $k \in \{0, 1, 2, \dots\}$. We will first write down the generating function for this statistic, $C(z|V)$ in terms of the $P(z|V)$. We follow the same definition as Eq. 1 to write

$$\begin{aligned} C(z|V) &= \sum_{k=0}^{\infty} P_{>k|V} z^k = \sum_{k=0}^{\infty} \sum_{m=k+1}^{\infty} P_{m|V} z^k \\ &= (P_{1|V} + P_{2|V} + \dots) + (P_{2|V} + P_{3|V} + \dots) z \\ &\quad + (P_{3|V} + P_{4|V} + \dots) z^2 + \dots \\ &= -P_{0|V} + (P_{0|V} + P_{1|V} + P_{2|V} + \dots) + \\ &\quad - (P_{0|V} + P_{1|V}) z + (P_{0|V} + P_{1|V} + P_{2|V} + \dots) z \\ &\quad - (P_{0|V} + P_{1|V} + P_{2|V}) z^2 + (P_{0|V} + P_{1|V} + \dots) z^2 \\ &\quad + \dots \\ &= -P(z|V)(1 + z + z^2 + \dots) + (1 + z + z^2 + \dots) \\ &= \frac{1 - P(z|V)}{1 - z} , \end{aligned} \tag{3}$$

where we have used the fact $\sum_{k=0}^{\infty} P_{k|V} = 1$, and $1/(1-z) = (1+z+z^2+\dots)$. Therefore, the generating function for the distribution of $P_{>k|V}$ is fully specified by the generating function of $P_{k|V}$. Note that, by definition

$$P_{k|V} = P_{>k-1|V} - P_{>k|V} \qquad \forall k \geq 1 . \tag{4}$$

From Eqs. 1, 3, and 4, it becomes clear there are three equivalent approaches to characterizing the clustering of a set of data points through the generating function: 1) by measuring all the connected N-point correlation functions $\xi^{(N)}$, *i.e.*, the second line in Eq. 1, 2) by measuring the distribution of the counts in cell, $P_{k|V}$, *i.e.* the first line in Eq. 1, and 3) by measuring the cumulative counts, $P_{>k|V}$, and connecting them to $P_{k|V}$ using Eq. 4. To fully characterize the generating function, each of these statistical measures has to be measured over the full range of scales of interest. In the next subsection, we present the method for efficiently calculating $P_{>k|V}$ concurrently over a large range in V for a set of data points. This is done by exploiting the connection between $P_{>k|V}$ and the k-Nearest Neighbor distributions of the data points from a set of volume-filling random points.

2.2 Efficient calculation of $P_{>k|V}$ in data using kNN-CDF

For a set of N_d data points distributed over a total volume $V_{\rm tot}$, we start by generating a volume-filling sample of random points. Typically, the total number of randoms, N_R is chosen to be larger than the number of data points, N_d, and using more randoms allows for better characterization of the tails of the distributions discussed below. Once the set of randoms have been generated, for each random, we compute the distance to, say, the nearest data point. This can be done extremely efficiently by constructing a k-d tree on the data (e.g. Wald & Havran 2006) in $N \log N$ operations. Using this tree structure allows to search for the k nearest neighbors in $\log N$

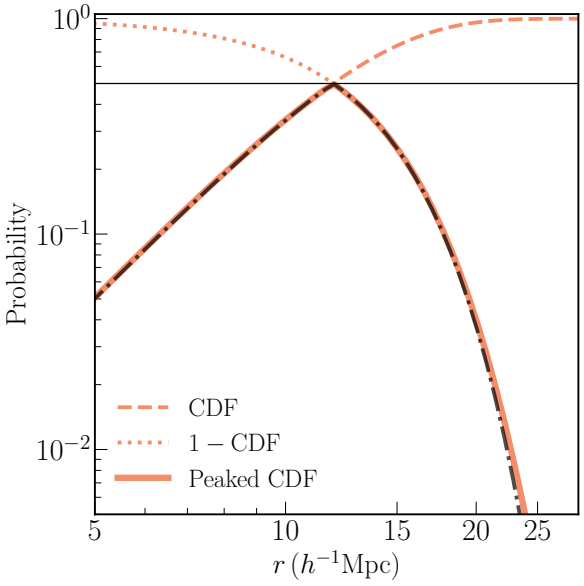

Figure 1. Peaked CDF (described in the text) of the nearest neighbor distribution (as defined in the text) for 10^5 random points distributed over a $(1h^{-1}\mathrm{Gpc})^3$ volume (solid curve). The dot-dashed curve is the analytic prediction for the distribution. The empirical CDF measured from the data is plotted using the dashed curve, while the Void Probability Function (VPF) is plotted using the dotted line.

time for each point. Most scientific software has efficient built–in functions or libraries such as e.g. SCIPY's CKDTREE implementation, or JULIA's NEARESTNEIGHBOR.JL[1] library. These typically will return an ordered list of the distances to the nearest k neighbors. Sorting the computed distances for each k, immediately gives the Empirical Cumulative Distribution Function (see *e.g.* Vaart (1998)) of the distribution of distances to the k-nearest data point from the set of volume-filling randoms. In the limit of large N_R, this converges to the true underlying Cumulative Distribution Function (CDF) of the distance to the k-nearest data point from any arbitrary point in the box. Notice that since the randoms are volume-filling, all regions are sampled equally, irrespective of whether the region is overdense or underdense in terms of the data points. This is especially relevant to applications in cosmology, where at late times, the volume is dominated by underdense regions, while the data points are usually concentrated in overdense regions. Measures such as the two-point correlation functions are typically dominated by the overdense regions, and have little information about the large underdense regions (e.g Oort 1983).

To understand the connection between the CDF of the distribution of distances to the nearest data point from a random point in the volume, and the probability of $P_{>k|V}$, consider all possible spheres of volume $V = 4/3\pi r^3$ (Kerscher et al. 1999). These spheres can be centered on any point in the total volume under consideration. The fraction of spheres with > 0 data points is exactly the fraction of sphere centers for which the distance to the nearest data point is $< r$. The nearest-neighbor CDF at some radius r is the precise measure of the fraction of points for which the nearest data point is at a distance $< r$. Therefore,

$$\mathrm{CDF}_{1\mathrm{NN}}(r) = P_{>0|V}\big|_{V=4/3\pi r^3}\,. \tag{5}$$

As mentioned previously, we will switch notations between radius r and the volume V throughout this paper, and we note that none of our results depend on the distinction in notation. Given the connection between nearest neighbor distances and finding a certain number of data points in a given volume, $1 - \mathrm{CDF}_{1\mathrm{NN}}(r)$ then represents the probability of finding a completely empty volume $V = 4/3\pi r^3$. That is, a randomly placed sphere of Volume V is empty with a probability of $1 - \mathrm{CDF}_{1\mathrm{NN}}(V)$, which is known as the Void Probability Function (VPF, White (1979)). Interestingly, this latter interpretation is what is customarily used (e.g. CORRFUNC code[2] Sinha & Garrison (2019); Sinha & Garrison (2020)) to measure the VPF using large numbers of randomly placed spheres. This approach is much slower, however, as compared to the kNN–CDF. It also only provides measurements at typically a small number of chosen volumes. So even if one were only interested in computing the VPF, using the nearest neighbor approach discussed here would be advisable since the measurements are desired at multiple volumes. Note that the empirical CDF directly measured from the data contains as many points as random points we chose to cover the full volume. Given the monotonic nature of the CDF, using linear or higher order interpolation between measured points allows one to carry out operations on these CDFs as if they were continuous functions.

We illustrate the shapes of these functions for a Poisson distribution of points in Fig. 1. While the Empirical CDF is the quantity which is directly computed from the simulations, in our plots we will usually display the Peaked CDF (PCDF) which is defined as

$$\mathrm{PCDF(r)} = \begin{cases} \mathrm{CDF}(r) & \mathrm{CDF}(r) \le 0.5 \\ 1 - \mathrm{CDF}(r) & \mathrm{CDF}(r) > 0.5\,. \end{cases} \tag{6}$$

The use of the PCDF allows for better visual representation of both tails of the CDF. This point is illustrated in Fig. 1 for a set of 10^5 points distributed according to a Poisson distribution over a $(1h^{-1}\mathrm{Gpc})^3$ volume. The dashed line represents the Empirical Cumulative Distribution Function measured from the set of particles. The solid line represents the Peaked CDF computed from the same data. The right hand tail of the distribution is difficult to differentiate using the Empirical CDF, since, by its very nature, it asymptotes to 1 smoothly. The Peaked CDF, on the other hand, illustrates clearly the behavior in the tails, especially when comparing to the analytic expectation, plotted using the dot-dashed line. We also plot the VPF $(= 1 - \mathrm{CDF}_{1\mathrm{NN}})$ using the dotted line for reference.

Importantly, apart from the distance to the 1st nearest neighbor data point, the same tree is used to find the distances to the k-th nearest data point for each random point in the volume. Once again, these distances can be sorted to produce the Empirical CDF of the k-th neighbor distances, and in the limit of large N_R, the true underlying CDF of the k-th neighbor distances. We can generalize the arguments presented above to connect the probability of finding spheres of volume $V = 4/3\pi r^3$ enclosing $> k - 1$ data points to the CDF of distances to the k-th nearest data point within radius r:

$$\mathrm{CDF}_{k\mathrm{NN}}(r) = P_{>k-1|V}\big|_{V=4/3\pi r^3}\,. \tag{7}$$

Eq. 4 can be recast as

$$P_{k|V} = \mathrm{CDF}_{k\mathrm{NN}}(r) - \mathrm{CDF}_{(k+1)\mathrm{NN}}(r) \qquad \forall k \ge 1\,, \tag{8}$$

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

[2] https://github.com/manodeep/Corrfunc

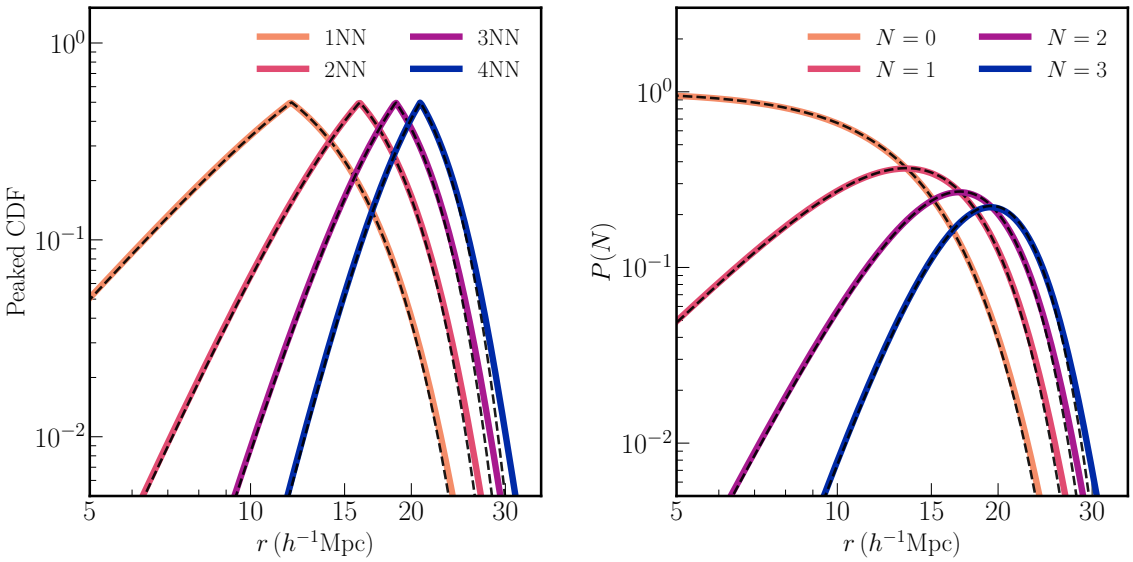

Figure 2. *Left*: Peaked CDF as a function of scale for 1st (1NN), 2nd (2NN), 3rd (3NN), and 4th (4NN) nearest neighbor distributions (solid lines) for a set of 10^5 Poisson distributed data points distributed over a $(1h^{-1}\mathrm{Gpc})^3$ volume from a set of volume filling randoms (see text for details). The dashed lines represent the analytic expectations for the distribution. *Right*: Probability of finding N points in a sphere with radius r given 10^5 Poisson distributed data points over a $(1h^{-1}\mathrm{Gpc})^3$ volume. Solid lined represent the probabilities computed using the CDFs from the left panel, while the dashed lines represent the analytic expectation.

$$P_{0|V} = 1 - \mathrm{CDF}_{1\mathrm{NN}}(r)\,. \tag{9}$$

Therefore, computing the k nearest neighbor distributions using the method outlined here is completely equivalent to measuring $P_{>k|V}$. Additionally, this procedure allows us to compute the probabilities over a range of scales, and for multiple values of k, in a single operation. To provide a sense of the time spent on a typical calculation, computing the CDFs for up to 8 nearest neighbors for 10^6 randoms and 10^5 data points distributed in a $(1h^{-1}\mathrm{Gpc})^3$ volume takes ~ 20 seconds on a single core. If the tree search is suitably parallelized, scaling up the number of cores can further reduce the runtime.

We note here that while we have used a set of random points to sample the entire volume, the method outlined above will also work for regularly spaced points, *e.g.* when using points placed on a finely-spaced grid, that is, with grid separations much smaller than the mean inter-particle separation of the data. Once the set of volume filling points are generated, whether using a random procedure, or on the regular grid, the calculation of the distances to the nearest data point, and the computation of the Empirical Cumulative Distribution Function proceeds exactly the same way.

3 APPLICATIONS TO VARIOUS DISTRIBUTIONS

In this section, we apply the NN-CDF formalism to tracers following different distributions, and point out various relevant features. We start with the simplest example of a Poisson sampling of a uniform field in Sec. 3.1. We then explore the Gaussian distribution in this framework in Sec. 3.2. Finally, in Sec. 3.3, we apply it to data from cosmological simulations, both simulation particles, and dark matter halos.

3.1 kNN-CDF for Poisson sampling of uniform field

For a sample of Poisson tracers of a uniform field, $\xi^{(1)} = 1$, and all higher order correlation functions are 0. In this case, Eq. 1 simplifies to

$$P(z|V) = \exp\left[\bar{n}(z-1)V\right]. \tag{10}$$

As can be anticipated for a pure Poisson process on a uniform field, the expression for the distribution of counts, $P_{k|V}$, in Eq. 2 becomes

$$P_{k|V} = \frac{1}{k!}\left[\left(\frac{d}{dz}\right)^k P(z|V)\right]_{z=0} = \frac{(\bar{n}V)^k}{k!}\exp(-\bar{n}V)\,. \tag{11}$$

The distribution of $P_{>k|V}$ can similarly be worked out by considering the derivatives of $C(z|V)$ from Eq. 3:

$$\begin{aligned}
P_{>k|V} &= \frac{1}{k!}\left[\left(\frac{d}{dz}\right)^k C(z|V)\right]_{z=0} \\
&= \frac{1}{k!}\left[\left(\frac{d}{dz}\right)^k \left(\frac{1-\exp\left[\bar{n}(z-1)V\right]}{1-z}\right)\right]_{z=0} \\
&= \frac{1}{k!}\left[\sum_{m=0}^{k}\frac{k!}{m!(k-m)!}\left(\frac{d}{dz}\right)^m \left(1-\exp\left[\bar{n}(z-1)V\right]\right)\right. \\
&\qquad \left.\left(\frac{d}{dz}\right)^{k-m}\frac{1}{1-z}\right]_{z=0} \\
&= 1 - \sum_{m=0}^{k}\frac{(\bar{n}V)^m}{m!}\exp(-\bar{n}V),
\end{aligned} \tag{12}$$

where we use the fact that $(d/dz)^m(1/(1-z)) = m!/(1-z)^{m+1}$. The form of $P_{>k|V}$ derived in Eq. 12 can also be anticipated by simply

noting that $P_{>k|V} = 1 - P_{<=k|V}$, and using Eq. 11. The form of Eq. 12 is known in the literature as the Cumulative Distribution Function of the Erlang distribution (M Evans & Peacock 2000). Since we consider the volumes V to be associated with spheres of radius r, all the equations above can also be trivially written in terms of r.

Eq. 12 can also be derived in a different way in the language of distances to nearest neighbors — by considering the distributions of successive nearest neighbors. Let us consider the CDF of the nearest neighbor distribution:

$$\mathrm{CDF}_{1\mathrm{NN}}(V) = P_{>0|V} = 1 - \exp(-\bar{n}V)\,. \tag{13}$$

Therefore, at fixed $\bar{n}$, the PDF of the distribution of distances to, or equivalently the distribution of volumes enclosed within, the nearest data point from a random point is given by

$$\mathrm{PDF}_{1\mathrm{NN}}(V) = \frac{d\big(\mathrm{CDF}_{1\mathrm{NN}}(V)\big)}{dV} = \bar{n}\exp(-\bar{n}V)\,. \tag{14}$$

Since there are no higher order correlations in the underlying continuous uniform field, the PDF of the distribution of volumes enclosed within the *second* nearest neighbor is a convolution of $\mathrm{PDF}_{1\mathrm{NN}}(V)$ with itself:

$$\begin{aligned}\mathrm{PDF}_{2\mathrm{NN}}(V) &= \int_0^V \mathrm{PDF}_{1\mathrm{NN}}(V')\mathrm{PDF}_{1\mathrm{NN}}(V-V')dV' \\ &= \bar{n}^2 V\exp(-\bar{n}V),\end{aligned} \tag{15}$$

and

$$\begin{aligned}\mathrm{CDF}_{2\mathrm{NN}}(V) &= \int_0^V \mathrm{PDF}_{2\mathrm{NN}}(V')dV' \\ &= 1 - \exp(-\bar{n}V) - (\bar{n}V)\exp(-\bar{n}V)\,.\end{aligned} \tag{16}$$

Comparing with the last line in Eq. 12, the expression above is indeed equivalent to $P_{>1|V}$ at fixed V. A similar result can be demonstrated for higher kNN CDFs of the Poisson distribution. For tracers of a Poisson distribution, therefore, computing the distribution of distances to the nearest neighbor data point from any arbitrary point in the volume under consideration is a complete description of the overall distribution of points - distributions of distances to all other neighbors can be generated from the former. Using the connection between nearest neighbor distributions and the probabilities of finding more than k data points in a volume V, the above result implies that, for the Poisson distribution, determining $P_{>0|V}$ or the VPF alone allows us to extract maximal information about distribution. This is consistent with the fact that the only variable for a Poisson distribution is the rate (= $\bar{n}V$ here), and that the full distribution can be generated once the rate is known, from say, the nearest neighbor distribution. Expressed in another way, there is no new information in any of the higher k-th neighbor distributions, once the nearest neighbor distribution is known.

We now compare the analytic predictions to the actual measurements from a set of 10^5 points distributed randomly over a $(1h^{-1}\mathrm{Gpc})^3$ volume. The results are shown in Fig. 2, where the first four nearest neighbor distributions are plotted in the left panel. The four lowest counts-in-cell distributions, computed from the nearest neighbor distributions on the left, are plotted in the right hand panel. In both panels, the solid lines represent the measurements, while the dashed line represent the analytic expectations. The measurements in the left panel agree extremely well with the analytic expectations for scales below, and comparable to, the mean inter-particle separation ($\sim 15h^{-1}\mathrm{Mpc}$). The right panel, where the solid curves have been computed using Eq. 8, clearly demonstrates that the nearest neighbor distributions and the counts-in-cells at a given radius are truly equivalent descriptions of the underlying data, and that one can easily be computed once the other is known.

On scales much larger than the mean inter-particle separation, the results begin to diverge for two reasons. The first is due to the limitations on the sampling of the total volume arising from the use a finite number of random particles. In the method outlined in Sec. 2.2, the number of randoms determines how well the Empirical CDF is sampled, and finite sampling can lead to discrepancies, especially in tails, where the true CDF is being approximated from a small number of measurements. For this plot, we use 8×10^6 random particles, and we have checked that using more randoms extends the agreement out to larger scales. The second effect is that of sample variance in the data itself. This arises because we evaluate the Empirical CDF on one specific realization of 10^5 points drawn from a Poisson distribution, and once again leads to departures from the analytic value in the tails. This effect can be reduced by appropriately averaging over many realizations of the data keeping the underlying distribution fixed. We note that amongst the distributions we have plotted, the nearest neighbor distribution most closely traces the analytic expectation in the tails, at fixed number of randoms, and is therefore expected to be the most robust measurement when we generalize to other distributions. In general, we will usually restrict our measurements and analysis to scales which are a few times the mean inter-particle separation of the tracers, and ensure that the measurements are not affected by the lack of sampling in the tail.

3.2 kNN-CDF for Gaussian Fields

For completely Gaussian continuous fields, the distribution is completely determined by the variance as a function of scale. Therefore the power spectrum, $P(k)$, or the two-point correlation function, $\xi(r)$, are complete descriptions of the underlying field, as well as for any set of tracers of the underlying field. As discussed in Sec. 1, these summary statistics have been employed extensively in the study of the CMB, as well as of Large Scale Structure on scales where non-linearities in the density field play a minor role. Therefore, a description of the clustering of a set of tracers of a Gaussian field in the language of NN-CDFs serves as a useful exercise to build up intuition about the connections between the two formalisms before examining a fully non-linear cosmological field.

For a Gaussian field, the generating function for $P_{k|V}$ can be written as follows:

$$\begin{aligned}P(z|V) &= \exp\Bigg[\bar{n}(z-1)V \quad + \\ &\qquad \frac{\bar{n}^2(z-1)^2}{2}\int_V\int_V d^3\mathbf{r}_1 d^3\mathbf{r}_2 \xi^{(2)}(\mathbf{r}_1,\mathbf{r}_2)\Bigg] \\ &= \exp\left[\bar{n}(z-1)V + \frac{1}{2}\bar{n}^2(z-1)^2V^2\sigma_V^2\right],\end{aligned} \tag{17}$$

where σ_V is the variance as a function of scale, defined in Eq. B7. Notice that the generating function for $P_{k>V}$, given by $C(z|V) = (1-P(z|V))/(1-z)$, therefore, contains only two possible unknowns - the mean number density $\bar{n}$, and the variance as a function of scale σ_V. The individual $P_{>k|V}$ distributions can be obtained by taking derivatives of $C(z|V)$. While there is no closed form expression of $P_{>k|V}$ for a general value of k, the individual terms are easy to

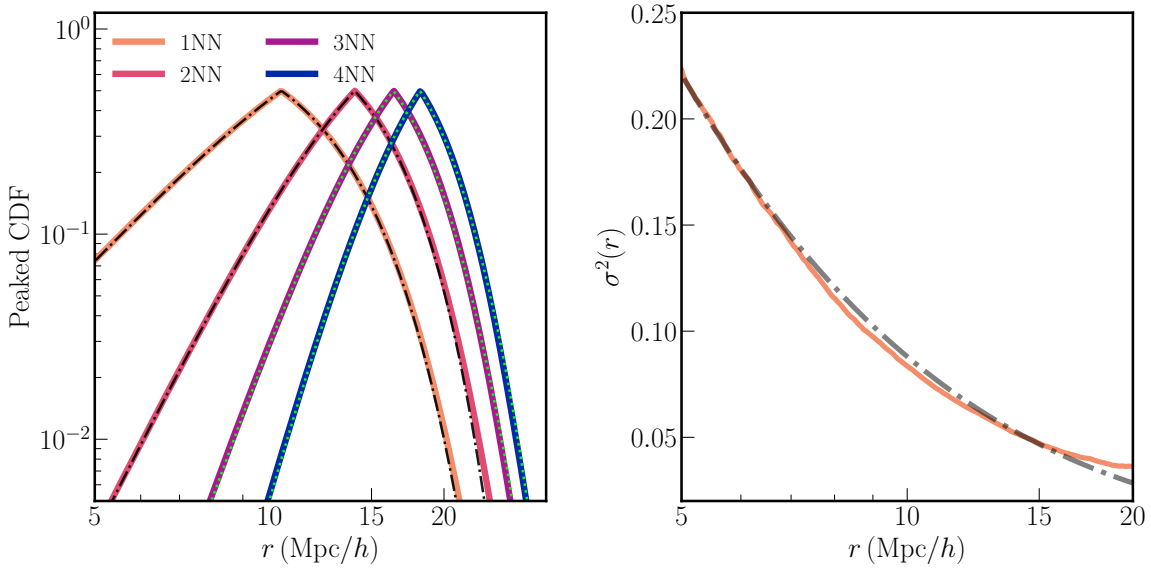

Figure 3. *Left*: Peaked CDF as a function of radius for 1st (1NN), 2nd (2NN), 3rd (3NN), and 4th (4NN) nearest neighbor distributions (solid lines) for a set of 10^5 points tracing an underlying (almost) Gaussian field distributed over a $(1h^{-1}\mathrm{Gpc})^3$ volume. The dash-dotted lines for the 1NN and 2NN represent the analytic expectations for distribution given the linear theory variance. The dotted lines represent the predictions for the 3NN and 4NN distributions from the actual measurements of the 1NN and 2NN distributions. *Right*: The solid line represents the variance of the Gaussian field as measured from the 1NN and 2NN CDFs. The dash-dotted line represents the theoretically computed variance from linear perturbation theory.

compute, especially for low values of k. For example,

$$P_{>0|V} = 1 - \exp\left[-\bar{n}V + \frac{1}{2}\bar{n}^2V^2\sigma_V^2 \right], \tag{18}$$

$$P_{>1|V} = 1 - P_{>0|V} - \left(\bar{n}V - \bar{n}^2V^2\sigma_V^2\right)\exp\left[-\bar{n}V + \frac{1}{2}\bar{n}^2V^2\sigma_V^2 \right], \tag{19}$$

and so on. Note that just by measuring the first two cumulative distributions, $P_{>0|V}$ and $P_{>1|V}$, one can constrain $\bar{n}$ and σ_V^2. Concretely,

$$\bar{n}V = -2\left(\log\left(1 - P_{>0|V}\right) + \frac{1}{2}\frac{P_{>0|V} - P_{>1|V}}{1 - P_{>0|V}} \right), \tag{20}$$

and

$$\sigma_V^2 = -2\left(\log\left(1 - P_{>0|V}\right) + \frac{P_{>0|V} - P_{>1|V}}{1 - P_{>0|V}} \right) \Big/ (\bar{n}V)^2 . \tag{21}$$

Using the relationship between $P_{>k|V}$ and $\mathrm{CDF}_{(k+1)\mathrm{NN}}(r)$ from Eq. 7, Eqs. 20 and 21 can also be expressed in terms of the 1st and 2nd nearest neighbor distributions. Once the relevant parameters have been uniquely defined in Eqs. 20 and 21, all higher k distributions can easily be derived in terms of the measured mean density and variance.

To compare measurements from a set of tracers of a Gaussian field with the theoretical predictions presented above, we consider a very coarse cosmological simulation of 128^3 particles in a $(1h^{-1}\mathrm{Gpc})^3$ box at the *Planck* best-fit cosmology run up to $z = 3$, when the density field is still roughly Gaussian. We then subsample 1.5×10^5 particles from the simulation particles and compute the four nearest neighbor distributions. We use the Colossus[3] code to generate the predictions for σ_V at this cosmology from linear perturbation theory. The results of the comparison are plotted in the left panel of Fig. 3. The solid lines represent the measurements of the nearest neighbor distributions from the data, while the dot-dashed line represent the predictions for the first and second nearest neighbor distributions using the theoretical σ_V^2. Once again, we find good agreement between the measurement and the predictions out to a few times the mean inter-particle separation. The dotted lines in the left panel represent the predictions for the 3rd and 4th nearest neighbor distributions, using $\bar{n}$ and σ_V^2 measured from the data using Eqs. 20 and 21. As anticipated from the arguments presented above, measurements of just the first and second nearest neighbor distributions allow us predict all other nearest neighbor distributions to a high degree of accuracy. In the right panel of Fig. 3, we plot the value of σ_V^2 that we recover from the measurements of the two nearest neighbor CDFs (solid line), and show that it agrees with the linear theory prediction for the continuous field from Colossus (dot-dashed line), once again on scales comparable to the mean inter-particle separation. It should be noted that the measurements allow us to correctly infer the variance of the underlying field even on scales smaller than the mean inter-particle separation ($\sim 12h^{-1}\mathrm{Mpc}$) — naively, these scales are expected to be dominated by the Poisson-like $\bar{n}V$ term in Eq. 17.

We conclude that, for tracers of a Gaussian density field, nearest neighbor distributions beyond the 2*nd* nearest neighbor distribution do not add any new statistical information about the underlying field. All higher neighbor distributions can be built up from convolutions of the two nearest neighbor distributions. In fact, these properties

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

of the NN-CDF for a formally Gaussian field can be turned into a test for the Gaussianity of a given field, or a set of tracers. If the mean density and the variance are computed from the data, then the field is completely Gaussian if and only if the nearest neighbor distributions are consistent with Eq. 18. Any departures from these expressions can be interpreted as evidence for non-Gaussianity in the field.

In general, therefore, for tracers of a field that is completely characterized by the first m connected N-point functions (or cumulants), measurements of only the lowest m nearest neighbor distributions are sufficient to capture the full statistical information of the underlying field. There is no independent information in the higher NN distributions. We note that there are certain distributions relevant to cosmological applications, most notably the log-normal PDF, which cannot be uniquely described by its cumulants. Further, at any given scale, we do not know *a priori* how many cumulants are needed to describe the distribution of tracers. However, even in these cases, measuring the first few NN distributions, captures a large fraction of total statistical information in the field. We will discuss these points further in Sections 3.3 and 4. For a discussion on the information in the VPF for a lognormal field, see Coles & Jones (1991).

Another aspect to note here is that our choice of downsampling the simulation particles to 1.5×10^5 for our measurements was arbitrary, and is not related to the number of particles with which we run the simulation. The choice was guided only the range of scales over which we wish to obtain robust measurements of the CDF. We can also recover the variance of the underlying continuous field for other choices of the mean number density, or equivalently, inter-particle separations. As shown previously, our measurements are most robust on scales comparable to the mean inter-particle separation, so a different choice of the mean number density allows us to accurately measure the variance on a different set of scales. However, the range of scales on which the linear theory variance can be reliably estimated from these measurements is limited on both large and small scales due to numerical and practical considerations. These limitations set the choice of scales displayed in the right panel of Fig. 3.

First consider the case when the inter-particle separation is much larger than considered here. In principle, this choice should allow us to measure the variance at large scales. However, the shape of the cosmological power spectrum is such that the variance decreases on large scales. If the variance is small on scales which can be well measured for a specific choice of the mean density of tracers, the distributions are dominated by $\bar{n}V$ term in the exponent on the RHS of Eq. 20, instead of the $\bar{n}^2V^2\sigma_V^2$ term. For small enough σ_V^2, this will be true even when considering scales above the mean inter-particle separation ($\bar{n}V > 1$). Such a scenario makes it numerically difficult to recover the variance of the continuous field using the techniques outlined above. However, it should be emphasized that this is a practical consideration, and not an inherent drawback of the overall method. If the tails of the distribution can be measured extremely accurately - *i.e.* with many more random points than typically used here, it is, in principle, possible to recover the full information about the variance, even at large scales.

At the other end, the problem arises from the fact that in linear theory, σ_V increases as we consider smaller scales, until $\sigma_V \sim 1$ on small enough scales. At this point, the field is no longer physical since the implied density PDF has tails which go below 0. In a simulation, of course, gravitational collapse leads to the generation of higher order correlation functions so that the tracer positions continue to represent a physical density field, as we will explore in more detail in Sec. 3.3. However, in this case, a measurement of the two nearest neighbor distributions will contain information not just about the mean and the variance, but all higher order moments that are present. This implies that the full clustering statistics can no longer be uniquely determined from just these measurements, and higher NN distributions need to be measured for a full statistical description of the field.

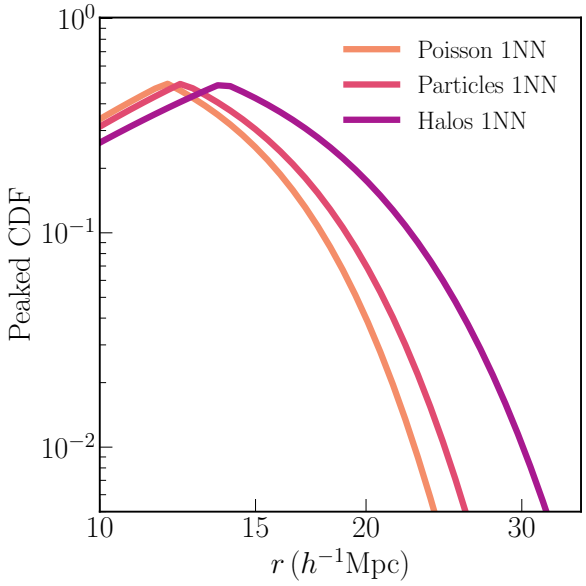

Figure 4. Comparison of the Peaked CDF for nearest neighbor (1NN) distributions of *a)* a Poisson distribution, *b)* particles from the $z = 0$ snapshot of an N-body simulation, and *c)* the most massive halos from the same simulation. In each case, 10^5 points were selected over a $(1h^{-1}\mathrm{Gpc})^3$ volume. For the particles, these were randomly selected from all the simulation particles, while for the halos, a cut was made on the 10^5 most massive halos in the box.

3.3 kNN-CDF for Large Scale Structure

We now move to analyzing realistic cosmological density fields using the nearest-neighbor formalism that we have set up. At low redshifts, the density field for matter is highly nonlinear, and the distribution of densities on small scales, especially, cannot be approximated by a Gaussian. The clustering of massive virialized dark matter halos, which host the visible galaxies, is usually even more non-Gaussian. In general, the distribution of Dark Matter at low redshifts form a cosmic web - with large empty regions (voids), and high density filaments and knots. In Fig. 4, we plot the nearest neighbor distributions for three different sets of points with the same total number of points, 10^5, spread over a $(1h^{-1}\mathrm{Gpc})^3$ volume. The corresponding mean inter-particle separation is $\sim 15h^{-1}\mathrm{Mpc}$. The range of scales represents those where the distributions are best measured for the choice of particle number and the number of randoms that are used to characterize the full volume. The first set of points are distributed randomly over the full volume, *i.e.* following a Poisson distribution. The second set of points are downsampled from the simulation particles in a cosmological simulation with 512^3 particles at $z = 0$. The third set of points are the halo centers of the 10^5 most massive halos in the simulation at $z = 0$. Even

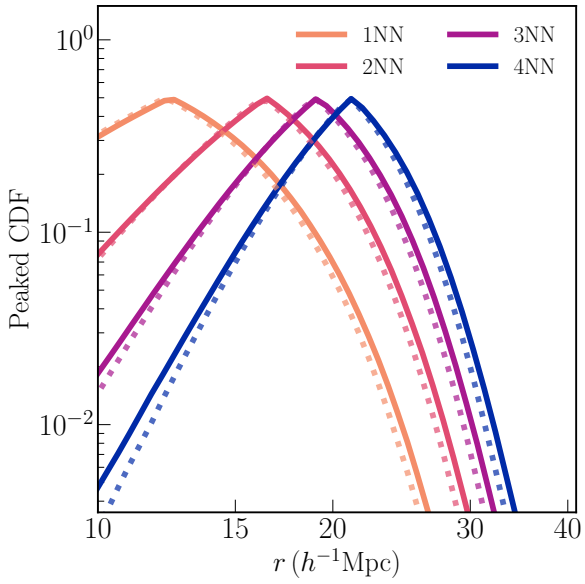

Figure 5. The Peaked CDFs for the first, second, third, and fourth nearest-neighbor distributions for 10^5 simulation particles in a $(1h^{-1}\mathrm{Gpc})^3$ volume. The solid lines represent these distributions at $z = 0$, while the dotted lines represent the distributions computed at $z = 0.5$.

though the mean number density is the same, it is clear that the nearest neighbor distributions look quite different. The NN-CDF of the simulation particles, and especially the halo positions, extends to larger scales. This happens due to the presence of large voids in these datasets - this implies that for a large faction of the volume filling set of randoms, the nearest data point is much further away than it would be for a Poisson distribution. We note that while the differences are the most pronounced on larger scales when plotted in terms of the Peaked CDF, the distributions are different even on scales smaller than the mean inter-particle separation.

It is also useful to plot the change in the nearest neighbor distributions with redshift. Fig. 5 shows the first four NN distributions for 10^5 simulation particles at z=0 (solid lines) and $z = 0.5$ (dotted line). Gravitational evolution of large scale structure drives overdense regions to become more overdense, while underdense regions become more underdense and expand in size. Both these effects can be seen in Fig. 5 - on large scales, the distributions extend out further at $z = 0$ compared to $z = 0.5$ as a result of the voids becoming larger. On small scales, the distributions, especially for the third and fourth nearest neighbors have larger values at a fixed scale for $z = 0$ compared to $z = 0.5$, as a result of the collapse of the overdense regions.

3.3.1 *Breaking the $b - \sigma_8$ degeneracy*

In the analysis of the clustering of halos using the two-point correlation function, there is a known degeneracy between the bias of the halos being considered and the amplitude of clustering of the underlying matter field, sometimes represented by the width of the density PDF at 8Mpc, σ_8. This is related to the fact that halos form at the peaks of the initial Gaussian random field (Bardeen et al. 1986). In other words, two halo populations can have the same two-point clustering signal, even when they are produced by very different underlying matter fields, just by appropriately choosing the bias of each sample. Since the bias is primarily dependent on the halo mass, this is equivalent to making choices about the mass cut for the samples. As a result, it is difficult to individually constrain the value of the bias b, and the amplitude of the clustering of the underlying field σ_8, given a two-point measurement of a sample of halos.

This degeneracy is illustrated in the left panel of Fig. 6. The dark solid line represents the ratio of the $\xi(r)$ at $z = 0.5$ to $\xi(r)$ at $z = 0$ for the 10^5 most massive halos for a single realization at a fixed cosmology. The halos are identified at the different redshifts separately, and so the actual objects that make the cut at different redshifts need not be the same. The dotted lines represent the ratio of $\xi(r)$ for 15 different realizations at $z = 0$ for the same cosmology, compared to the mean $\xi(r)$ over the 15 realizations. These curves serve as a measure of the sample variance for this measurement at $z = 0$. We can conclude that the two-point functions of the 10^5 most massive halos in the box at $z = 0$ and $z = 0.5$ cannot be distinguished over these range of scales to within sample variance, even though the amplitude of clustering of the underlying matter field is quite different at the two redshifts.

As we have shown in Sec. 2.1, the nearest neighbor distributions are sensitive not just to the two-point correlation functions, but all higher order correlations. This extra information can be used to break the degeneracy outlined above, as has already been demonstrated, *e.g.* in Pan & Szapudi (2005) and Sefusatti et al. (2006), by including the bispectrum or three point correlation function in the analysis. We illustrate this in the middle and right panels of Fig. 6. In the middle panel, the dark solid line represents the ratio of the nearest neighbor CDF of the 10^5 most massive halos at $z = 0.5$ to the nearest neighbor CDF of the 10^5 most massive halos at $z = 0$, for a single realization at a fixed cosmology. The lighter dotted lines represent the ratio of the nearest neighbor CDFs of the 10^5 most massive halos at $z = 0$ for 15 different realizations of the same cosmology to the mean distribution. Once again the dotted curves serve as a visual representation of the sample variance for the measurement. The right panel repeats the same calculation for the *second* nearest neighbor distribution. It easy to distinguish between the distributions at the two redshifts, either by considering the nearest neighbor distribution or the second nearest neighbor distribution, even though the clustering of the two samples cannot be distinguished by measurements of the two-point correlation function. This result can also be interpreted as proof that at $z = 0.5$ and $z = 0$, the clustering is quite non-Gaussian on the scales considered. If the halo field was completely Gaussian, Eq. 17 would imply that two fields - one at $z = 0.5$ and at $z = 0$ - with the same variance, as is the case here, would have exactly the same kNN statistics. The fact that the two fields have different kNN statistics, despite having the same variance, implies the non-Gaussianity of the clustering.

We note that the fact that the same number density cut - 10^5 in a $(1h^{-1}\mathrm{Gpc})^3$ volume - yields the same two-point correlation measurement is a coincidence. The results above would hold even if the number density cuts were different. However, it is also important to note that for the NN-CDF analysis, care should be taken to use the same number of data points. The distributions depend sensitively on the mean number density, and a change in the mean number density can be misinterpreted as a change in the clustering signal. It is easy to ensure the two datasets have the same number of points by randomly downsampling the larger dataset. We will use this strategy of keeping the number density fixed when using the NN distributions to analyze different cosmologies and obtain constraints in Sec. 4.

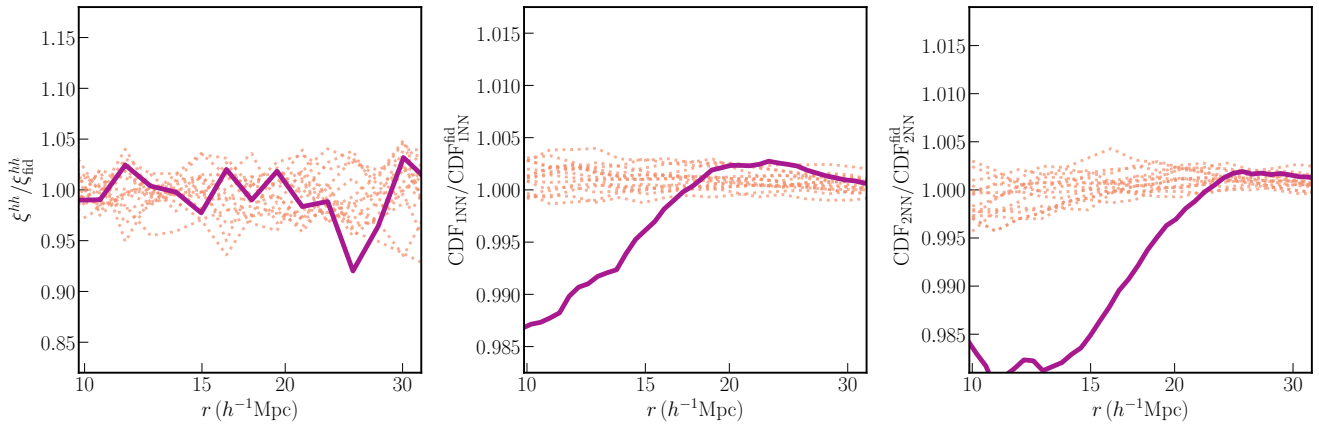

Figure 6. *Left*: The darker line represents the ratio of correlation function of the 10^5 most massive halos in a $(1h^{-1}{\rm Gpc})^3$ box at redshifts $z = 0$, and $z = 0.5$. The lighter shaded lines represent the ratio of the correlation functions at $z = 0$ for 15 different realizations of the same cosmology, divided by the mean correlation function at that cosmology. *Center*: The darker line represents the ratio of the nearest neighbor CDF of the 10^5 most massive halos in a $(1h^{-1}{\rm Gpc})^3$ box at redshifts $z = 0$, and $z = 0.5$. The lighter shaded lines represent the ratio of the nearest neighbor CDFs at $z = 0$ for 15 different realizations of the same cosmology, divided by the mean nearest neighbor CDF at that cosmology. *Right*: Same measurements as the center panel, except with second nearest neighbor distances instead of the first. Even though the correlation function of the two samples at different redshifts are almost indistinguishable within sample variance uncertainties, the NN CDFs are clearly separated.

4 COSMOLOGICAL PARAMETER CONSTRAINTS

In this section, we explore the degree to which constraints on various cosmological parameters improve when the same simulation datasets, using the same scale cuts, are analyzed using the nearest neighbor framework developed above, compared to the traditional two-point analysis. We focus on the following range of scales, $10h^{-1}$Mpc to $40h^{-1}$Mpc, *i.e.*, scales smaller than those where the linear, Gaussian approximation is valid, but larger than the typical sizes of even the largest virialized structures in the universe. We avoid using smaller scales, where the gains from using the nearest neighbor statistics are potentially even larger, in the analysis to avoid any possible systematics related to the resolution of the simulations used in the analysis.

4.1 Fisher formalism

The Fisher matrix formalism has been widely used to estimate the constraints on cosmological parameters given a set of summary statistics, from which the relevant "data vector" is constructed, and the expected error bars on the measurement of these summary statistics. For a cosmological survey, the error bars depend on specifications such as the sky area covered by the survey, the depth, and the number density of tracers. When the summary statistics under consideration are two-point correlation functions, or power spectra, the error bars are relatively easy to compute once the survey specifications are known. Therefore, the Fisher formalism can be used to estimate constraints on cosmological parameters around some fiducial cosmology, even before the survey starts collecting data.

For summary statistics other the two-point correlation function, it is often not possible to analytically compute the error bars, or more generally, the covariance matrix between various entries in the data vector. The Fisher framework has been applied in such situations to estimate the parameter constraints from mock datasets. These datasets are often generated from cosmological simulations, where these non-trivial summary statistics can computed directly from the simulation outputs. This is the spirit in which we use the Fisher formalism in this work.

Formally, the elements of the Fisher matrix ($\mathbf{F}$) is defined as

$$\mathbf{F}_{\alpha\beta} = \sum_{i,j} \frac{\partial D_i}{\partial p_\alpha} \Big[\mathbf{C}^{-1}\Big]_{ij} \frac{\partial D_j}{\partial p_\beta}, \tag{22}$$

where D_i are the entries of the data vector, p_α represent various cosmological parameters, and $\mathbf{C}$ is the covariance matrix for the data vector, evaluated at some fiducial cosmology. The Fisher matrix can then be inverted to determine the constraints on individual parameters, while marginalizing over the uncertainties in all other parameters. as well as the covariances between different parameters. In particular,

$$\sigma_\alpha = \sqrt{\left(\mathbf{F}^{-1}\right)_{\alpha\alpha}}, \tag{23}$$

where σ_α represents the 1-σ constraint on parameter α.

To construct both the derivatives of the data vector with respect to the cosmological parameters, as well as the covariance matrix, we use the Quijote[4] simulations (Villaescusa-Navarro et al. 2019). These N-body simulations were run on $(1h^{-1}{\rm Gpc})^3$ volumes with 512^3 CDM particles in cosmologies with no massive neutrinos, and with 512^3 CDM and 512^3 neutrino particles in cosmologies with massive neutrinos. The mean inter-particle separation for the particles is therefore $\sim 2h^{-1}$Mpc, and to be conservative in our analysis, we only use measurements above $10h^{-1}$Mpc in our analysis. The cosmological parameters that are included in the analysis are $\{\Omega_m, \Omega_b, \sigma_8, n_s, h, M_\nu, w\}$. The Quijote simulations have been run in a way that the derivatives with respect to each of these parameters can be easily computed around a fiducial cosmology. These simulations have already been used to estimate the information content of various non-trivial statistics of the cosmological field (Hahn et al. 2020; Uhlemann et al. 2019).In general, this is done by running simulations with one parameter larger (and smaller) than the fiducial cosmology, while all other parameters are held at their fiducial value. Special care has to be taken to compute the derivatives with respect to the total neutrino mass M_ν, and this is discussed in detail in Villaescusa-Navarro et al. (2019).

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

The data vector for the analysis is constructed from the measurement of the kNN-CDF for $k = \{1, 2, 4, 8\}$. Each distribution has a different functional dependence on all the n-point functions present in the data. It also worth reiterating that we do not *a priori* know which n-point functions are relevant for the clustering. Using multiple values of k, spread over a relatively wide range, as with our particular choice, ensures that the data vector has fewer degeneracies. We have checked that using slightly different combinations of k do not have a major effect on the cosmological constraints. We discuss this further below. We use 10^5 data points (simulation particles or halos), and 10^6 volume-filling random points to generate the Empirical CDF as outlined in Sec. 2.2. Each CDF is interpolated to determine its value at 16 logarithmically spaced values of r. Since the analysis is focused on scales from $10h^{-1}$Mpc to $40h^{-1}$Mpc, the values of r for all the CDFs lie within this range. However, to ensure that we do not go too deep into the tails where the finite number of random points starts to affect the ability to accurately measure the distributions, we impose stricter scale cuts for each NN-CDF. For each k, we determine the range of scales for which the Erlang CDF distribution (see Eq. 12) for that k, and the same mean number density, lies between 0.005 and 0.995. As can be seen from Fig. 4, using the Erlang distribution is a conservative choice for the analysis of both simulation particles, and halos, since the tails of the latter distributions extend out further than the reference Poisson distribution. We then choose the 16 logarithmically spaced r for each k over the range of scales that are allowed after taking into account both the cuts of $10h^{-1}$Mpc to $40h^{-1}$Mpc, and the cuts implied by the Erlang distribution tails. We then append the 16 measurements for each k into a single data vector with 64 entries. In most of analysis, we combine measurements from $z = 0$ and $z = 0.5$, by once again combining the 64 entry data vector computed for each redshift into a single data vector with 128 entries.

Once the data vector is defined and computed on every simulation, the derivative term for each parameter in Eq. 22 is computed by averaging over 100 realizations at the relevant cosmologies. This helps reduce the noise from both sample variance, and the fact that these derivatives are computed numerically. For Fisher matrix analysis, care needs to be taken that the derivatives are smooth, since numerical noise can lead to spurious features in derivatives which are then interpreted as artificially tight constraints. We inspect the derivatives to ensure that no such pathological features exist over the range of scales we use in the analysis. Some of the derivatives are discussed further in Sec. 4.2.

The covariance matrix is computed from 1000 realizations at the fiducial cosmology. The entries of the raw covariance matrix are given by

$$\mathbf{C}'_{ij} = \left\langle \Big(D_i - \langle D_i \rangle\Big)\Big(D_i - \langle D_i \rangle\Big) \right\rangle, \tag{24}$$

where the labels i, j represent various rows in the data vector and $\langle ... \rangle$ represents an average over realizations. Note that we set the off-diagonal terms of data vector entries corresponding to two different redshifts to 0. This is done to avoid spurious covariances arising from the use of the same realizations of the cosmology at $z = 0$ and $z = 0.5$. To compute the correct inverse covariance matrix that goes into the Fisher analysis, we use the Hartlap correction factor (Hartlap et al. 2007):

$$\mathbf{C}^{-1} = \frac{n - p - 2}{n - 1} \left(\mathbf{C}'\right)^{-1}, \tag{25}$$

where n is total number of entries in the data vector, and p is the number of simulation realizations used in the evaluation of $\mathbf{C}'$. Notice that this constrains the allowed length of the data vector, given the number of simulations available. For accurate error forecasts, the Hartlap factor should be as close to unity as possible. Since our aim is to only demonstrate the gain in constraining power when using kNN analysis over $\xi(r)$ analysis for the same analysis choices, and not the actual values of the constraints, we use the native simulation volume of $(1h^{-1}\text{Gpc})^3$ throughout our analysis. Current and future cosmological surveys typically have much larger volumes, and consequently can lead to much tighter absolute constraints than the ones presented here, even after accounting for possible systematics arising in the data. While inverting the covariance matrix, we check that the condition number is within acceptable limits, for each of the analyses presented below. We also explicitly check that the distribution of deviations of the data vector around the mean are roughly Gaussian for different entries in the data vector, implying that the Fisher error estimates should be roughly valid.

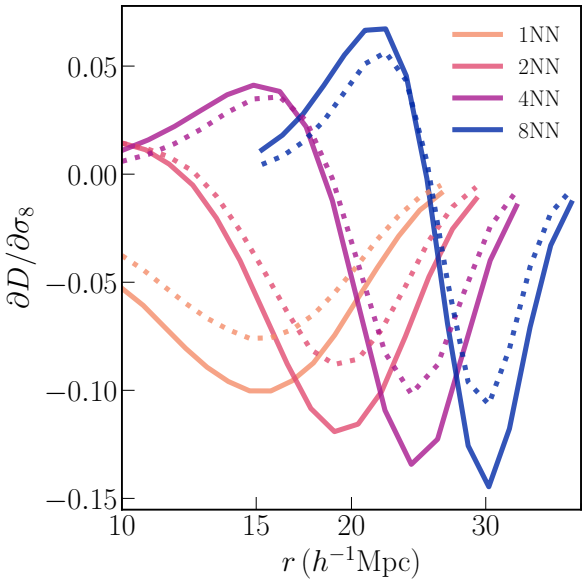

Figure 7. The derivative of the data vector with respect to the cosmological parameter σ_8. The different colored curves represent the portions of the data vectors coming from $k = \{1, 2, 4, 8\}$ nearest neighbor CDF distributions. The solid lines represent the derivative at $z = 0$, while the dotted lines represent the derivative at $z = 0.5$.

We use exactly the same framework to compute the constraints on cosmological parameters from the two-point correlation function over the range $10h^{-1}$Mpc to $40h^{-1}$Mpc. At each redshift, we compute $\xi(r)$ in 30 logarithmically spaced bins. We then combine the data vectors at $z = 0$ and $z = 0.5$ into a single data vector. The rest of the analysis proceeds the same way as outlined above. In principle, the Fisher analysis on the two-point function could also have been carried out using analytic estimates of the covariance matrix. However, using the same pipeline for both observables helps keep both analyses consistent in terms of numerics, while also serving as a systematics check on the kNN-CDF results.

4.2 Constraints from Dark Matter density field

In this section, we present the constraints on the cosmological parameters when considering the 3-dimensional Dark Matter density

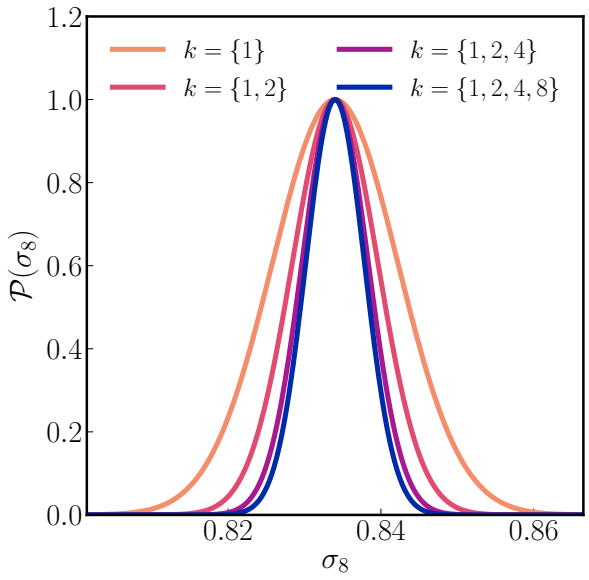

Figure 8. The posterior distribution for σ_8, marginalized over all other parameters. The different colors represent different kNN combinations from which the constraint was obtained. The constraints improve as more nearest neighbor distributions are added, but the gain saturates by the time we add all four CDFs that are computed from the data.

field, where the simulation particles are tracers of the field. For the nearest neighbor analysis, we downsample the simulation particles to 10^5 before computing the different distributions. As shown in Sec. 3.2, the downsampling determines the range of scales over which we can accurately measure the features of the underlying continuous field. The specific choice we make here is to ensure that the measurements are robust in the range of scales that enter the Fisher analysis, i.e. $10h^{-1}$Mpc to $40h^{-1}$Mpc. To reduce the sampling variance caused by downsampling, for every cosmology, we create 16 distinct downsamplings of 10^5 particles each from the original 512^3 particles in the simulation, and then compute the Empirical CDF by combining the nearest neighbor measurements from individual downsampled datasets. For the correlation functions, we use the Corrfunc code[5] (Sinha & Garrison 2019; Sinha & Garrison 2020) to compute $\xi(r)$ using all 512^3 particles in the simulation. Once the data vector is defined in this way, we compute the covariance matrix and the derivatives of the data vector with respect to the cosmological parameters. Note that we use only the CDM particles, for both the nearest neighbor analysis, and the $\xi(r)$ analysis, even for cosmologies with massive neutrinos. In massive neutrino cosmologies, the CDM density field, traced by the CDM particles, and the total matter field, which determines quantities relevant for gravitational lensing, for example, are different. The total matter field includes the contribution from the clustering of neutrinos, and is known to be more constraining on the total neutrino mass M_ν. For this work, we only present the constraints on the cosmological parameters, including the neutrino mass, from the CDM field only. We will explore the possible stronger constraints when using the total matter field in a future work.

[5] https://github.com/manodeep/Corrfunc

In Fig. 7, we plot the derivative of the data vector with respect to σ_8. The different colors represent the parts of the data vector that come from different kNN distributions. The solid lines represent the part of the data vector computed from particles at $z = 0$, and the dashed lines represent the parts of the data vector computed from particles at $z = 0.5$. The sign of the derivative for the part of the data vector coming from nearest neighbor distribution can be understood in the following way — when the amplitude of clustering at 8Mpc, σ_8 is higher, there are more empty void-like regions on larger scales. In other words, on these scales, the probability of finding 0 particles (VPF) in a sphere of radius r is higher when σ_8 is higher. Since $\mathrm{CDF}_{1\mathrm{NN}}(r) = P_{>0|r} = 1 - \mathrm{VPF}(r)$, a higher σ_8 implies a lower value for the data vector entry at these scales if all other cosmological parameters are held fixed. This leads to the negative sign of the derivative seen in Fig. 7. The derivative becomes larger in magnitude as the redshift decreases, as is expected from the growth of structure with time. We check that all the other derivatives are smooth, and do not suffer from numerical artifacts on scales that enter the analysis.

Before moving to the combined constraints on all cosmological parameters, we first show how the constraints on one of the parameters, σ_8, changes as we add information from different kNN distributions at $z = 0$. For this example, we first use a data vector constructed only from the nearest neighbor CDF at z=0, and calculate the constraints on σ_8 from the Fisher analysis on this data vector. Next, we use the measurements from $k = \{1, 2\}$ nearest neighbors, and repeat the calculation. We repeat this until we use all four $k = \{1, 2, 4, 8\}$ nearest neighbor CDFs in our analysis. The posterior distribution for σ_8 from each of these calculations is plotted in Fig. 8. Note that this posterior is marginalized over all other cosmological parameters even though they are not shown here. We find that the constraints improve as more NN distributions are added to the data vector. However, the gain from adding new NN distributions diminishes by the time we use all four of the computed CDFs, $k = \{1, 2, 4, 8\}$. A similar trend is observed for other cosmological parameters as well. We conclude that up to the lowest scales in the analysis, the choice of $k = \{1, 2, 4, 8\}$ in our nearest neighbor analysis is sufficient to extract most of the information on the cosmological parameters. If smaller scales are to be included in the analysis, higher k neighbors may have to be considered to ensure maximal constraints on the parameters down to those scales.

The joint constraints on all the cosmological parameters from the simulation particles, using the kNN analysis and the $\xi(r)$ analysis is presented in Fig. 9. Here we use the full data vector outlined in Sec. 4.1. Since the fiducial cosmology has $M_\nu = 0$, and negative M_ν values are unphysical, we follow the example in Uhlemann et al. (2019), and plot constraints on ΔM_ν instead, where ΔM_ν is the change in the total neutrino mass from the fiducial value. The 1-σ constraints on individual parameters is listed in Table 1. For all the cosmological parameters, we find that the kNN analysis improves the constraints by almost a factor of 2 over the $\xi(r)$ analysis. While some of the degeneracy directions between pairs of parameters are somewhat different between the two analyses, we find that both have are affected by a stronge $M_\nu - \sigma_8$ degeneracy, as can be expected when utilizing information only from the relatively small scales used in these analyses. Note that our choice of leaving out the larger scales, including the BAO peak from the analysis also affects the shape and size of the other contours and the degeneracy directions compared to other works. We conclude that on the scales between $10h^{-1}$Mpc and $40h^{-1}$Mpc, using only a two-point function analysis fails to capture at least half the total information about cosmological parameters that is available on these scales. The kNN analysis, on the

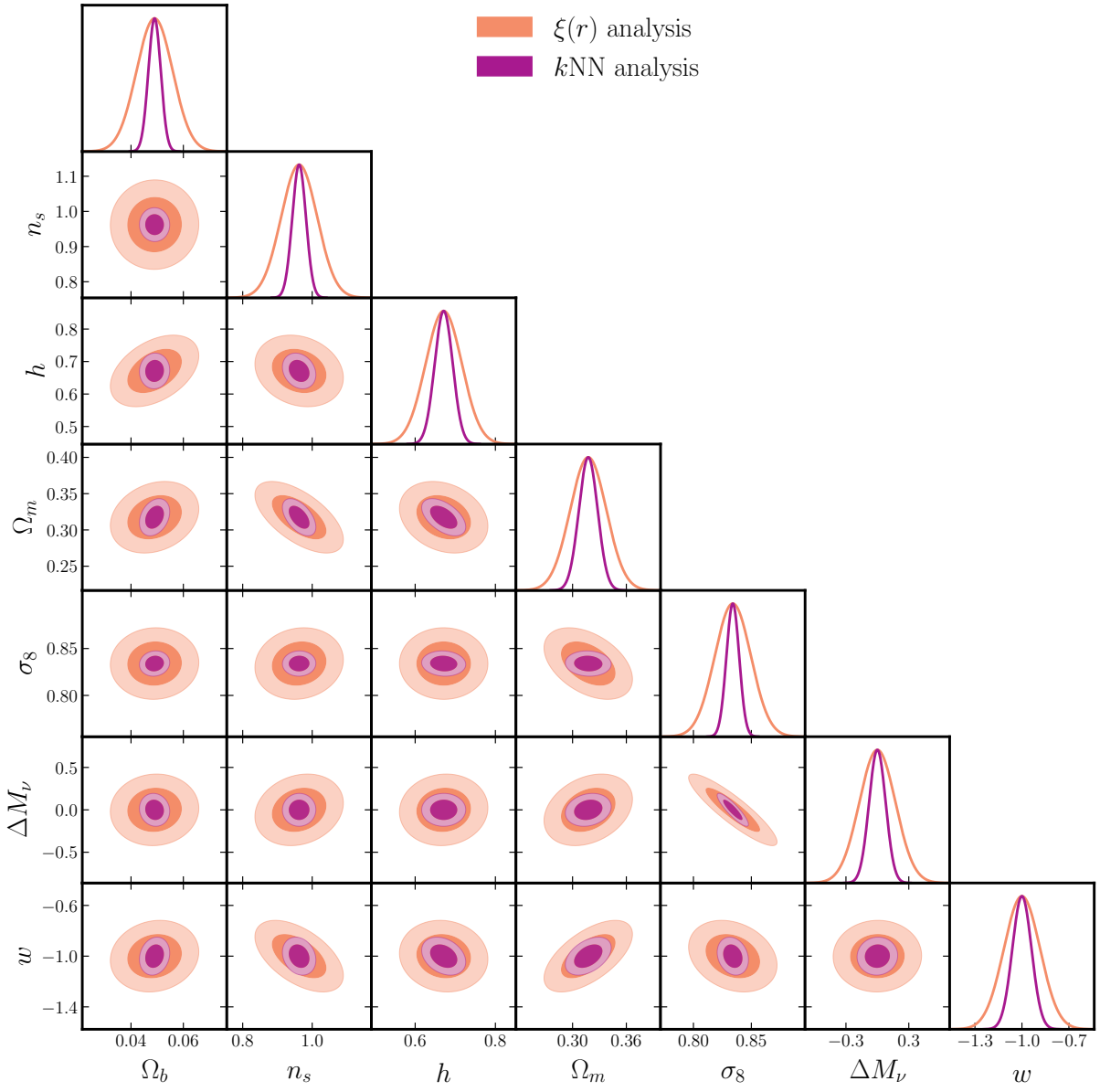

Figure 9. Constraints on the cosmological parameters derived from the Fisher analysis on simulation particles, combining information from $z = 0$ and $z = 0.5$, using scales in the range $10h^{-1}$Mpc to $40h^{-1}$Mpc. The constraints from the nearest neighbor analysis (using the same set of scales) are tighter by more than a factor of 2 for most of the parameters. The values of the constraints on individual parameters are listed in Table 1.

other hand, is sensitive to higher order clustering, and is much more sensitive to changes in the cosmological parameters. In Appendix C, we show that the results presented here are consistent with those obtained from a slightly different formalism where only the nearest neighbor distribution is used, but with different mean densities for the data points.

4.3 Simulation Halos

We now turn to the parameter constraints from the analysis of halos in the simulation.Throughout this section, we will consider fixed number density samples across different cosmologies. Specifically, we focus on the clustering statistics of the 10^5 most massive halos in each simulation volume. The halos are identified using an FoF algorithm from the simulation particles. Even for comsologies with neutrino particles in the simulation, the halo finding is run only on the CDM particles. At $z = 0$, this number cut corresponds to mass cuts around $5\times10^{13}M_\odot/h$, and at $z = 0.5$, corresponds to mass cuts around $3 \times 10^{13}M_\odot/h$. As discussed previously, the kNN statistics is sensitive to the mean number density, and using a fixed number density ensures that differences in the kNN statistics at different cosmologies arise only from changes in the underlying clustering. In the language of the Fisher formalism, this analysis quantifies the response of the clustering observables — kNN distributions or $\xi(r)$ measurements — constructed from the 10^5 *most massive halos* in a $(1h^{-1}\text{Gpc})^3$ volume, to a change in the cosmological parameters, and then converts the amplitude of the response into parameter constraints. Once again, we only use scales in the range $10h^{-1}$Mpc to $40h^{-1}$Mpc in our analysis.

If the two-point function is used as the summary statistic, pa-

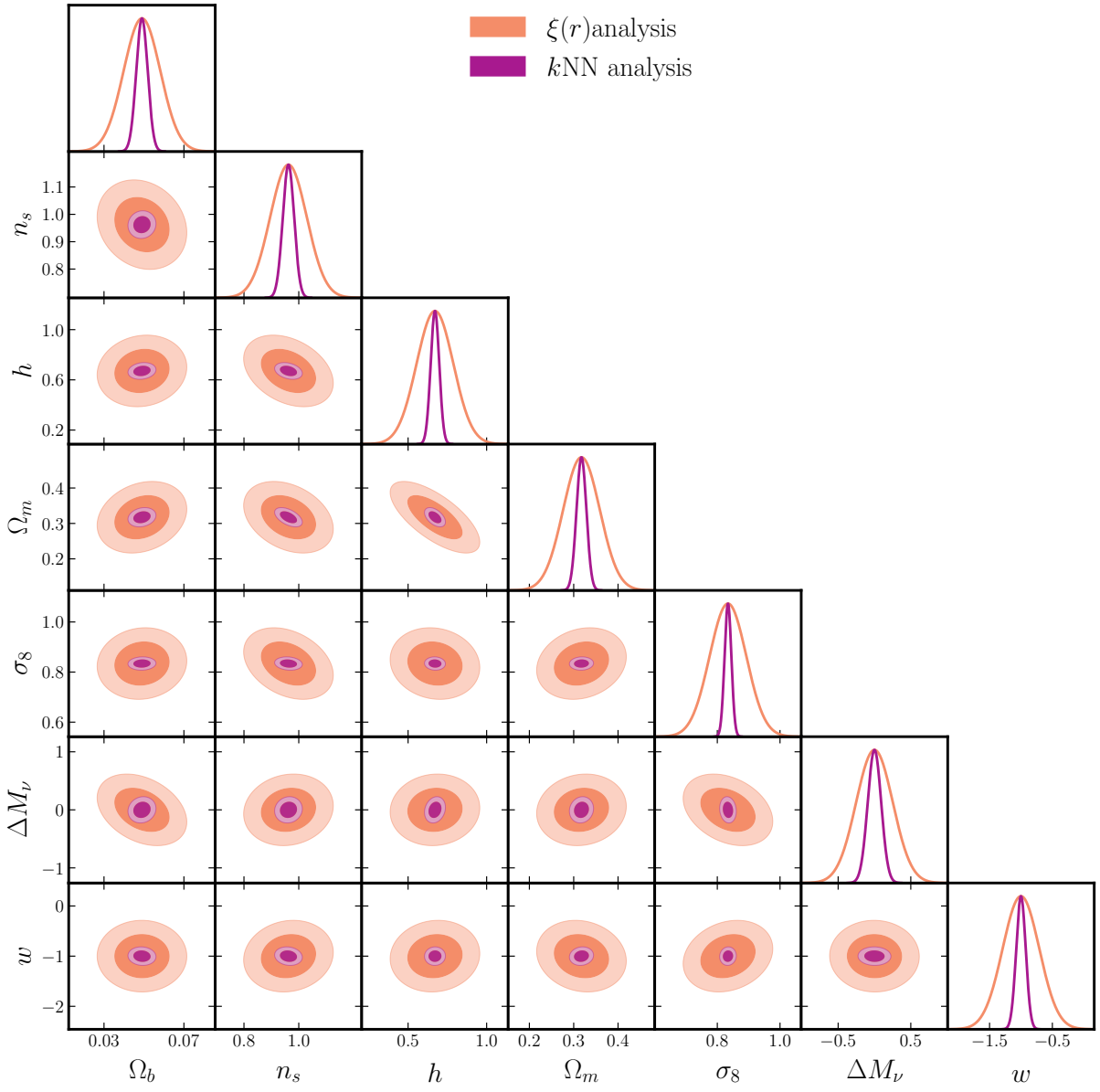

Figure 10. Constraints on the cosmological parameters derived from the Fisher analysis of the clustering of the 10^5 most massive halos at different cosmologies, combining information from $z = 0$ and $z = 0.5$, and using scales in the range $10h^{-1}$Mpc to $40h^{-1}$Mpc. The constraints from the kNN analysis are much tighter than those from the $\xi(r)$ analysis. The values of the constraints on individual parameters are listed in Table 2.

rameter constraints degrade significantly when analyzing halo clustering compared to matter clustering. This is due to halo bias, which, in general, is unknown. Since the effect of the linear bias term on the two-point function, especially on large scales, is degenerate with a number of cosmological parameters, like σ_8 and M_ν, marginalizing over the unknown bias term relaxes the constraints on these. It is only possible to break the degeneracy by either using other observables (usually lensing) which have a different dependence on bias, or by using information from smaller scales, where the linear bias approximation breaks down. Note that, in general, it is not necessary for the linear bias approximation to break down at the same scale as where the underlying clustering becomes non-Gaussian (Modi et al. 2020).

We have already demonstrated in Sec. 3.3.1 that the kNN statistics can break the degeneracy between linear bias and the amplitude of clustering of the underlying field. The kNN statistics were able to distinguish between two halo samples where the underlying clustering of the matter field were different, but the bias (or mass cut) adjusted such that they have the same two-point clustering signal. It is therefore, possible, that more information about the cosmological parameters is retained in the kNN analysis of massive halos, compared to a $\xi(r)$ analysis.

This is indeed what we find from the Fisher analysis on the clustering of halos in real space, and the results are presented in Fig. 10. The values of the 1-σ constraints on individual parameters are tabulated in Table 2. Unsurprisingly, we find that for $\xi(r)$, the constraints are significantly relaxed when compared to the matter clustering case in Fig. 9 and Table 1. This is most significantly apparent for σ_8, where the constraints degrade by almost a factor of 8. On the other hand, the constraints yielded by the kNN analysis

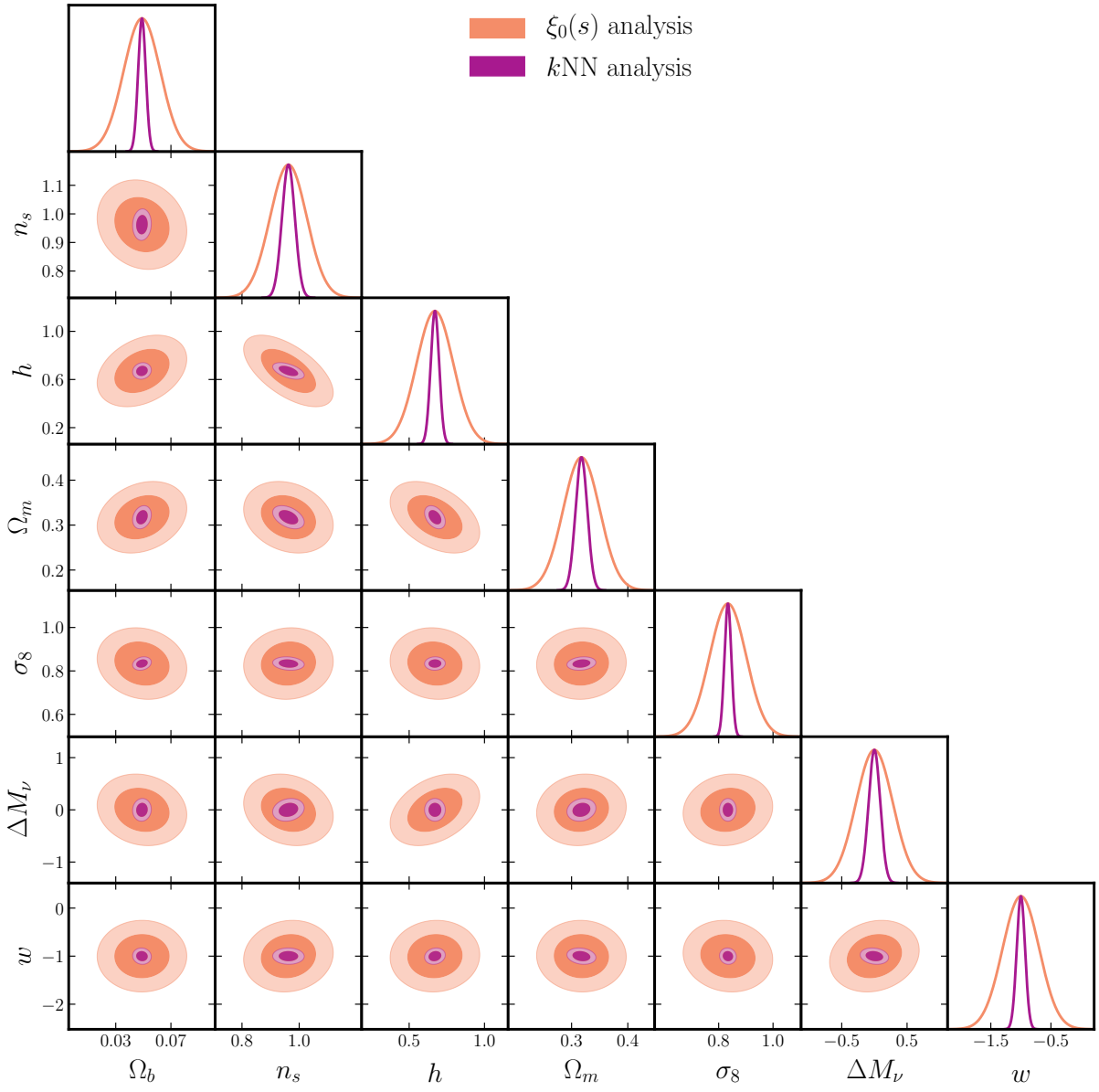

Figure 11. Constraints on the cosmological parameters derived from the Fisher analysis of the monopole of clustering of the 10^5 most massive halos in redshift space, combining information from $z = 0$ and $z = 0.5$, and using scales in the range $10h^{-1}$Mpc to $40h^{-1}$Mpc. Similar to Fig. 10, the constraints from the kNN analysis are much tighter than those from the $\xi(r)$ analysis.

are much closer to those from the analysis of the matter field. Even though the constraint on σ_8 is less stringent than for the matter field analysis, it is still tighter than the $\xi(r)$ constraint by a factor of ~ 5.

Therefore, the constraints on all cosmological parameters are significantly improved when using the kNN statistics, and the improvement over the two-point analysis using the same scales is larger when considering the 10^5 most massive halos than when considering the clustering of the underlying matter field. This has many potential implications for the analysis of galaxy clustering in cosmological surveys, both photometric, and spectroscopic.

While the analysis presented above considers the clustering of halos in real space, spectroscopic surveys generally measure the clustering of galaxies in redshift space. It is, therefore, worthwhile to understand, if the projection to redshift space affects the relative improvements on parameter constraints from the kNN analysis over the $\xi(r)$ analysis. To convert the real space positions $\mathbf{r}$ of the halos in the simulation volume to redshift space positions $\mathbf{s}$, we simply use

$$\mathbf{s} = \mathbf{r} + \frac{1+z}{H(z)} v_x \hat{x}\,, \tag{26}$$

where $H(z)$ is the Hubble parameter at redshift z. In the above equation, we have assumed that the $\hat{x}$ direction is the line-of-sight direction.

Because the line of sight projection breaks the rotational symmetry, clustering in redshift is no longer isotropic, and it is common to include the quadrupole and the hexadecapole of the correlation function in addition to the monopole. To keep the analysis simple, we focus here only on the monopole for the two point correlation function, and compare it to the kNN statistics in redshift space. To construct the CDFs for the kNN distributions, we proceed exactly

Table 1. 1-σ constraints on cosmological parameters, from the kNN and $\xi(r)$ analysis of simulation particles using scales in the range $10h^{-1}$Mpc to $40h^{-1}$Mpc.

Parameter	$\sigma_{k\mathrm{NN}}$	$\sigma_{\xi(r)}$
Ω_b	0.0023	0.0069
n_s	0.0197	0.0521
h	0.0224	0.0450
Ω_m	0.0105	0.0201
σ_8	0.0055	0.0156
ΔM_ν	0.0792	0.173
w	0.0612	0.115

Table 2. 1-σ constraints on cosmological parameters, from the kNN and $\xi(r)$ analysis of the real space clustering of the 10^5 most massive halos in the simulation volume of $(1h^{-1}\mathrm{Gpc})^3$, using scales in the range $10h^{-1}$Mpc to $40h^{-1}$Mpc.

Parameter	$\sigma_{k\mathrm{NN}}$	$\sigma_{\xi(r)}$
Ω_b	0.0029	0.0092
n_s	0.0206	0.0667
h	0.0273	0.1165
Ω_m	0.0111	0.0413
σ_8	0.0108	0.0584
ΔM_ν	0.0925	0.2520
w	0.0756	0.2916

as for the real space calculations, once all the halo positions have been transformed according to Eq. 26.

The results of the analysis are plotted in Fig. 11. Once again, we find that using the kNN statistics yields much tighter constraints on the cosmological parameters compared to the two-point function analysis. The gain is very similar to those that were obtained in the real space analysis above. Adding in the quadrupole and hexadecapole two-point functions should add more constraining power to the two-point analysis, but it should be noted that the kNN statistics can also be expanded in the presence of a special line-of-sight direction. For example, the nearest neighbor distributions along the LOS direction and the perpendicular directions can be computed separately - the differences in the distribution along individual directions arise exactly from the redshift projection. We will explore these issues in detail in future work.

5 SUMMARY AND FUTURE DIRECTIONS

In this paper we have presented the nearest neighbor distributions (kNN-CDF), the empirical cumulative distribution function of distances from a set of volume-filling Poisson distributed random points to the k–nearest data points, as a set of new summary statistics that can be used in cosmological clustering analysis. We have shown how these are related to the counts-in-cell distributions, and to the various N-point correlation functions of the clustering of a set of data points. We have outlined how these distributions can be efficiently computed over a range of scales using the data points, or a subsample of them, and a set of volume filling randoms. We have applied the kNN-CDF formalism to the clustering of tracers drawn from two different continuous distributions, *i.e.* the Poisson sampling of an uniform field, and a Gaussian field. We also demonstrate its application to data from a cosmological simulation - both for simulation particles, and dark matter halos identified in the simulations. We have explored how the nearest neighbor distributions can break the degeneracy between linear halo bias and the amplitude of clustering of the underlying matter distribution, which is a major limitation of the two-point statistics on large scales. Then, employing the Fisher matrix formalism, we have quantified the constraints that can be placed on various cosmological parameters by using the kNN-CDF as the summary statistics. Using simulation particles, and scales between $10h^{-1}$Mpc and $40h^{-1}$Mpc, we demonstrate that the use of kNN-CDF distributions improve the constraints by a factor of ~ 2 over those from the two-point correlation function analysis. We have demonstrated that these gains are even larger, roughly a factor of 4, when applied to the analysis of the clustering of halos in real space with a fixed number density, and that these gains are expected to persist even after projecting to redshift space, making it a promising tool for analyzing cosmological data.

In addition to the sensitivity of these kNN-CDF statistics to changes in the underlying cosmology, the fact that they can computed extremely efficiently, is highly advantageous. Other methods which aim to harness the extra constraining power, beyond the two-point information, usually are computationally expensive. For example, measuring the bispectrum is much more expensive than the power spectrum, while the trispectrum is even more so. Other approaches, such as measuring the density PDF at different scales usually require multiple Fourier transforms, at least two for each scale at which the PDF is to be computed. For the kNN-CDFs on the other hand, search times for the k-th nearest neighbor does not scale with k once the tree has been constructed, and so, information about higher order cumulants can be accessed at negligible additional computational cost. Further, a single computation is sufficient to provide information on a wide range of scales, and a range of values for k. While we have focused in this paper on clustering in 3 dimensions, the formalism presented can easily be applied to non-Gaussian data in 2 dimensions, where the calculations are computationally even cheaper due to the reduction in the dimensionality. This framework is therefore also well suited to the study the tomographic clustering of galaxies in a photometric survey.

For distributions with only a finite number of connected N-point functions, say m, measuring only the m nearest neighbor distributions is sufficient to capture the full clustering. For cosmological applications, one does not *a priori* know the value of k, or if k is even finite, but as has been demonstrated, we exhaust most of the independent information by the time, *e.g*, the 8-th nearest neighbor distribution is used in the analysis. This value is of course dependent on the choice of scale cuts used in the analysis, and if smaller scales are used, it is likely that a few more nearest neighbor measurements will be needed to capture all the information. As we have pointed out, we have been conservative in our choice of scale cuts based on the resolution of the simulations used in the analysis, and including smaller scales will yield even greater improvements in parameter constraints over the two-point analysis.

While the NN-CDF formalism lends itself most naturally to the analysis of clustering of discrete points, we have demonstrated that it can also be used to study statistical properties of continuous fields. The matter density field has been analyzed by downsampling the simulation particles. Similarly, statistical properties of other continuous fields, such as the convergence fields for weak lensing measurements, can also be analyzed in the same way by appropriately sampling the field. The sampling rate is determined by the scales of interest, such that different distribution functions are robustly measured, and have to be tailored for the application at hand. However, once the sampling rate has been fixed, all other aspects

of the analysis presented here can be carried through without major changes in the workflow.

In the current work, we have used the Fisher formalism to demonstrate the improvement in constraining power when NN-CDFs are used, the Fisher formalism is only reliable when certain approximations are valid. Further, note that the empirical CDF is composed of $\sim 10^6$ measurements, but in the Fisher analysis, we only use the CDFs interpolated to 16 radial scales. This is needed such that the Hartlap factor in Eq. 25, given the number of simulations, remains close to unity, and that the Fisher constraints are reliable. Any analysis of actual data will likely demand more sophisticated statistical techniques. Since the actual measurements in the nearest neighbor calculations are those of Cumulative Distribution Functions, the Kolmogorov–Smirnov (KS) test (Smirnov 1948) lends itself quite naturally as an alternative. Another advantage of the KS test, when a large number of measurements is at hand is that the test has more statistical power when the empirical CDF is sampled at higher rates. Other measures such as the Kullback–Leibler divergence (Kullback & Leibler 1951) can also potentially be applied in conjunction with the NN-CDF measurements, and these issues will be explored in more detail in future work.

We have focused here on the improvement in parameter constraints for samples where the number densities are high enough for the two-point function to be computed using standard methods without being strongly affected by shot noise considerations. However, as demonstrated in Sec.3.2, the NN-CDF method also allows for the measurement of the clustering signal even on scales where shot noise is expected to dominate, i.e. on scales smaller than the mean inter-particle separations. The stronger the clustering, the easier it is to robustly detect the clustering at a fixed mean number density. In cosmology, interesting samples with low number density, such as the most massive clusters, also tend to be highly clustered with respect to the underlying matter field, implying that the NN-CDF method can be applied to the study of the clustering of extremely rare objects such as the most massive galaxy clusters, which are usually completely noise-dominated when computed using the standard methods.

Another possible application of nearest neighbor statistics is as a test of non-Gaussianity in the clustering of objects. If the number density of a sample is well known, and the variance of the distribution can be computed as a function of scale, the nearest neighbor measurements can be used to check if the clustering matches that of a fully Gaussian distribution, as discussed in Sec. 3.2. Any departures from the expressions for the nearest neighbor distributions derived in Sec. 3.2 can be interpreted as the presence of non-Gaussian terms. On small scales, this can help test the range of validity of analysis methods that rely on assumptions of Gaussianity, while on larger scales such measures could be relevant for the search for primordial non-Gaussianity in Large Scale Structure.

At a fixed cosmology, the NN-CDF measurements can be used to study the clustering of different halo samples. One such application is in the context of halo and galaxy assembly bias. For example, different galaxy samples targeted in spectroscopic surveys are known to occupy different parts of the cosmic web (Orsi & Angulo 2018; Alam et al. 2019). Under these conditions, a single linear bias parameter in the measured two-point correlation function does not reflect the full difference in the clustering (Wang et al. 2019). The nearest neighbor framework offers a way to capture these differences more clearly. On a related note, adding NN-CDF measurements to the data vectors used in studies of the galaxy-halo connection can add to the power of these models, and can improve constraints on cosmological parameters even after marginalizing over all the galaxy-halo connection parameters.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Energy SLAC Contract No. DE-AC02-76SF00515. The authors would like to thank Susmita Adhikari and Michael Kopp for useful discussions, and Daniel Gruen, Michael Kopp, Johannes Lange, Jeff Scargle, Istvan Szapudi, Cora Uhlemann, and Francisco Villaescusa-Navarro for comments on an earlier version of the manuscript. 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[6] analysis library was used extensively in this paper. We acknowledge the use of the GetDist[7] (Lewis 2019) software for plotting.

REFERENCES

Abbott T. M. C., et al., 2018, Phys. Rev. D, 98, 043526
Alam S., et al., 2017, MNRAS, 470, 2617
Alam S., Zu Y., Peacock J. A., Mand elbaum R., 2019, MNRAS, 483, 4501
Armijo J., Cai Y.-C., Padilla N., Li B., Peacock J. A., 2018, Monthly Notices of the Royal Astronomical Society, 478, 3627
Baldauf T., Smith R. E., Seljak U. c. v., Mandelbaum R., 2010, Phys. Rev. D, 81, 063531
Balian R., Schaeffer R., 1989, A&A, 220, 1
Banerjee A., Dalal N., 2016, J. Cosmology Astropart. Phys., 2016, 015
Banerjee A., Castorina E., Villaescusa-Navarro F., Court T., Viel M., 2019, arXiv e-prints, p. arXiv:1907.06598
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Bernardeau F., 1994, A&A, 291, 697
Bernardeau F., Pichon C., Codis S., 2014, Phys. Rev. D, 90, 103519
Carrasco J. J. M., Hertzberg M. P., Senatore L., 2012, Journal of High Energy Physics, 2012, 82
Castorina E., Feng Y., Seljak U., Villaescusa-Navarro F., 2018, Phys. Rev. Lett., 121, 101301
Chiang C.-T., LoVerde M., Villaescusa-Navarro F., 2019, Phys. Rev. Lett., 122, 041302
Coles P., Jones B., 1991, MNRAS, 248, 1
Colombi S., 1994, ApJ, 435, 536
Coulton W. R., Liu J., Madhavacheril M. S., Böhm V., Spergel D. N., 2019, J. Cosmology Astropart. Phys., 2019, 043
Dalal N., Doré O., Huterer D., Shirokov A., 2008, Phys. Rev. D, 77, 123514
Desjacques V., Seljak U., Iliev I. T., 2009, MNRAS, 396, 85
Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007, The Astrophysical Journal, 664, 675
Euclid Collaboration et al., 2019, MNRAS, 484, 5509
Fluri J., Kacprzak T., Sgier R., Refregier A., Amara A., 2018, J. Cosmology Astropart. Phys., 2018, 051
Friedrich O., et al., 2018, Phys. Rev. D, 98, 023508
Fry J. N., 1986, ApJ, 306, 358
Fry J. N., Colombi S., 2013, Monthly Notices of the Royal Astronomical Society, 433, 581
Gaztañaga E., Fosalba P., Elizalde E., 2000, ApJ, 539, 522
Gil-Marín H., Noreña J., Verde L., Percival W. J., Wagner C., Manera M., Schneider D. P., 2015, MNRAS, 451, 539

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

Gruen D., et al., 2015, Monthly Notices of the Royal Astronomical Society, 455, 3367
Gruen D., et al., 2018, Phys. Rev. D, 98, 023507
Gualdi D., Gil-Marín H., Schuhmann R. L., Manera M., Joachimi B., Lahav O., 2019, MNRAS, 484, 3713
Hahn C., Villaescusa-Navarro F., Castorina E., Scoccimarro R., 2020, J. Cosmology Astropart. Phys., 2020, 040
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hernández-Aguayo C., Baugh C. M., Li B., 2018, MNRAS, 479, 4824
Hildebrandt H., et al., 2017, MNRAS, 465, 1454
Hinshaw G., et al., 2013, ApJS, 208, 19
Horowitz B., Seljak U., Aslanyan G., 2019, J. Cosmology Astropart. Phys., 2019, 035
Ivanov M. M., Simonović M., Zaldarriaga M., 2020, J. Cosmology Astropart. Phys., 2020, 042
Kerscher M., Pons-Bordería M. J., Schmalzing J., Trasarti-Battistoni R., Buchert T., Martínez V. J., Valdarnini R., 1999, ApJ, 513, 543
Klypin A., Prada F., Betancort-Rijo J., Albareti F. D., 2018, MNRAS, 481, 4588
Kofman L., Bertschinger E., Gelb J. M., Nusser A., Dekel A., 1994, ApJ, 420, 44
Kullback S., Leibler R. A., 1951, Ann. Math. Statist., 22, 79
Lam T. Y., Sheth R. K., 2008, MNRAS, 386, 407
Lawrence E., et al., 2017, ApJ, 847, 50
Lewis A., 2019, arXiv e-prints, p. arXiv:1910.13970
LoVerde M., 2016, Phys. Rev. D, 93, 103526
M Evans N. H., Peacock B., 2000, Measurement Science and Technology, 12, 117
Massara E., Villaescusa-Navarro F., Ho S., Dalal N., Spergel D. N., 2020, arXiv e-prints, p. arXiv:2001.11024
Matsubara T., 2010, Phys. Rev. D, 81, 083505
Mead A. J., Peacock J. A., Heymans C., Joudaki S., Heavens A. F., 2015, MNRAS, 454, 1958
Modi C., Feng Y., Seljak U., 2018, J. Cosmology Astropart. Phys., 2018, 028
Modi C., Chen S.-F., White M., 2020, MNRAS, 492, 5754
Munshi D., van Waerbeke L., Smidt J., Coles P., 2012, MNRAS, 419, 536
Nishimichi T., Bernardeau F., Taruya A., 2017, Phys. Rev. D, 96, 123515
Oort J. H., 1983, ARA&A, 21, 373
Orsi Á. A., Angulo R. E., 2018, MNRAS, 475, 2530
Padmanabhan N., White M., Cohn J. D., 2009, Phys. Rev. D, 79, 063523
Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A. J., Mehta K. T., Kazin E., 2012, MNRAS, 427, 2132
Pan J., Szapudi I., 2005, MNRAS, 362, 1363
Paranjape A., Alam S., 2020, MNRAS, 495, 3233
Peacock J. A., 1998, Cosmological Physics. Cambridge University Press, doi:10.1017/CBO9780511804533
Peel A., Lin C.-A., Lanusse F., Leonard A., Starck J.-L., Kilbinger M., 2017, A&A, 599, A79
Petri A., Haiman Z., Hui L., May M., Kratochvil J. M., 2013, Phys. Rev. D, 88, 123002
Petri A., Liu J., Haiman Z., May M., Hui L., Kratochvil J. M., 2015, Phys. Rev. D, 91, 103511
Planck Collaboration et al., 2018, arXiv e-prints, p. arXiv:1807.06209
Planck Collaboration et al., 2019, arXiv e-prints, p. arXiv:1905.05697
Repp A., Szapudi I., 2020, arXiv e-prints, p. arXiv:2006.01146
Schmittfull M., Feng Y., Beutler F., Sherwin B., Chu M. Y., 2015, Phys. Rev. D, 92, 123522
Scoccimarro R., Colombi S., Fry J. N., Frieman J. A., Hivon E., Melott A., 1998, ApJ, 496, 586
Sefusatti E., Crocce M., Pueblas S., Scoccimarro R., 2006, Phys. Rev. D, 74, 023522
Seljak U., Aslanyan G., Feng Y., Modi C., 2017, J. Cosmology Astropart. Phys., 2017, 009
Sharp N. A., 1981, MNRAS, 195, 857
Sinha M., Garrison L., 2019, in Majumdar A., Arora R., eds, Software Challenges to Exascale Computing. Springer Singapore, Singapore, pp 3–20, https://doi.org/10.1007/978-981-13-7729-7_1
Sinha M., Garrison L. H., 2020, MNRAS, 491, 3022
Slepian Z., et al., 2017, MNRAS, 469, 1738
Smirnov N., 1948, Ann. Math. Statist., 19, 279
Szapudi I., Szalay A. S., 1993, ApJ, 408, 43
Takada M., Jain B., 2004, MNRAS, 348, 897
Taruya A., Bernardeau F., Nishimichi T., Codis S., 2012, Phys. Rev. D, 86, 103528
Tegmark M., 1997, Phys. Rev. Lett., 79, 3806
Uhlemann C., Codis S., Pichon C., Bernardeau F., Reimberg P., 2016, Monthly Notices of the Royal Astronomical Society, 460, 1529
Uhlemann C., Friedrich O., Villaescusa-Navarro F., Banerjee A., Codis S. r., 2019, arXiv e-prints, p. arXiv:1911.11158
Vaart A. W. v. d., 1998, Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, doi:10.1017/CBO9780511802256
Verde L., Heavens A. F., 2001, ApJ, 553, 14
Villaescusa-Navarro F., Marulli F., Viel M., Branchini E., Castorina E., Sefusatti E., Saito S., 2014, J. Cosmology Astropart. Phys., 2014, 011
Villaescusa-Navarro F., et al., 2019, arXiv e-prints, p. arXiv:1909.05273
Vlah Z., Seljak U., Baldauf T., 2015, Phys. Rev. D, 91, 023508
Wald I., Havran V., 2006, in 2006 IEEE Symposium on Interactive Ray Tracing. pp 61–69
Walsh K., Tinker J., 2019, MNRAS, 488, 470
Wang K., et al., 2019, MNRAS, 488, 3541
Way M. J., Gazis P. R., Scargle J. D., 2011, ApJ, 727, 48
White S. D. M., 1979, MNRAS, 186, 145
White M., 2016, J. Cosmology Astropart. Phys., 2016, 057
d'Amico G., Gleyzes J., Kokron N., Markovic K., Senatore L., Zhang P., Beutler F., Gil-Marín H., 2020, J. Cosmology Astropart. Phys., 2020, 005

APPENDIX A: DERIVATION OF THE GENERATING FUNCTION FOR DISCRETE COUNTS

We derive the expression for Eq. 1 by summarizing the arguments presented in Szapudi & Szalay (1993). The generating functional for a continuous field $\rho(\mathbf{r})$ can be written as an integral over all field configurations of ρ:

$$\mathcal{Z}[J] = \int \big[D\rho(\mathbf{r})\big] P\big[\rho(\mathbf{r})\big] \exp\left[i \int d^3\mathbf{r} \rho(\mathbf{r}) J(\mathbf{r}) \right] \tag{A1}$$

where $\big[D\rho(\mathbf{r})\big]$ represents the functional integral over field configurations, and $P\big[\rho(\mathbf{r})\big]$ represents the probability of a specific field configuration. For a mean density $\bar{\rho}$, the cumulants of the distribution are related to the generating functional through the following equation:

$$\bar{\rho}^k \xi^{(k)}(\mathbf{r}_1 ... \mathbf{r}_k) = \frac{i^k \delta^k \big(\ln \mathcal{Z}[J] \big)}{\delta J(\mathbf{r}_1) ... \delta J(\mathbf{r}_k)} \,. \tag{A2}$$

Eq. A2 can be inverted to express the generating functional itself in terms of the cumulants:

$$\mathcal{Z}[J] = \exp\left[\sum_{k=0}^{\infty} \frac{(i\bar{\rho})^k}{k!} \int d^3\mathbf{r}_1 ... d^3\mathbf{r}_k \xi^{(k)}(\mathbf{r}_1 ... \mathbf{r}_k) \times J(\mathbf{r}_1) ... J(\mathbf{r}_k) \right] . \tag{A3}$$

We now consider a set of points generated by a local Poisson process where the number of points generated in a volume V around $\mathbf{r}$ depends only on the local integrated density over that volume, $\rho_V(\mathbf{r})$:

$$\mathcal{M}_V(\mathbf{r}) = \int_V d^3\mathbf{r}' \rho(\mathbf{r}') W(\mathbf{r}, \mathbf{r}') , \tag{A4}$$

where $W(\mathbf{r}, \mathbf{r}')$ defines the window function for the smoothing procedure. Therefore, at a given point $\mathbf{r}$ with integrated density $\mathcal{M}_V(\mathbf{r})$, the probability of finding k points within radius r is

$$P_{k|\mathcal{M}_V} = \frac{(\mathcal{M}_V/m)^k}{k!} \exp\left[-\mathcal{M}_V/m\right], \tag{A5}$$

where m is the "mass" associated with each particle. While this has a relatively straightforward meaning when applied to simulation particles, for halos, this can be thought of as a normalization factor. The overall probability of finding k points within volume V of $\mathbf{r}$ needs to average over all possible field configurations:

$$P_{k|V} = \left\langle \frac{(\mathcal{M}_V/m)^k}{k!} \exp\left[-\mathcal{M}_V/m\right] \right\rangle. \tag{A6}$$

The generating function for the discrete distribution is then defined as

$$\begin{aligned} P(z|V) &= \sum_{k=0}^{\infty} P_{k|V} z^k = \left\langle \frac{(\mathcal{M}_V/m)^k}{k!} \exp\left[-\mathcal{M}_V/m\right] \right\rangle z^k \\ &= \left\langle \exp\left[\mathcal{M}_V(z-1)/m\right] \right\rangle. \end{aligned} \tag{A7}$$

Note that in the last line of Eq. A7, all quantities are in terms of the continuous variables. Therefore this expectation value can be computed through the functional integral over all configurations of the continuous field ρ:

$$P(z|V) = \int \left[D\rho(\mathbf{r})\right] P\left[\rho(\mathbf{r})\right] \times \exp\left[\frac{(z-1)}{m} \int d^3\mathbf{r} \rho(\mathbf{r}) W(\mathbf{r}, \mathbf{r}')\right]. \tag{A8}$$

Therefore, Eqs. A1 and A8 match each other when $J(\mathbf{r}) = W(\mathbf{r}, \mathbf{r}')(z-1)/(im)$. We now use the fact that the RHS of Eqs. A1 and A3 are equivalent, to write

$$P(z|V) = \exp\left[\sum_{k=0}^{\infty} \frac{\left(\bar{n}(z-1)\right)^k}{k!} \int d^3\mathbf{r}_1 ... d^3\mathbf{r}_k \xi^{(k)}\left(\mathbf{r}'_1, ..., \mathbf{r}'_k\right) \times W(\mathbf{r}_1, \mathbf{r}'_1) ... W(\mathbf{r}_k, \mathbf{r}'_k)\right], \tag{A9}$$

where $\bar{n} = \bar{\rho}/m$. For the special case $W(\mathbf{r}, \mathbf{r}') = 1$ only when both points are in the volume considered, and 0 everywhere else, such as in top-hat smoothing, we get the expression in Eq. 1,

$$P(z|V) = \exp\left[\sum_{k=1}^{\infty} \frac{\bar{n}^k (z-1)^k}{k!} \times \int_V ... \int_V d^3\mathbf{r}_1 ... d^3\mathbf{r}_k \xi^{(k)}(\mathbf{r}_1, ..., \mathbf{r}_k)\right]. \tag{A10}$$

APPENDIX B: CUMULANTS

The moment generating function for the discrete distribution is given by

$$\begin{aligned} M(t|V) &= \sum_{k=0}^{\infty} \exp\left[tk\right] P_{k|V} \\ &= \exp\left[\sum_{k=1}^{\infty} \frac{\bar{n}^k (e^t - 1)^k}{k!} \times \int_V ... \int_V d^3\mathbf{r}_1 ... d^3\mathbf{r}_k \xi^{(k)}(\mathbf{r}_1, ..., \mathbf{r}_k)\right]. \end{aligned} \tag{B1}$$

The kth cumulant C_k is given by

$$C_k = \left[\left(\frac{d}{dt}\right)^k \ln\left(M(t|V)\right)\right]_{t=0}. \tag{B2}$$

The first and second cumulants therefore work out to be

$$C_1 = \bar{n}V \tag{B3}$$

$$C_2 = \bar{n}V + \bar{n}^2 \int_V \int_V d^3\mathbf{r}_1 d^3\mathbf{r}_2 \xi^{(2)}(\mathbf{r}_1, \mathbf{r}_2), \tag{B4}$$

where $\xi^{(2)}$ is the usual two-point correlation function. Note that $\xi^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = \xi^{(2)}(r = |\mathbf{r}_1 - \mathbf{r}_2|)$ due to isotropy. For the special case when the volume V is associated with a sphere of radius r_0, we can write C_2 as a one dimensional integral over $\xi^{(2)}(r)$ with some weight function $W(r|r_0)$:

$$C_2 = \bar{n}V + \bar{n}^2 \int_0^{2r_0} dr\, r^2 \xi^{(2)}(r) W(r|r_0), \tag{B5}$$

where

$$W(r|r_0) = 8\pi^2 \left(\frac{1}{12}(2r_0 - r)^3 + \frac{1}{8} r (2r_0 - r)^2\right). \tag{B6}$$

The second moment or variance can also be expressed in terms of the power spectrum, $P(k)$, smoothed on scale r corresponding to volume V:

$$C_2 = \bar{n}V + \frac{(\bar{n}V)^2}{2\pi^2} \int dk\, k^2 P(k) W^2(kr) = \bar{n}V + \bar{n}^2 V^2 \sigma_V^2. \tag{B7}$$

$W(kr)$ represents the Fourier transform of the tophat window function for a filter with radius r:

$$W(kr) = \frac{3}{(kr)^3}\left(\sin(kr) - (kr)\cos(kr)\right). \tag{B8}$$

Note that Eq. B7 is true for any distribution - the distribution need not be a Gaussian. Therefore, $\xi(r)$ or $P(k)$, along with the knowledge of the mean number density, encodes information about the second cumulant of the distribution, even in the nonlinear regime.

The second cumulant can also be computed from the CIC distributions:

$$C_2(V) = \frac{\sum_{k=0}^{\infty} k^2 P_{k|V}}{\sum_{k=0}^{\infty} P_{k|V}} - \left(\frac{\sum_{k=0}^{\infty} k P_{k|V}}{\sum_{k=0}^{\infty} P_{k|V}}\right)^2. \tag{B9}$$

While the sum formally runs from 0 to ∞, in practice, for any realistic distribution function, a finite number of terms is sufficient to determine C_2 accurately. Fig. B1 shows the measurement of $C_2(V)$ for a set of 10^5 simulation particles in a $(1h^{-1}\mathrm{Gpc})^3$ volume at $z = 0$. The solid line represents the measurement from the data using Eq. B9. The CIC measurements themselves are derived from the kNN measurements through Eq. 8. The dotted line represents the measurement of C_2 using Eq. B5, where $\xi(r)$ was computed from all 512^3 particles in the simulation, but $\bar{n}$ was set to match the number density considered for the CIC measurement. The two measurements produce consistent results at the $\sim 1\%$ level, even though we use only 200 nearest neighbor distributions to compute the solid curve - where, formally, the sums in Eq. B9 run from 0 to ∞.

Similarly, higher cumulants can also be obtained directly from the NN-CDF by taking higher moments of the $P_{k|V}$, as in Eq. B9. These cumulants are directly related to the connected higher N-point functions and their Fourier equivalents, just as the second cumulant is related to the two-point function or $P(k)$. We plot the skewness and excess kurtosis, which are related to the third and fourth cumulants,

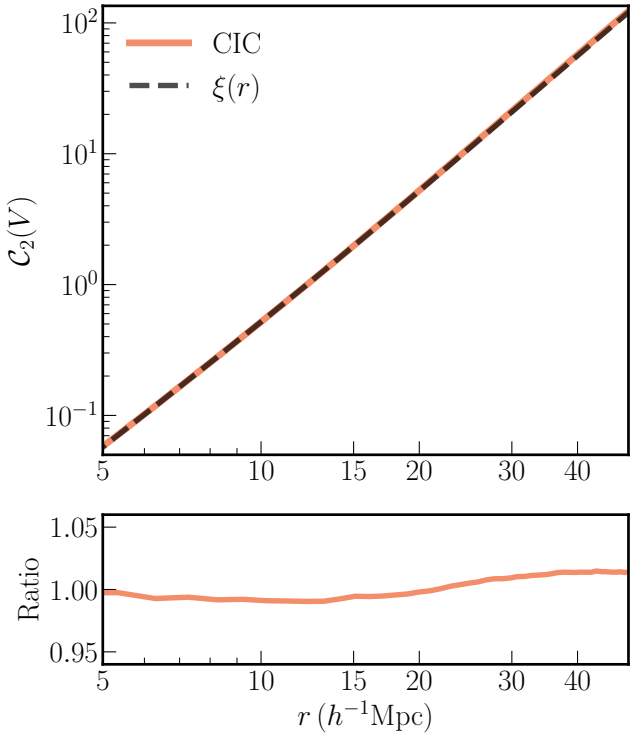

Figure B1. *Top*: Second cumulant C_2, as a function of scale, for the distribution of 10^5 simulation particles in a $(1h^{-1}\mathrm{Gpc})^3$ volume at $z = 0$. The solid line represents the result for C_2 computed using the CIC distributions, according to Eq. B9. The CIC distributions themselves have been computed from the kNN distributions, according to Eq.8. The dashed line represents C_2 computed from the measured $\xi(r)$ of the full set of simulation particles, using Eq. B5. *Bottom*: Ratio of the two measurements from the top panel.

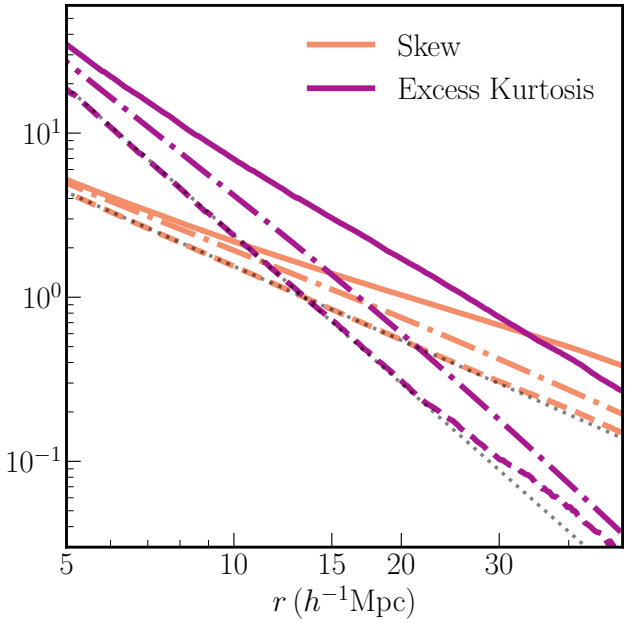

Figure B2. Solid lines represent the skew and excess kurtosis (as a function of scale) of the distribution of 10^5 simulation particles in a $(1h^{-1}\mathrm{Gpc})^3$ volume at $z = 0$. The dot-dashed lines represent the expected skew and kurtosis values for a set of discrete points with the same number density, and the same two point function, but no higher connected $n-$ point functions. The dashed lines represent the measurements for a set of Poisson distributed points with the same number density. The thin dotted lines represent the analytic predictions for the skew and excess kurtosis of the Poisson distribution.

of the same sample of particles from the simulations using the solid lines in Fig. B2. The dot-dashed lines are the expected skew and kurtosis for a set of points with the same number density, and the same two-point function, but with no higher connected $n-point$ functions. The dashed lines of the same color represent the measurements of the the same quantities from a set of Poisson distributed points with the same mean density. The dotted lines represent the analytic predictions for the skew and excess kurtosis for the Poisson distribution. We find very good agreement between the measurements and the predicted values for the Poisson distribution, while the measurements from the cosmological distribution (solid lines) are clearly different from that of the Poisson distribution on all scales displayed on the plot. This is true even in scales below the mean inter-particle separation, where shot noise could potentially be important. Further, comparing the solid lines and the dot-dashed lines on the plot illustrates the fact that this measurement is indeed sensitive to the presence of higher cumulants in the data. Once again, we have terminated the calculation at 200 nearest neighbors, just as we did for the calculation of the second cumulant. It is also worth noting that the time taken to compute quantities related to the second, third, fourth, and potentially higher cumulants is the same - *i.e.* once the NN-CDF distributions for the 200 neighbors has been computed using the tree, the rest of the calculation is computationally trivial, irrespective of the order of the cumulant. Of course, sample variance limitations of the nature discussed in the text imply that the lower cumulants are more robustly measured over a larger range of scales. This can be seen in the departure of the measurement and analytic expectation of the excess kurtosis, or the fourth cumulant, of the Poisson distribution on large scales in Fig. B2.

APPENDIX C: DATA DOWNSAMPLING

In the paper introducing the Void Probability Function, White (1979) showed that the VPF itself can serve as the generating function for the $P_{k|V}$ distribution. In this formalism, however, the derivatives had to be taken with respect to the mean number density $\bar{n}$, unlike Eq. 2, where the derivatives are taken with respect to the dummy variable z. As we have shown earlier,

$$\mathrm{VPF}(r) = 1 - \mathrm{CDF}_{1\mathrm{NN}}(r)\,. \tag{C1}$$

In theory, therefore, the information in the k-th neighbor distribution can also be accessed by computing the nearest neighbor distribution at a different mean number density.

We now compare the actual parameter constraints from the Fisher analysis of kNN distributions for simulation particles, presented in Sec. 4.2, with those that can be obtained by computing the VPF at different number densities. In order to do this, we compute the VPF for subsamples of the simulation particles with $\bar{n} = \{1\times10^{-4}, 0.5\times10^{-4}, 0.25\times10^{-4}, 0.125\times10^{-4}\}(h^{-1}\mathrm{Mpc})^{-3}$, using Eq. C1. Note that the $1\times10^{-4}(h^{-1}\mathrm{Mpc})^{-3}$ is the mean number density used in Sec. 4.2. The other values of $\bar{n}$ are also chosen in a way that they correspond to the kNN measurements presented there. The data vector is created by appending the VPF measurements at 16 logarithmically spaced scales for each $\bar{n}$ in the scale range of $10h^{-1}$Mpc to $40h^{-1}$Mpc. We once again use the relevant Erlang

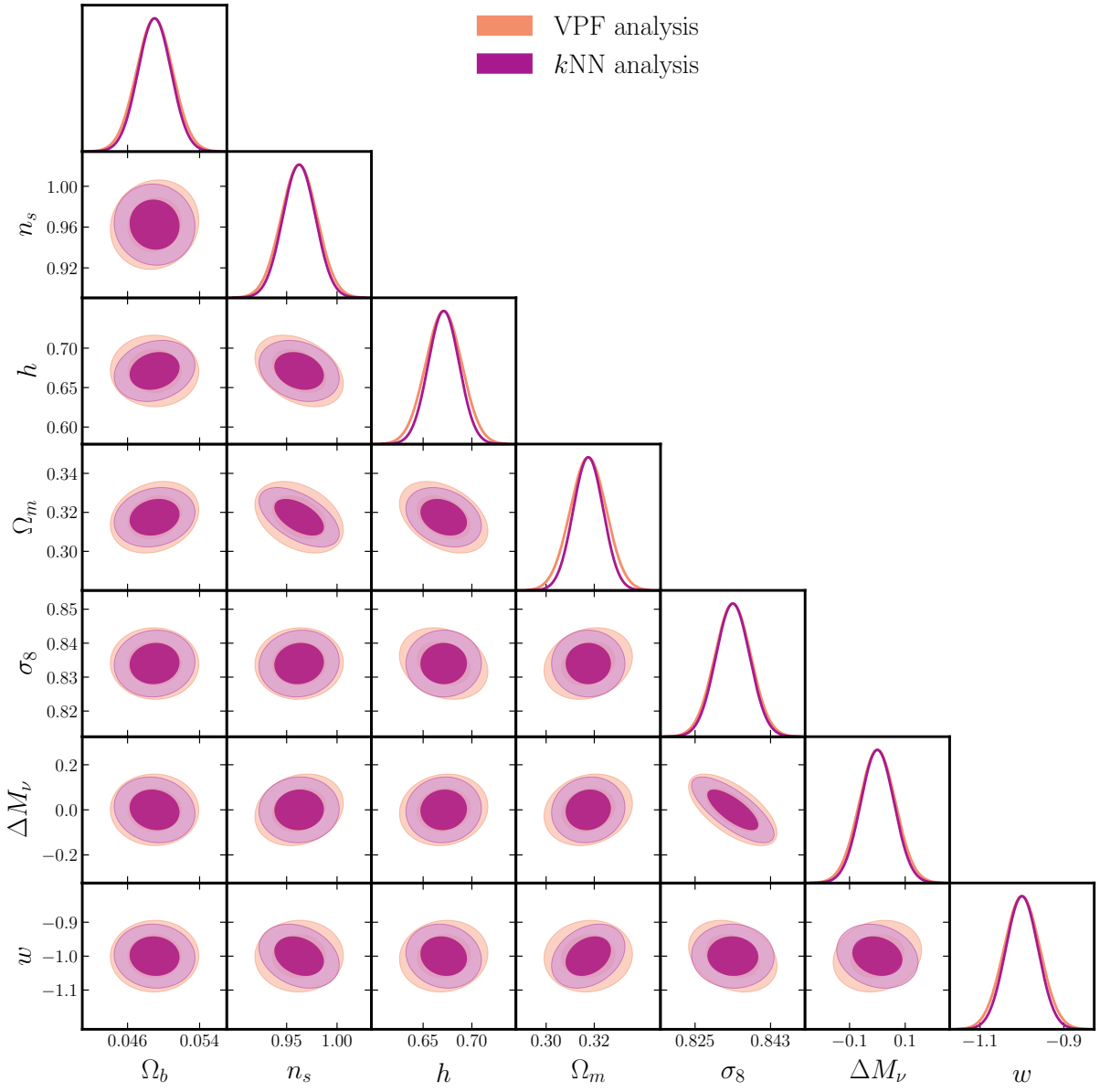

Figure C1. Fisher constraints on cosmological parameters from the kNN analysis in Fig. 9 and the VPF analysis outlined above. The two formalisms yield very similar constraints on all the cosmological parameters.

distribution to ensure that the scales do not include measurements of the VPF deep into the tails. As earlier, we combine measurements from $z = 0$ and $z = 0.5$.

The results of the Fisher analysis are presented in Fig. C1, where the constraints from the VPF formalism above is contrasted with the constraints from the kNN analysis presented in Fig. 9 and Table 1. The final constraints on individual parameters are very similar, though some of the degeneracy directions are slightly different. The agreement in the constraints hold true even though the different derivatives and the covariance matrices between the two analyses look very different. This is a useful check on the robustness of the constraints obtained from the kNN distributions, as a well as a test of the understanding of the underlying statistical methods.

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