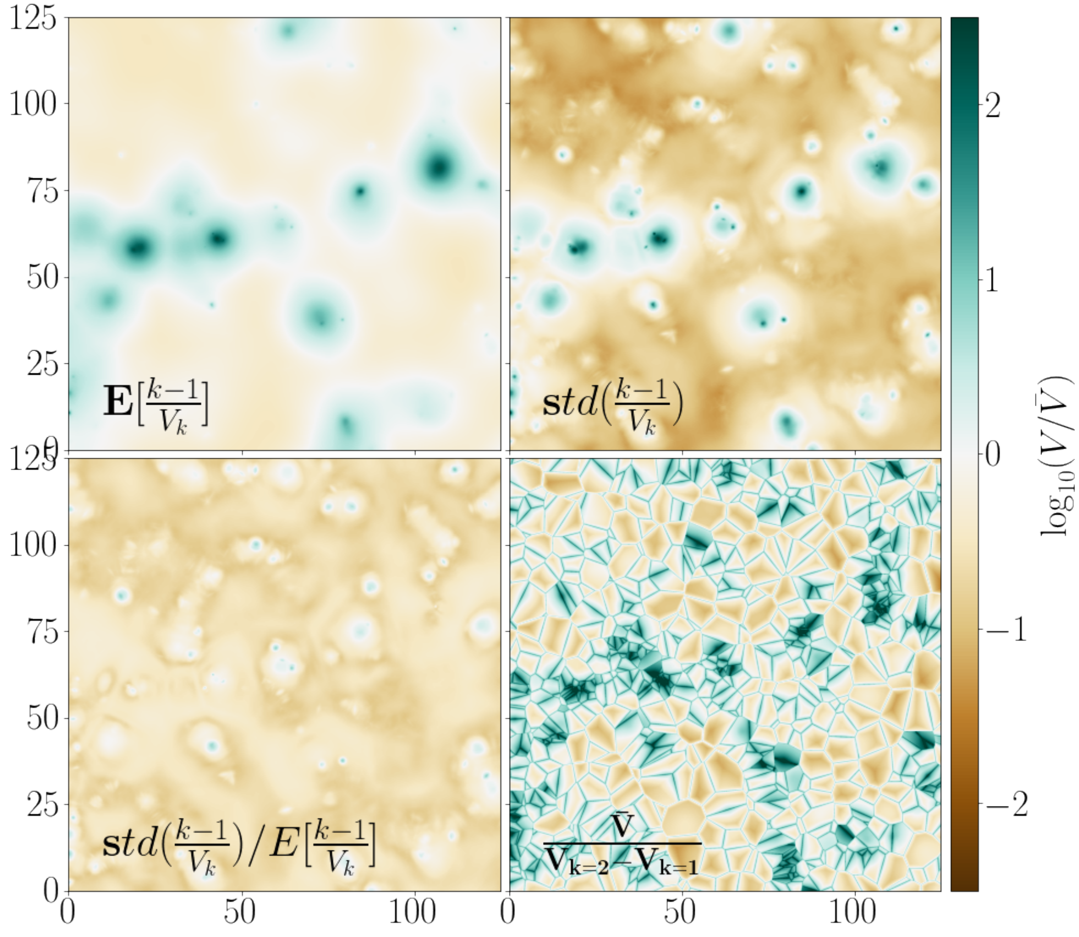
Novel Practical Statistics for Large Scale Structure:

Nearest-Neighbor Cumulative Distribution Functions KNN-CDFs

Arka Banerjee and Tom Abel
With
Richie Wang, Al Zamora, Sandy Yuan, Nick Kokron, Susmita Adikhari, Lehman Garrison, Jessie Muir, …



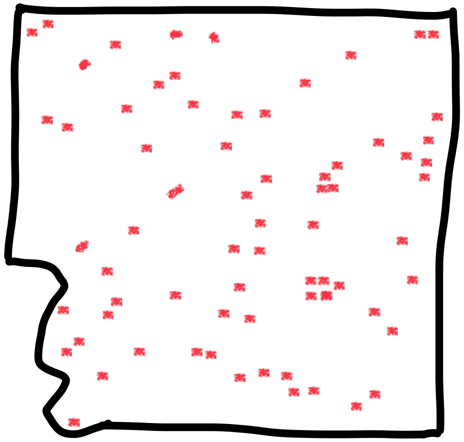
Minimal Problem setup

- We have N equal weight data points labeled by their position in d-dimensional Space of Volume V

  - Goals:

    - Describe their spatial distribution in a way that facilitates the comparison between multiple such realizations

    - Also allow to ask whether the point distribution is a fair sample of continuous density field for which we know its statistics



Metrics [design goals] of useful summary statistics

- Informative
- Interpretable
- Predictable
- Robust
- Minimal number of nuisance parameters (cut off scale, bin widths, smoothing, power, mark, etc.)

- Fast to compute
- No binning/averaging
- Complete with respect to symmetries
- **Discuss: What is Missing?**

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **500,** 5479–5499 (2021)
Advance Access publication 2020 November 20

doi:10.1093/mnras/staa3604

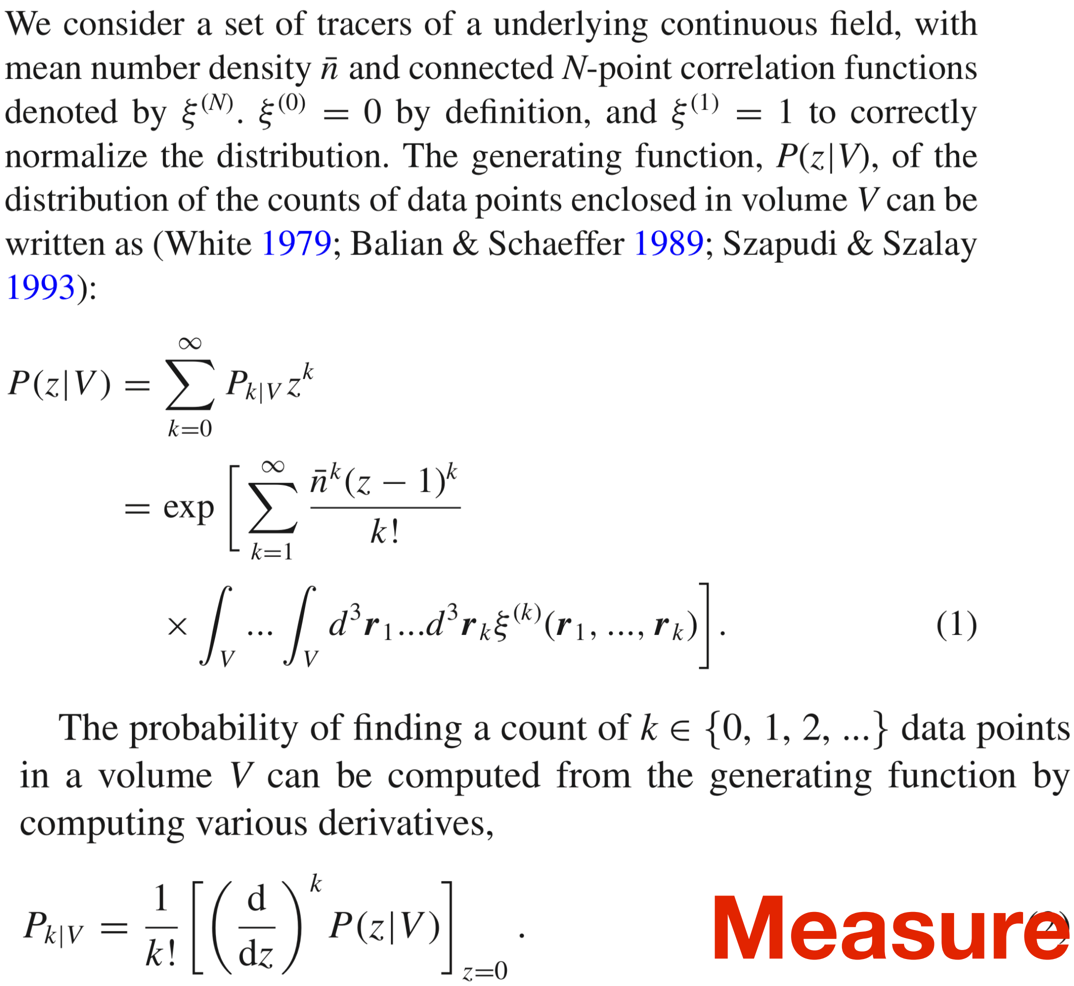
Nearest neighbour distributions: New statistical measures for cosmological clustering

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

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

arXiv:2007.13342

https://github.com/yipihey/kNN-CDFs

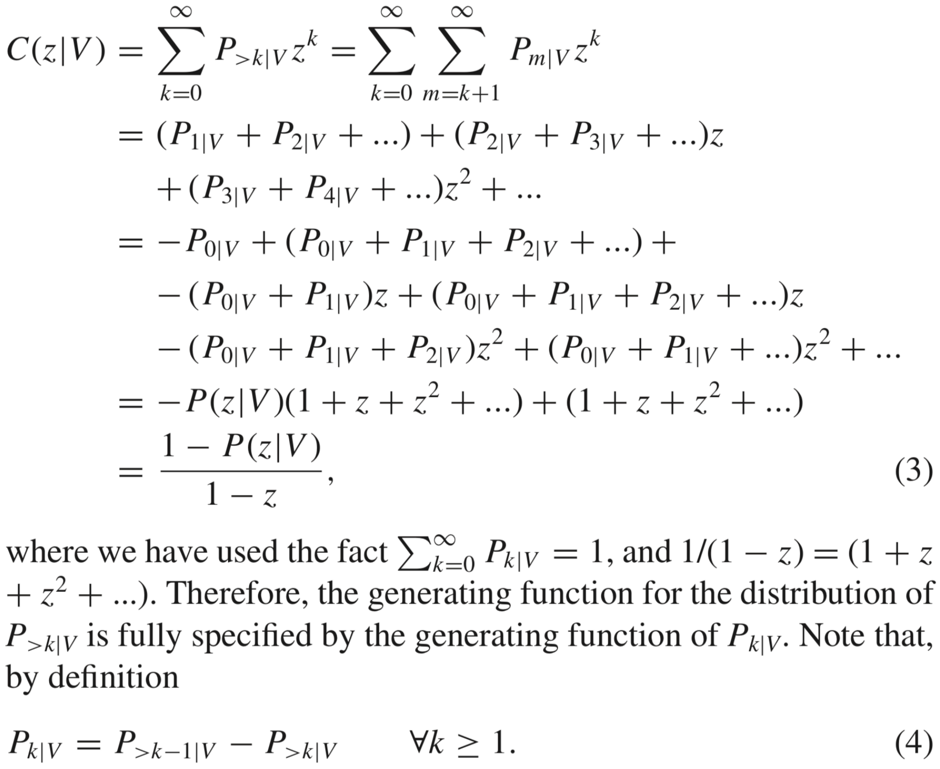


We consider a set of tracers of a underlying continuous field, with mean number density $\bar{n}$ and connected N-point correlation functions denoted by $\xi^{(N)}$. $\xi^{(0)} = 0$ by definition, and $\xi^{(1)} = 1$ to correctly normalize the distribution. The generating function, $P(z|V)$, of the distribution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

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

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,

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



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

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

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

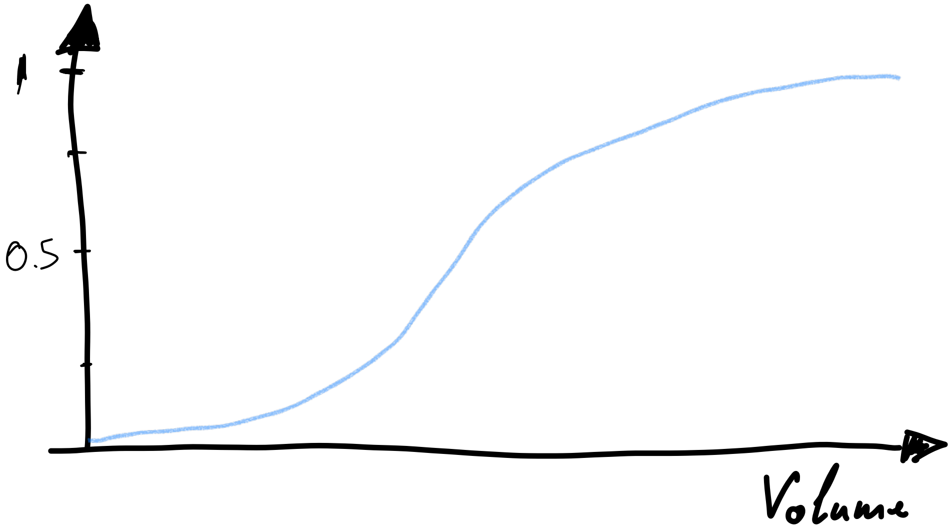
Measure all $P_{k|V}$ or $P_{>k|V}$ or $\xi^{(n)}$

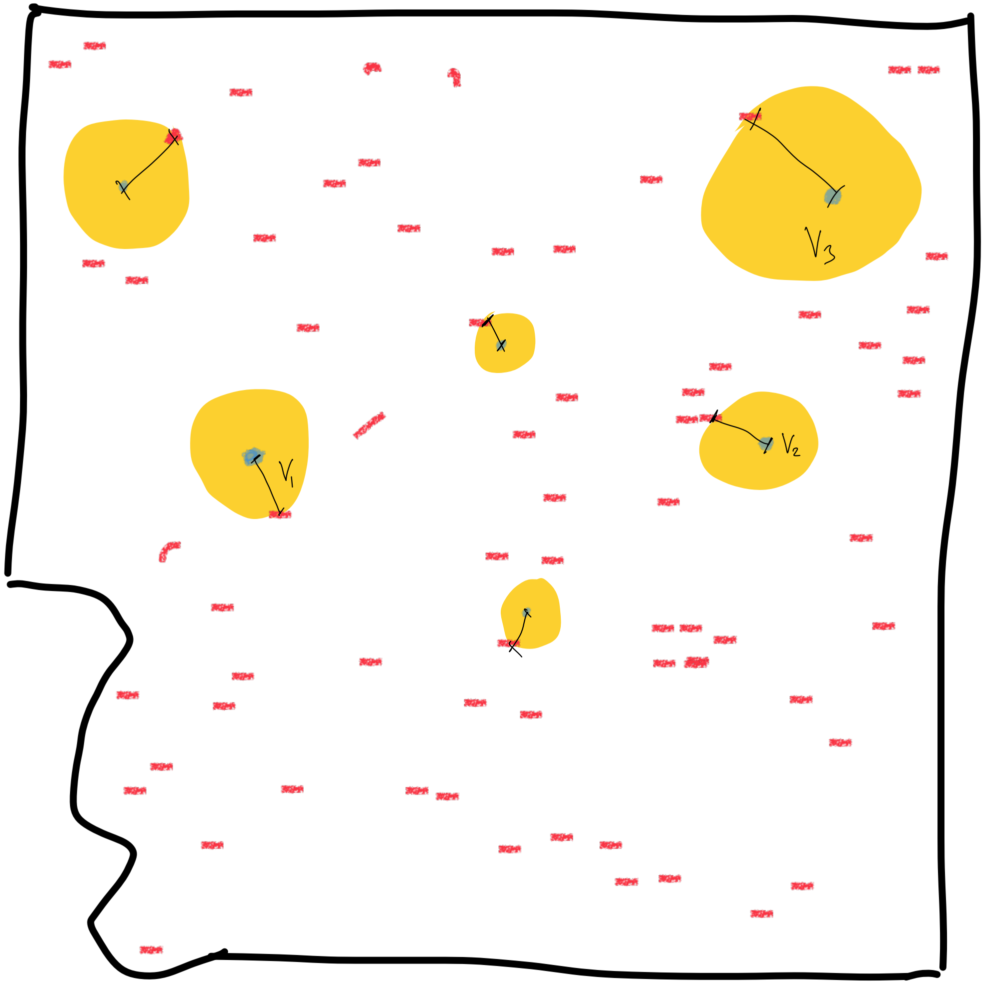
to characterize all counts in cell

Measure volume enclosed by sphere with radius given by the distance to the nearest data point from a very large number of random points R.

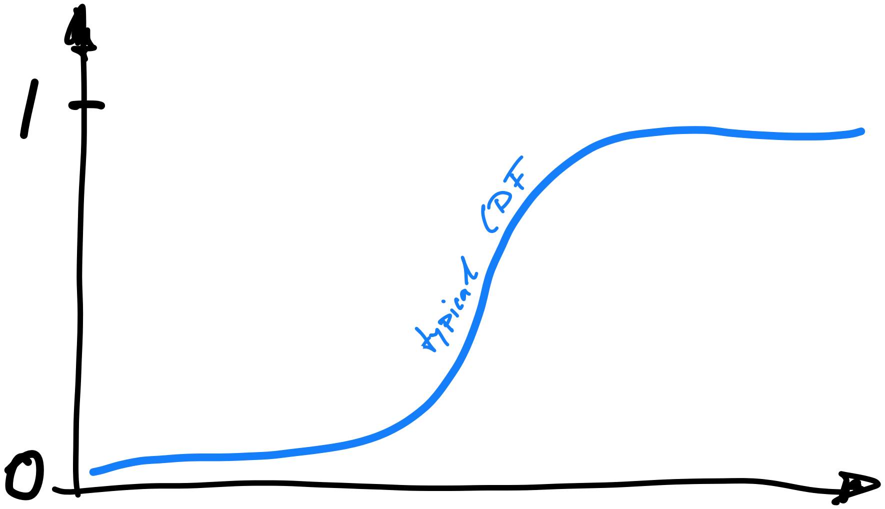
Fraction of points for which the volume of the sphere to the nearest neighbors is less than V is the empirical cumulative distribution function we use.

I.e. we sort resulting volume values which yields the 1NN empirical cumulative distribution function. No binning, no averaging!





Peaked CDFs



- To visualize both tails at small and large distances/ volumes we plot the peaked CDF

$$\mathrm{PCDF(r)} = \begin{cases} \mathrm{CDF}(r) & \mathrm{CDF}(r) \leq 0.5 \\ 1 - \mathrm{CDF}(r) & \mathrm{CDF}(r) > 0.5 \end{cases}$$

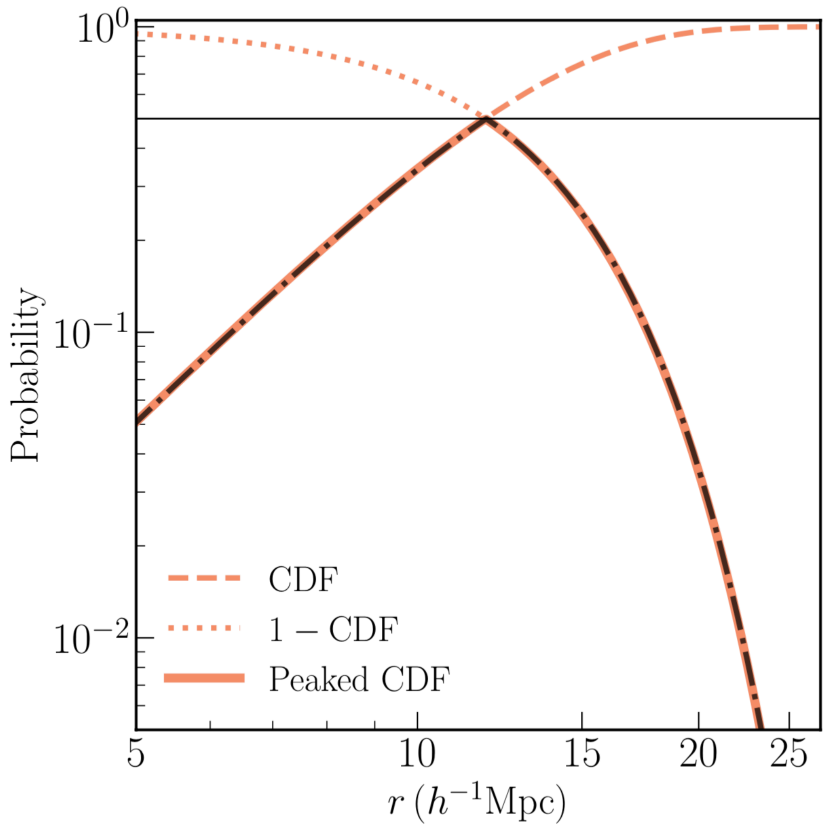


Figure 1. Peaked CDF (described in the text) of the nearest neighbor distribution (as defined in the text) for 10^5 random points distributed over a $(1h^{-1}\mathrm{Gpc})^3$ volume (solid curve). The dot-dashed curve is the analytic prediction for the distribution. The empirical CDF measured from the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

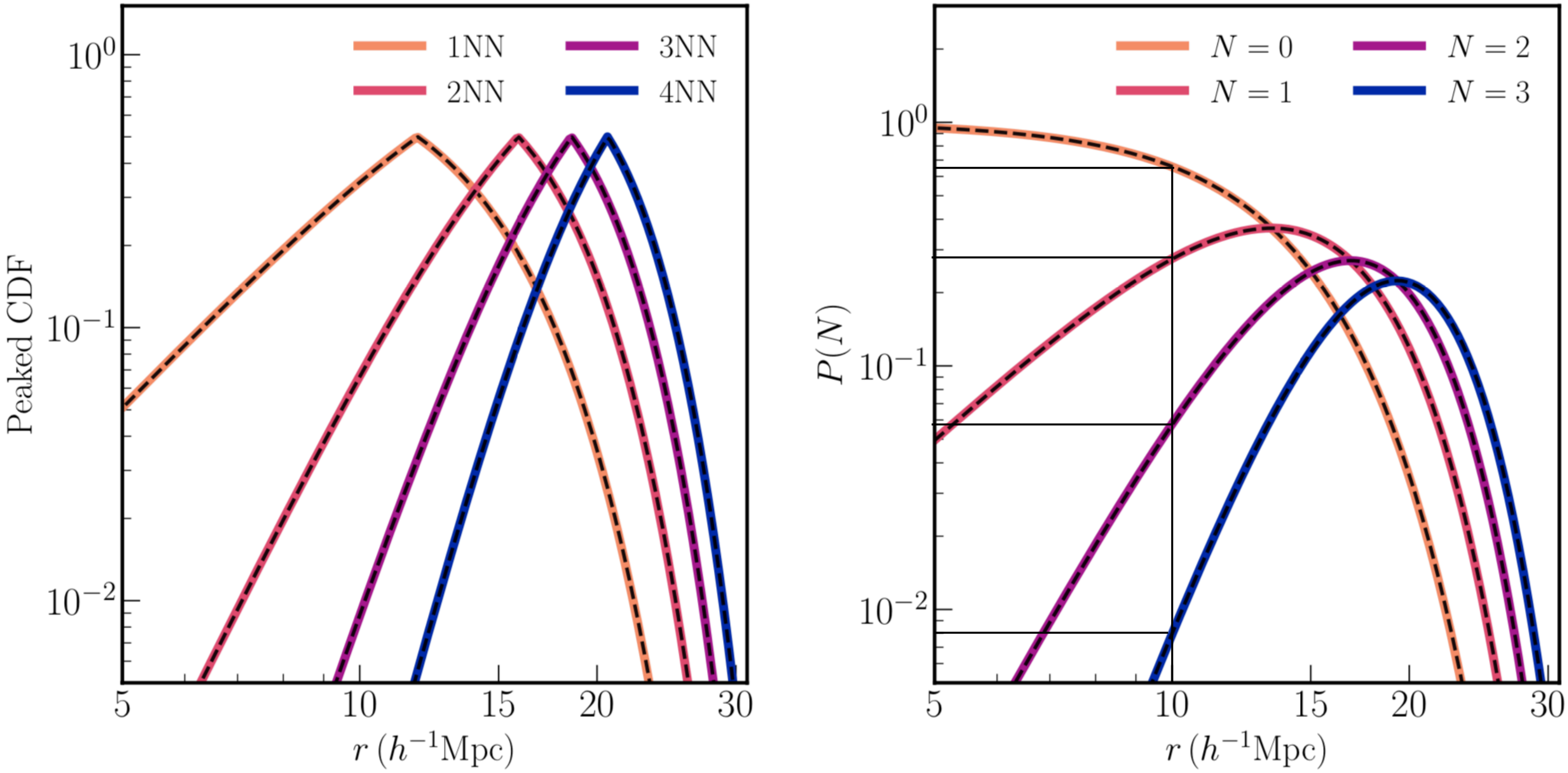
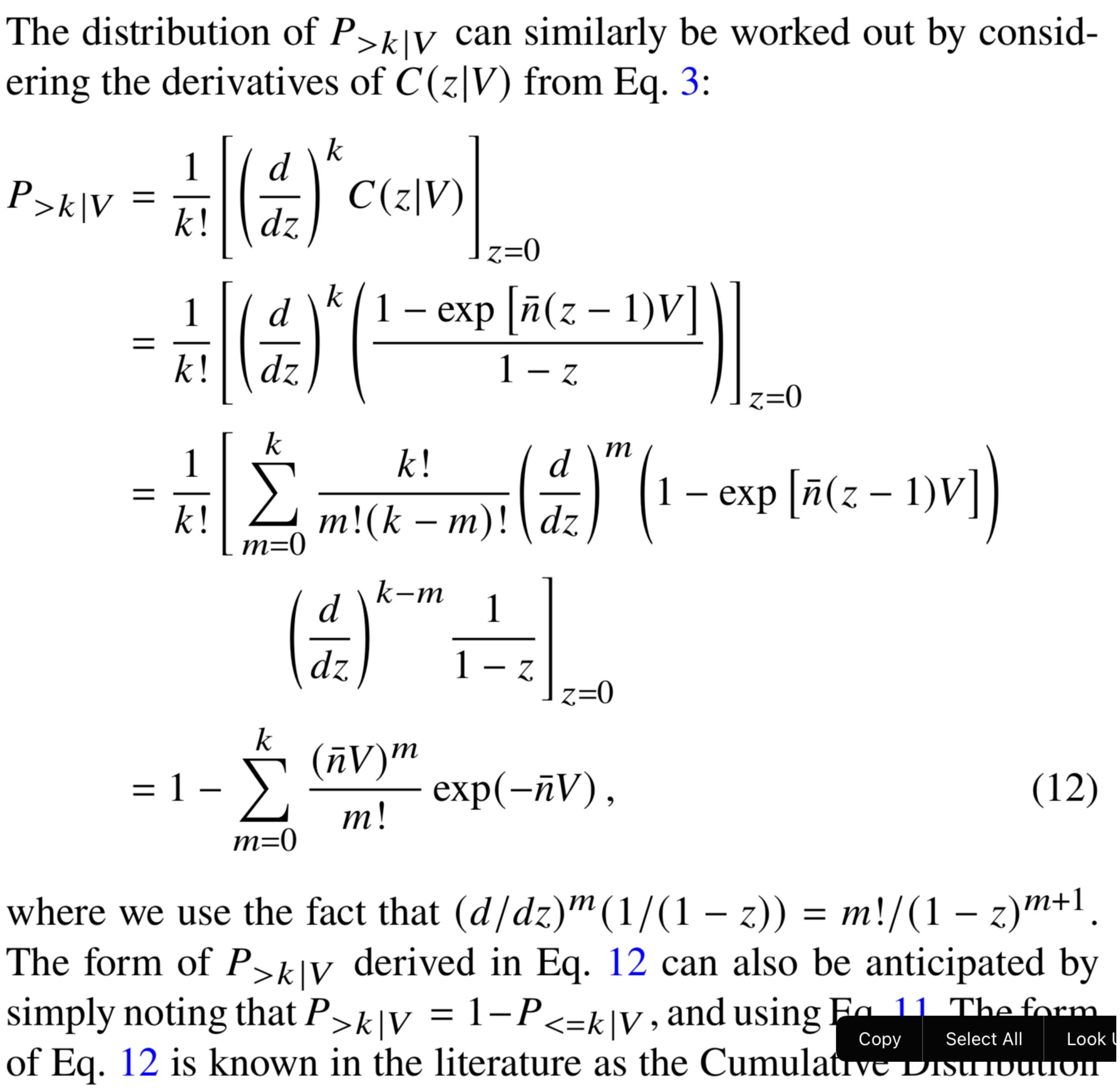


Figure 2. *Left*: Peaked CDF as a function of scale for 1st (1NN), 2nd (2NN), 3rd (3NN), and 4th (4NN) nearest neighbor distributions (solid lines) for a set of 10^5 Poisson distributed data points distributed over a $(1h^{-1}\mathrm{Gpc})^3$ volume from a set of volume filling randoms (see text for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

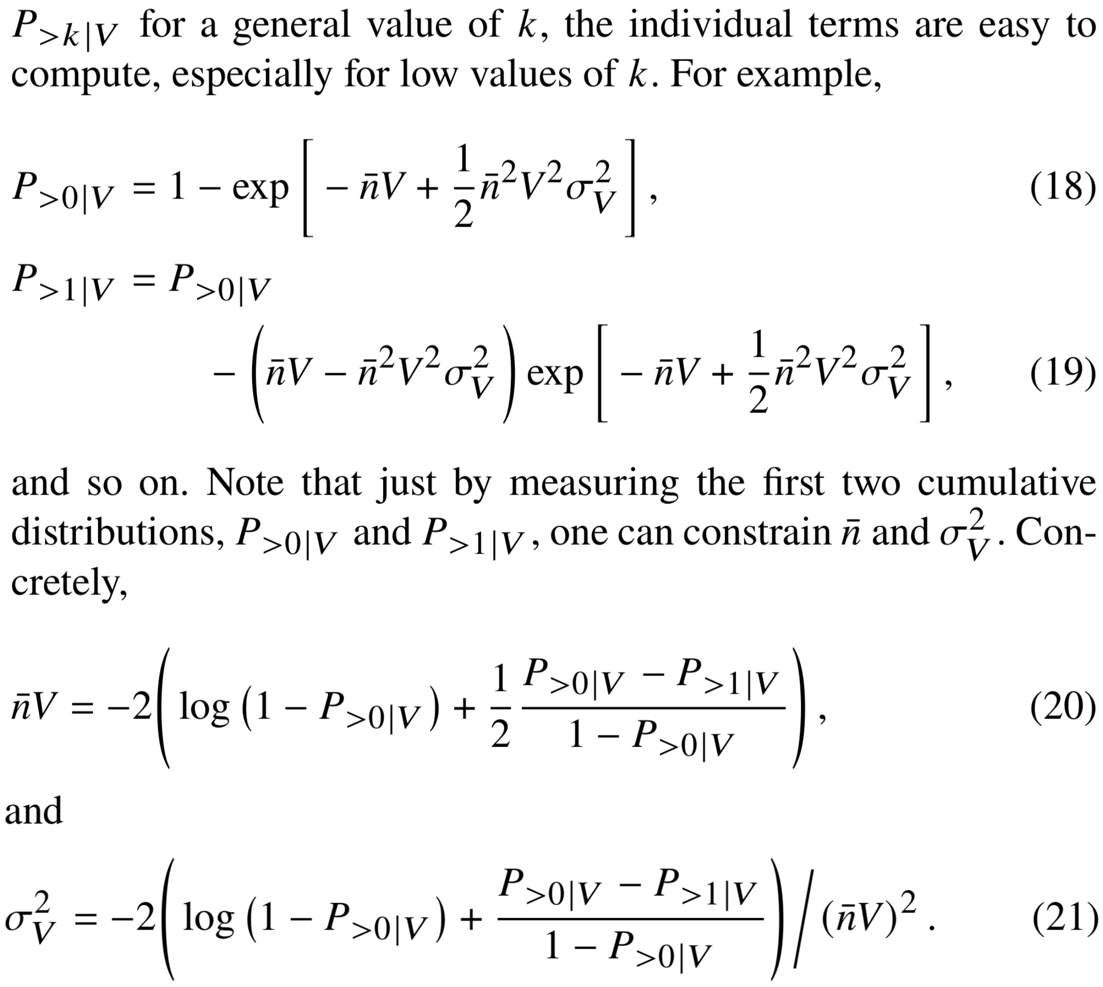


The distribution of $P_{>k|V}$ can similarly be worked out by considering the derivatives of $C(z|V)$ from Eq. 3:

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

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

Knowing $P_{0|V}$ enough to predict all of them in the Poisson Case



$P_{>k|V}$ for a general value of k, the individual terms are easy to compute, especially for low values of k. For example,

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

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

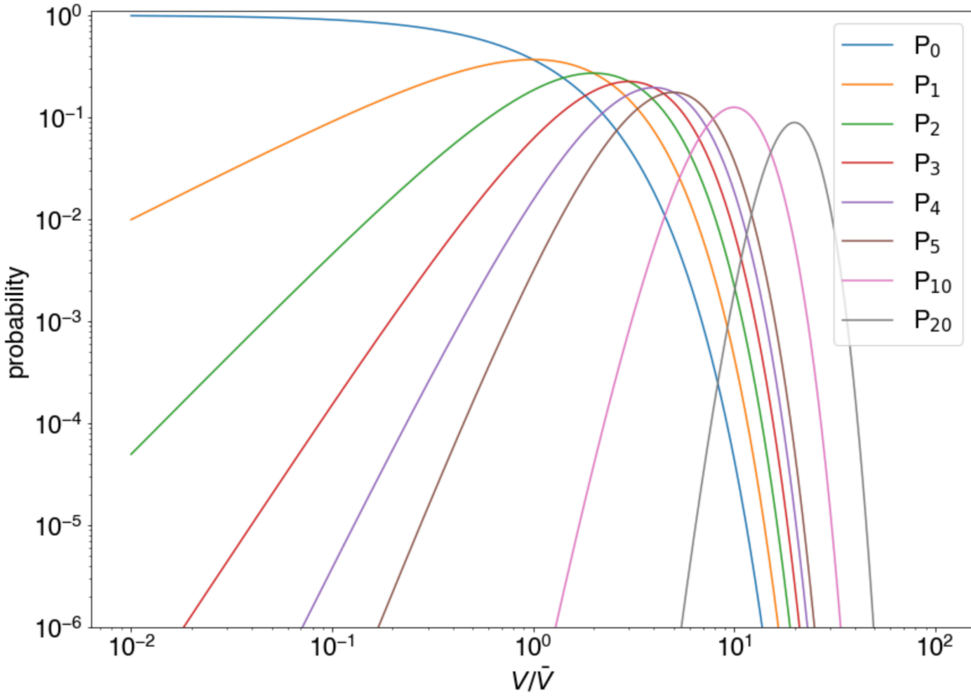
and so on. Note that just by measuring the first two cumulative distributions, $P_{>0|V}$ and $P_{>1|V}$, one can constrain $\bar{n}$ and σ_V^2. Concretely,

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

and

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

Knowing $P_{>0|V}$ and $P_{>1|V}$ enough to predict all of them for a Gaussian random field



https://github.com/yipihey/kNN-CDFs

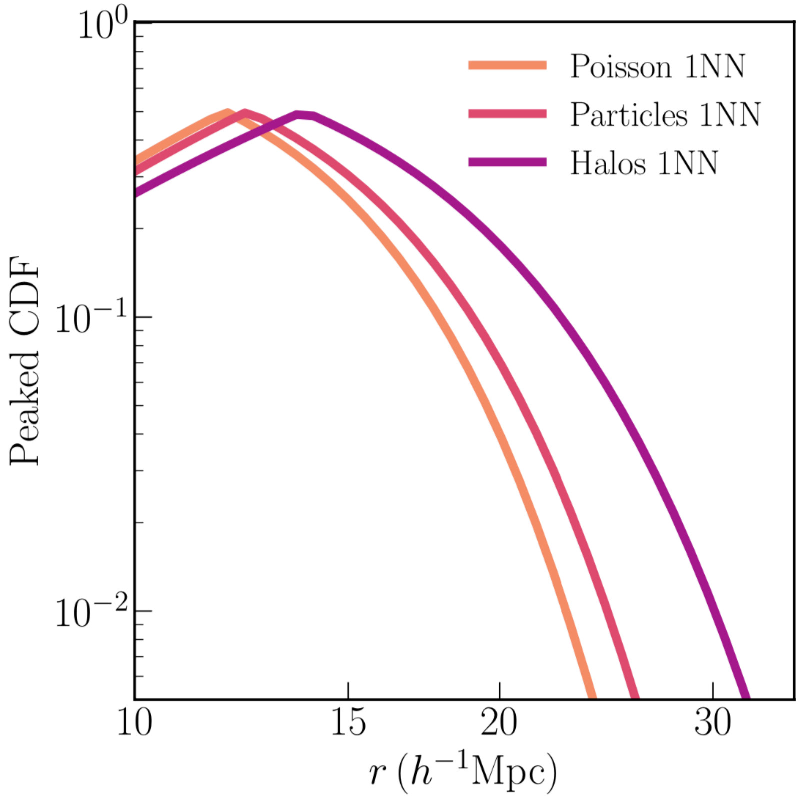


Figure 4. Comparison of the Peaked CDF for nearest neighbor (1NN) distributions of *a)* a Poisson distribution, *b)* particles from the $z = 0$ snapshot of an N-body simulation, and *c)* the most massive halos from the same simulation. In each case, 10^5 points were selected over a $(1h^{-1}\mathrm{Gpc})^3$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

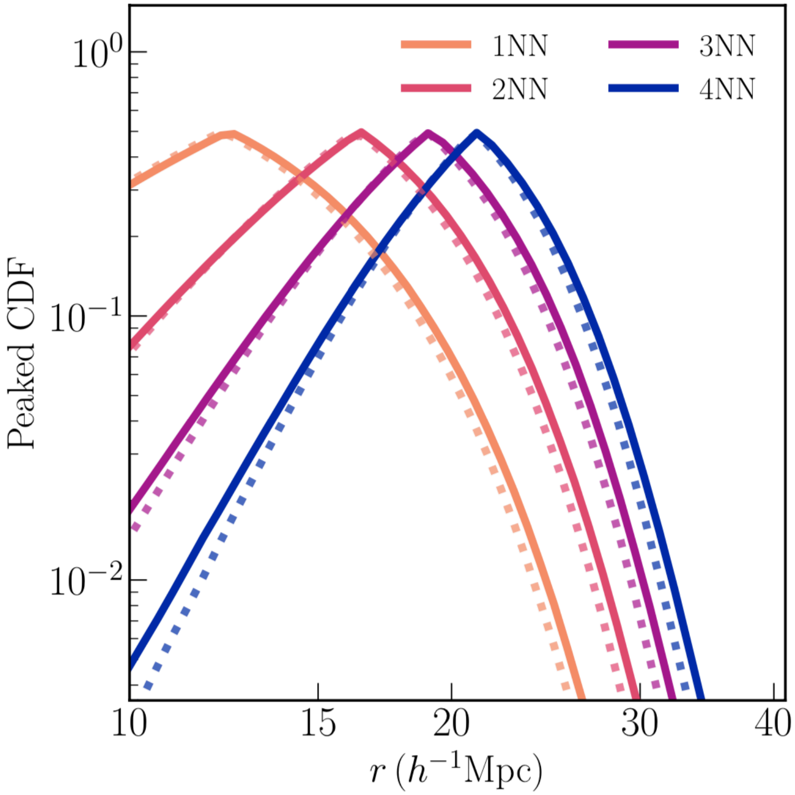


Figure 5. The Peaked CDFs for the first, second, third, and fourth nearest-neighbor distributions for 10^5 simulation particles in a $(1h^{-1}\mathrm{Gpc})^3$ volume. The solid lines represent these distributions at $z = 0$, while the dotted lines represent the distributions computed at $z = 0.5$.

How SENSITIVE ARE kNN-CDFs compared with ξ(R)?

Significantly MORE!

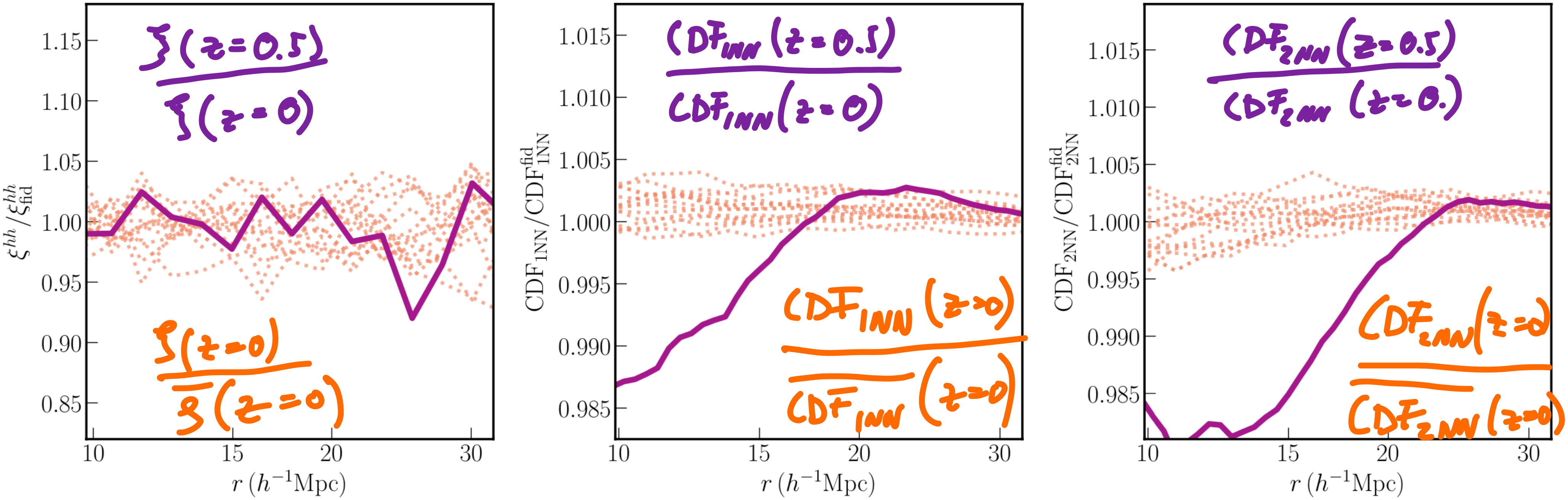


Figure 6. *Left*: The darker line represents the ratio of correlation function of the 10^5 most massive halos in a $(1h^{-1}{\rm Gpc})^3$ box at redshifts $z = 0$, and $z = 0.5$. The lighter shaded lines represent the ratio of the correlation functions at $z = 0$ for 15 different realizations of the 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.

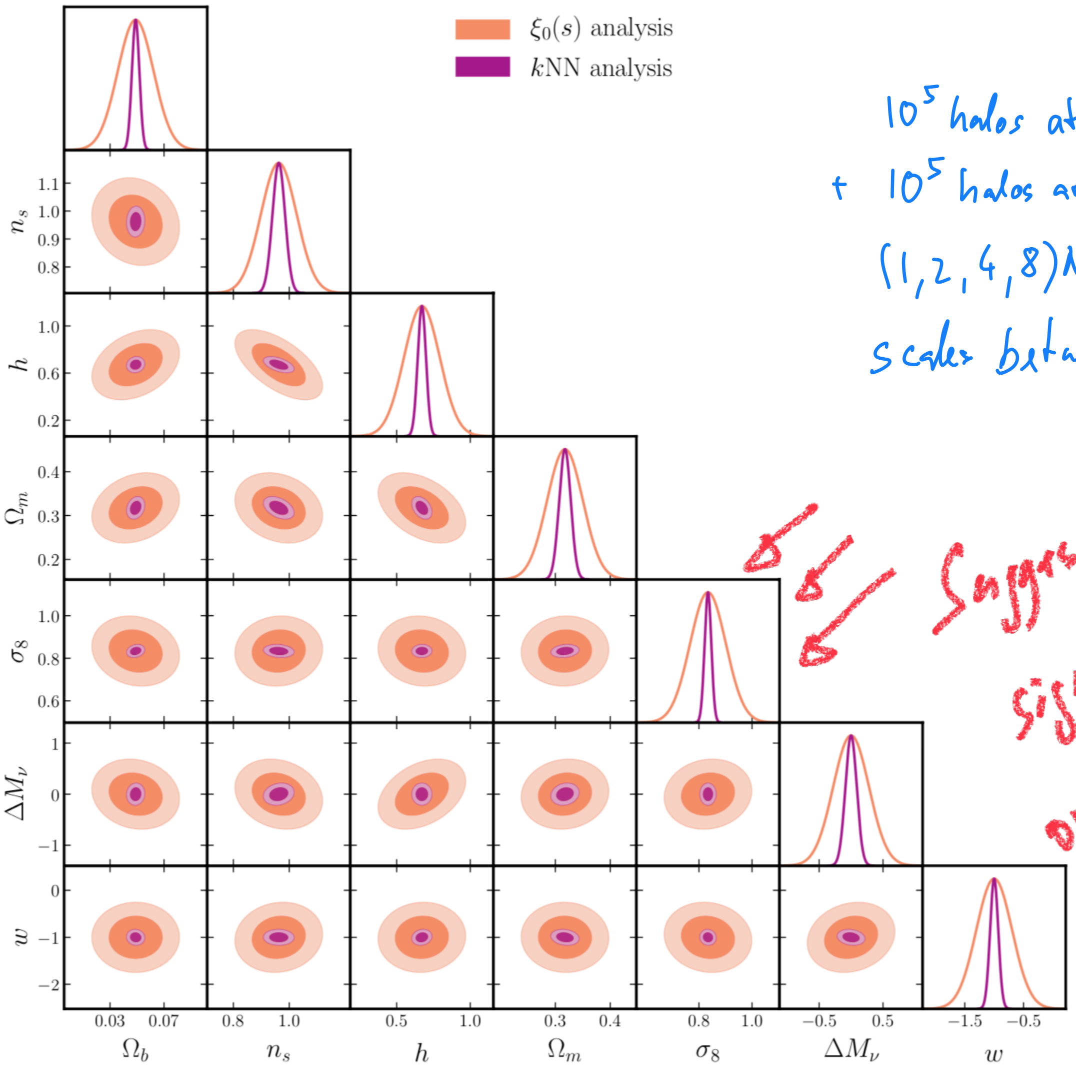


Figure 12. Constraints on the cosmological parameters derived from the Fisher analysis of the monopole of clustering of the 10^5 most massive halos in redshift space, combining information from $z = 0$ and $z = 0.5$, and using scales in the range $10h^{-1}$Mpc to $40h^{-1}$Mpc. Similar to Fig. 11, the constraints from the kNN analysis are much tighter than those from the $\xi(r)$ analysis.

10^5 halos at $z = 0.5$
+ 10^5 halos at $z = 0.5$
(1,2,4,8)NN-CDFs only for
scales between 10 and 40 $\frac{\text{Mpc}}{h}$

Suggests a very
significant improvement
over standard analysis
may be possible

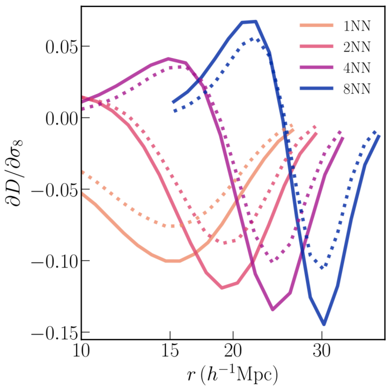


Figure 7. The derivative of the data vector with respect to the cosmological parameter σ_8. The different colored curves represent the portions of the data vectors coming from $k = \{1, 2, 4, 8\}$ nearest neighbor CDF distributions. The solid lines represent the derivative at $z = 0$, while the dotted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

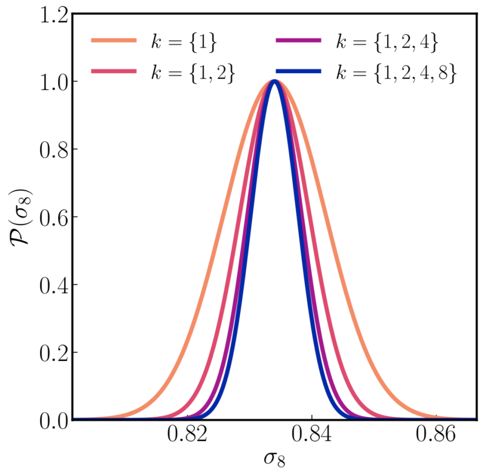
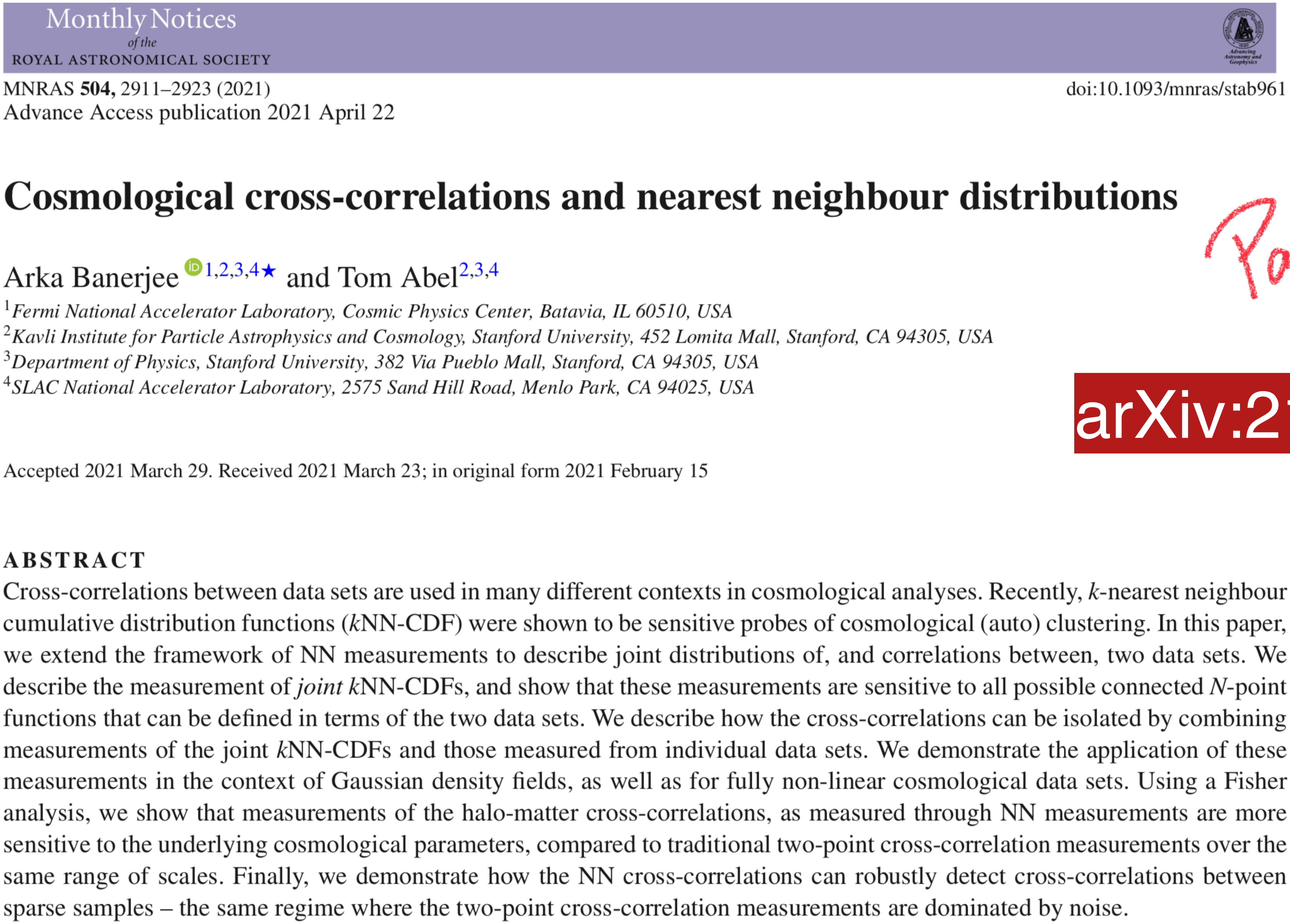


Figure 9. The posterior distribution for σ_8, marginalized over all other parameters. The different colors represent different kNN combinations from which the constraint was obtained. The constraints improve as more nearest neighbor distributions are added, but the gain saturates by the time we add all four CDFs that are computed from the data.



Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

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

Paper 2

Cosmological cross-correlations and nearest neighbour distributions

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

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

arXiv:2102.01184

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

ABSTRACT

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

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

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

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

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

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

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

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

UNCORRELATED CASE

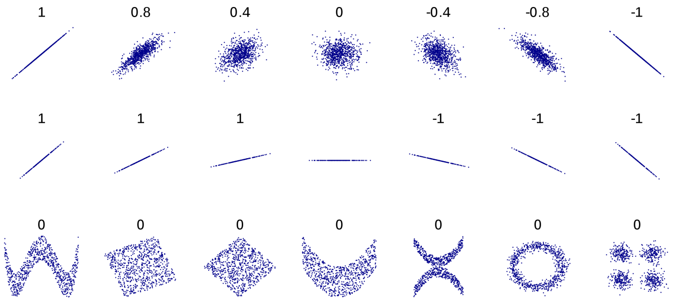
For each random point, we can associate two distances - one to the nearest neighbor data point from the first set, and the other to the nearest neighbor data point from the second set. Now, for every random point, we choose the larger of the two distances. These distances are then sorted to get the empirical Cumulative Distribution Function (CDF) of the distances chosen in this manner. We will refer to this distribution as the **joint Nearest Neighbor CDF**, $\mathrm{CDF}_{k1,k2}$

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

TESTS STATISTICAL DEPENDENCE

STRONGER TEST THAN FOR

"LINEAR" CORRELATION



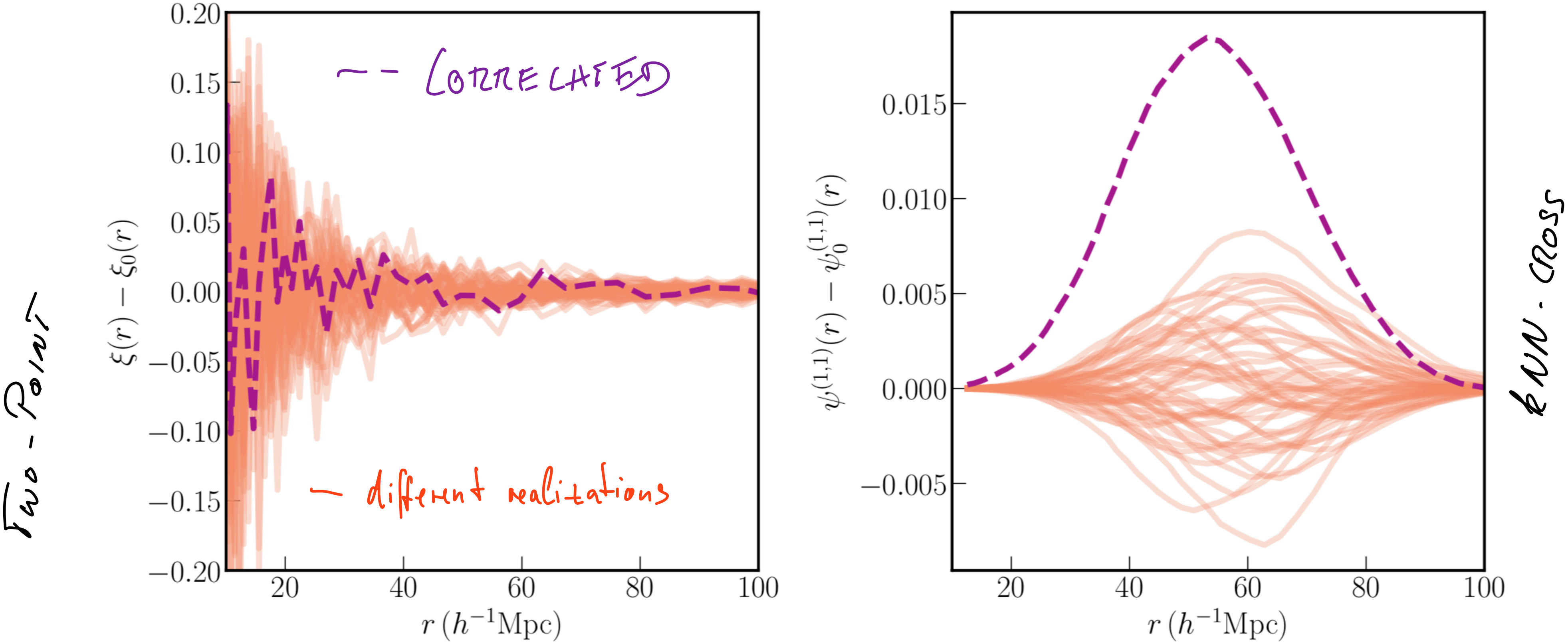


Figure 5. *Left panel:* Difference of the two-point cross correlation measurements of two sets of dark matter halos (1000 halos each) from the mean of 1000 such measurements where the two sets are spatially uncorrelated (drawn from different realizations). The lighter solid lines represent the difference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

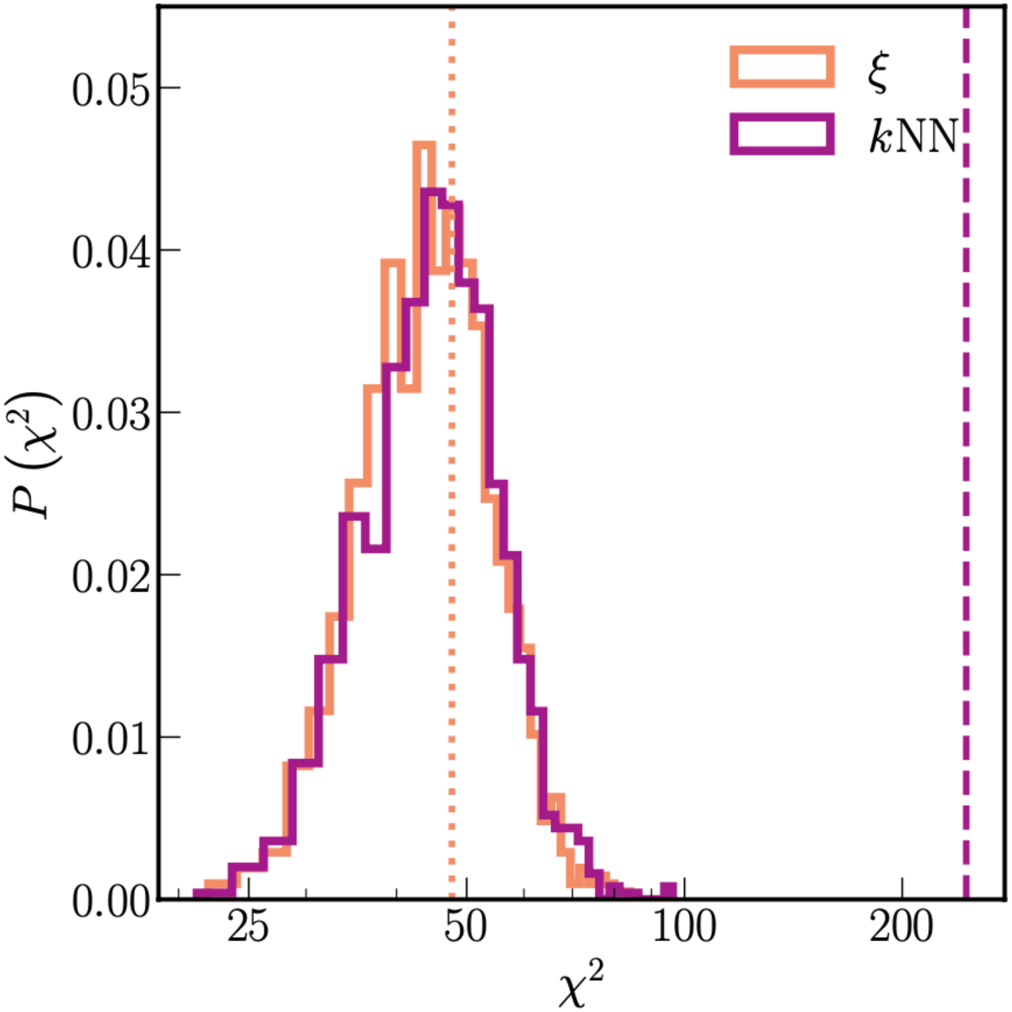


Figure 6. The solid lines represent the binned distribution of χ^2 values for 1000 measurements of cross-correlations between two samples of halos (1000 halos each from a $(1h^{-1}\mathrm{Gpc})^3$ volume) which are spatially uncorrelated. The lighter shaded line represents the distribution when cross-correlations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

Joint CDFs for Gaussian fields

Dotted lines: Analytic predictions

Solid lines: Measurements in sims

Dashed: 1NN and 2NN of individual fields

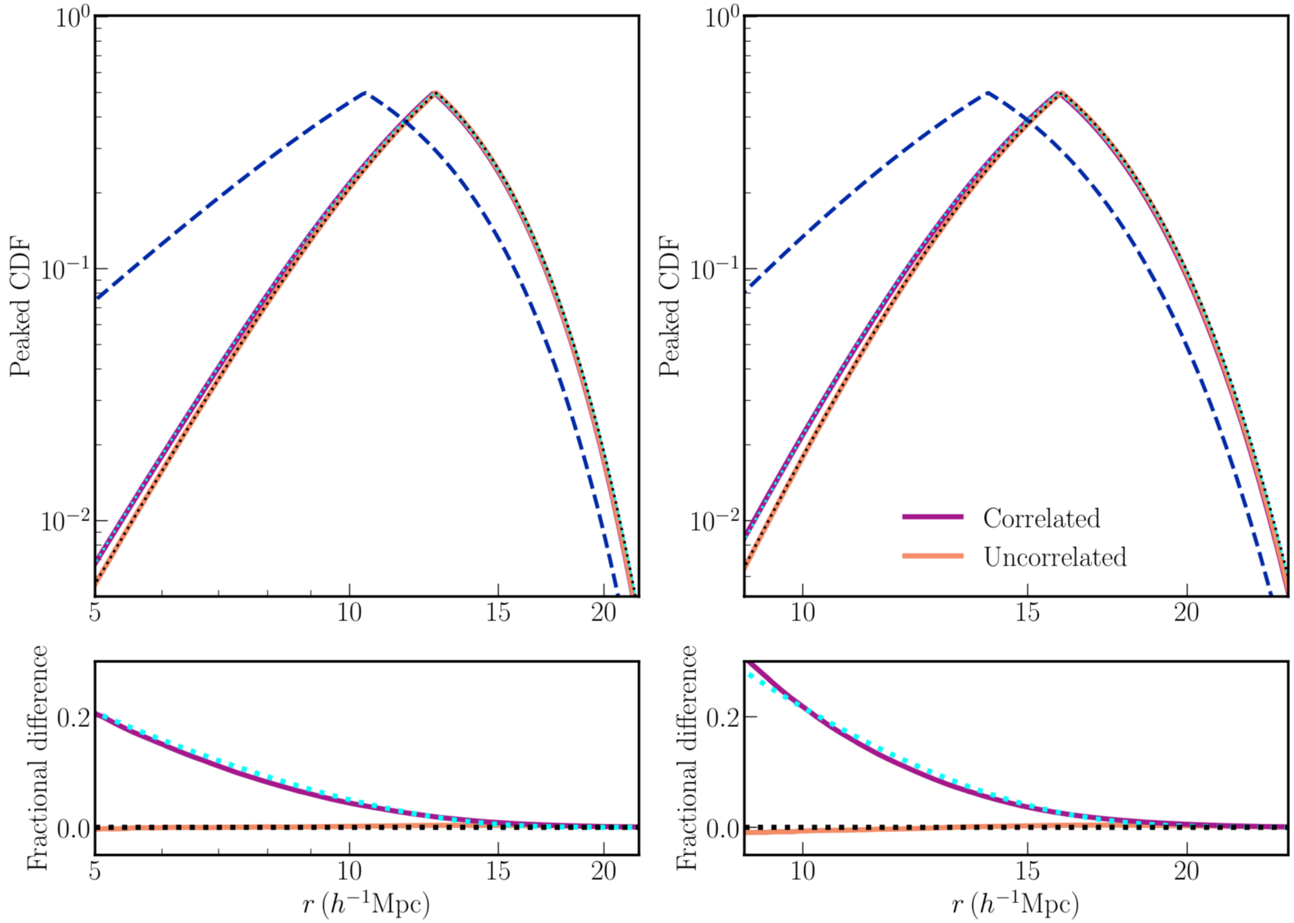


Figure 1. *Top panels*: Solid lines represent the peaked CDF of the *joint* 1NN (top-left panel) and *joint* 2NN (top-right panel) distributions for two correlated (darker lines) and two uncorrelated (lighter lines) sets, each composed 2×10^5 tracers of a Gaussian random field over a $(1h^{-1}\mathrm{Gpc})^3$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

NONLINEAR FIELD

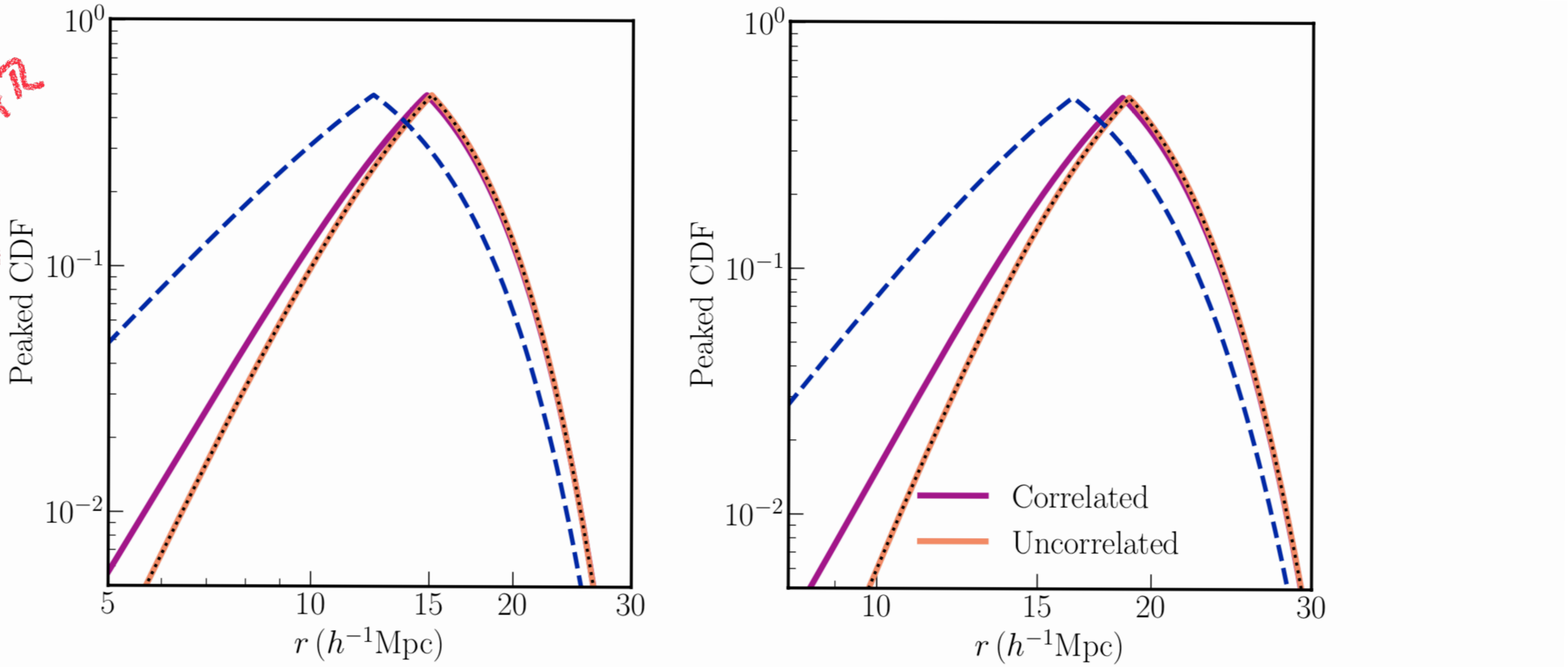


Figure 2. Solid lines represent the peaked CDF of the *joint* 1NN (left panel) and *joint* 2NN (right panel) distributions for two correlated (darker lines) and two uncorrelated (lighter lines) sets of simulation particles at $z = 0$, when the matter field is highly nonlinear on small scales. Each set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

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

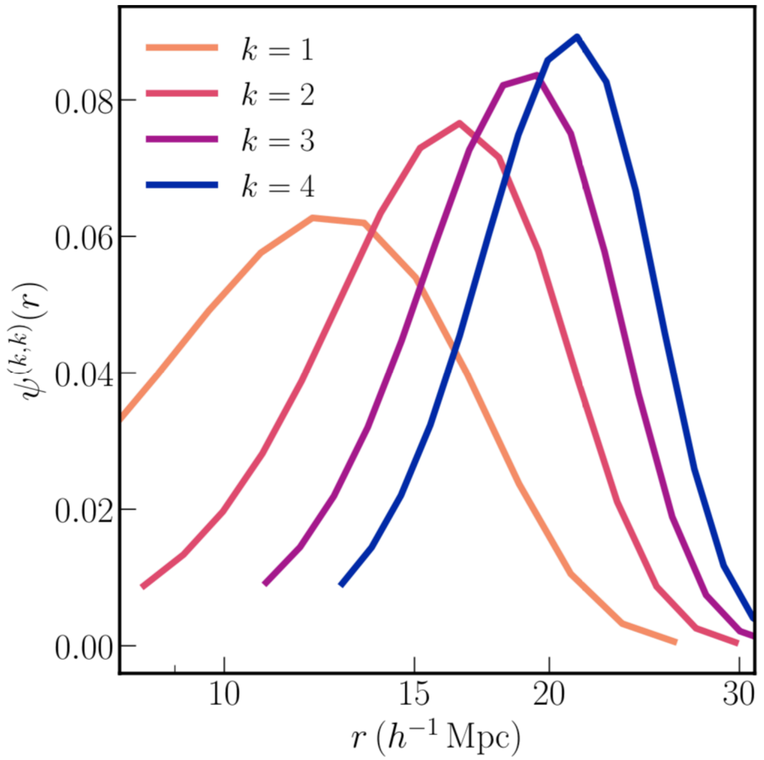


Figure 3. $\psi^{(k,k)}(r)$ (see Eq. 9), which measures the spatial correlation between two samples, for various k measured from the 10^5 most massive halos and 10^5 randomly chosen particles from a $(1h^{-1}\mathrm{Gpc})^3$ simulation at $z = 0$. These measurements are used in the Fisher matrix calculations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

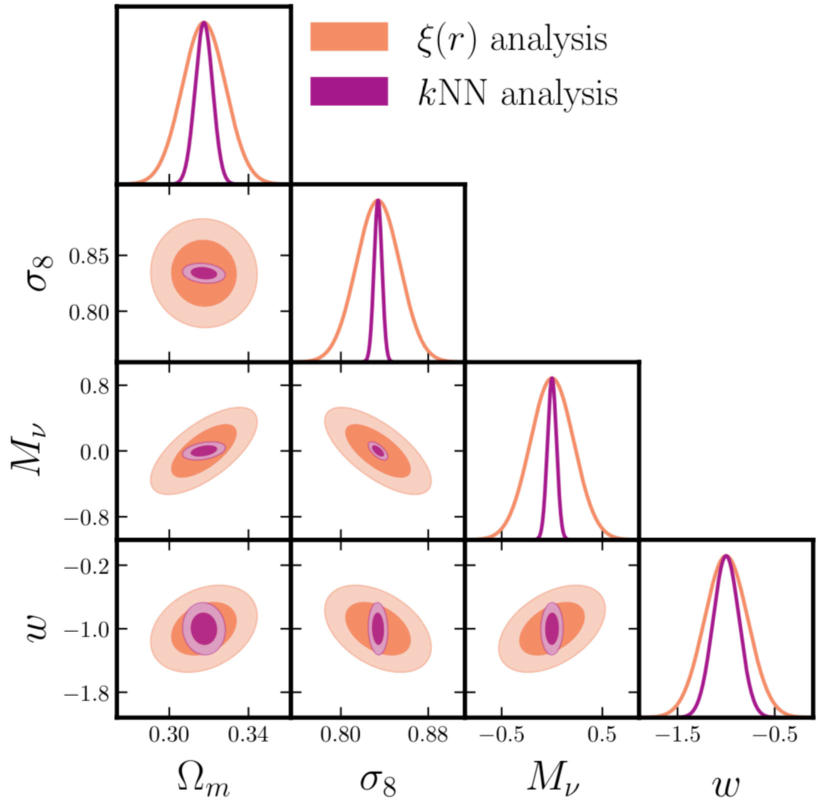
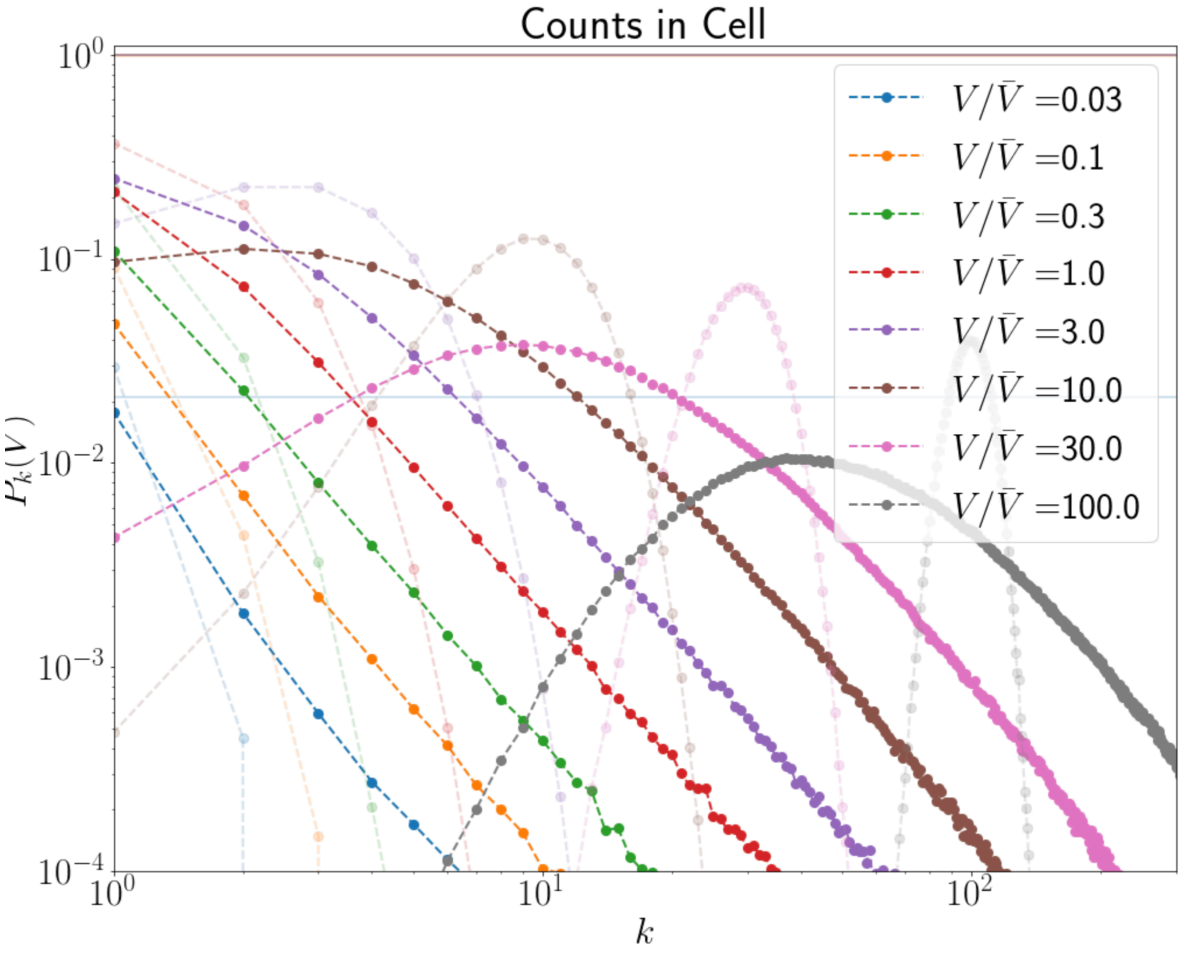


Figure 4. Constraints on various cosmological parameters from the Fisher analysis in Sec. 4.1. These are obtained from the cross-correlations of the 10^5 most massive halos in the simulation volume ($(1h^{-1}\mathrm{Gpc})^3$) with the matter distribution. There is a marked improvement in the constraints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

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

Predictability

- Analytic predictions by Uhlemann, Codis, et al, Cataneo et al (2016,2017, -) using extremal value statistics
- Use analytic models of the PDF such as Lam & Sheth 2008ab, 2010, Klypin, Prada et al 2018
- Linear theory (some of it in paper I)
- Spherical collapse modeling (Banerjee & Abel in prep.)
- Build emulators from simulations (McLaughlin et al 2021, Yuan, Abel et al in prep.)



https://github.com/yipihey/kNN-CDFs **https://github.com/OliverFHD/CosMomentum** **https://github.com/mcataneo/pyLDT-cosmo**

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

Predictability

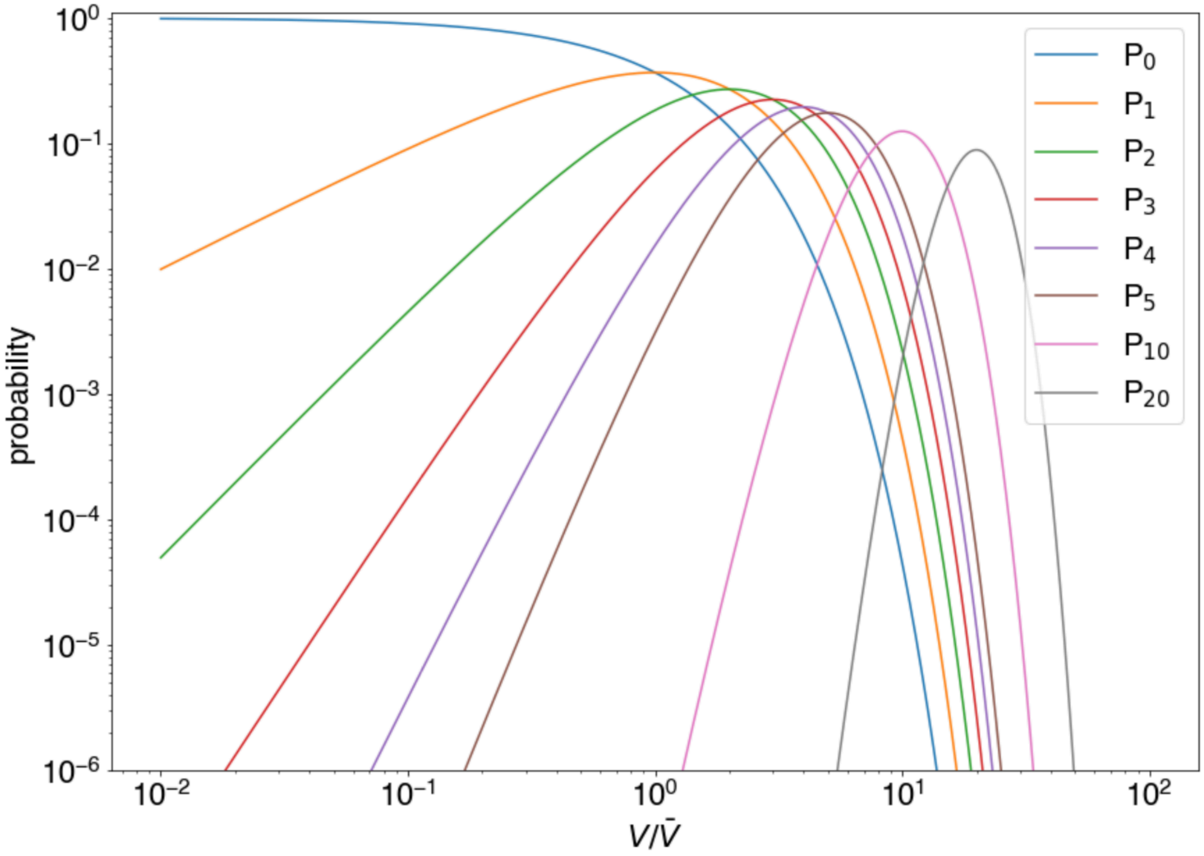
- Linear theory (some of it in paper I) for Gaussian field implemented in kNN_analytic.py in https://github.com/yipihey/kNN-CDFs

$P_{>k|V}$ for a general value of k, the individual terms are easy to compute, especially for low values of k. For example,

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

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

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,



https://github.com/yipihey/kNN-CDFs

**“All theories are legitimate, no matter.
What matters is what you do with them.”**

–Jorge Luis Borges

Self-Similarity of k-Nearest Neighbor Distributions in Scale-Free Simulations

Lehman H. Garrison,[1] Tom Abel,[2,3,4] and Daniel J. Eisenstein[5]

[1] *Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*
[2] *Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA*
[3] *Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[4] *SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*
[5] *Center for Astrophysics | Harvard & Smithsonian 60 Garden Street, Cambridge, MA 02138, USA*

ABSTRACT

We use the k-nearest neighbor probability distribution function (kNN-PDF, Banerjee & Abel 2021) to assess convergence in a scale-free N-body simulation. Compared to our previous two-point analysis, the kNN-PDF allows us to quantify our results in the language of halos and numbers of particles, while also incorporating non-Gaussian information. We find good convergence for 32 particles and greater at densities typical of halos, while 16 particles and fewer appears unconverged. Halving the softening length extends convergence to higher densities, but not to fewer particles. Our analysis is less sensitive to voids, but we analyze a limited range of underdensities and find evidence for convergence at 16 particles and greater even in sparse voids.

MNRAS **511,** 2765–2781 (2022) https://doi.org/10.1093/mnras/stac193
Advance Access publication 2022 January 27

Modelling nearest neighbour distributions of biased tracers using hybrid effective field theory

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

Accepted 2022 January 18. Received 2022 January 14; in original form 2021 August 9

ABSTRACT

We investigate the application of hybrid effective field theory (HEFT) – which combines a Lagrangian bias expansion with subsequent particle dynamics from *N*-body simulations – to the modelling of *k*-nearest neighbour cumulative distribution functions (*k*NN-CDFs) of biased tracers of the cosmological matter field. The *k*NN-CDFs are sensitive to all higher order connected *N*-point functions in the data, but are computationally cheap to compute. We develop the formalism to predict the *k*NN-CDFs of discrete tracers of a continuous field from the statistics of the continuous field itself. Using this formalism, we demonstrate how *k*NN-CDF statistics of a set of biased tracers, such as haloes or galaxies, of the cosmological matter field can be modelled given a set of low-redshift HEFT component fields and bias parameter values. These are the same ingredients needed to predict the two-point clustering. For a specific sample of haloes, we show that both the two-point clustering *and* the *k*NN-CDFs can be well-fit on quasi-linear scales ($\gtrsim 20h^{-1}$Mpc) by the second-order HEFT formalism with the *same values* of the bias parameters, implying that joint modelling of the two is possible. Finally, using a Fisher matrix analysis, we show that including *k*NN-CDF measurements over the range of allowed scales in the HEFT framework can improve the constraints on σ_8 by roughly a factor of 3, compared to the case where only two-point measurements are considered. Combining the statistical power of *k*NN measurements with the modelling power of HEFT, therefore, represents an exciting prospect for extracting greater information from small-scale cosmological clustering.

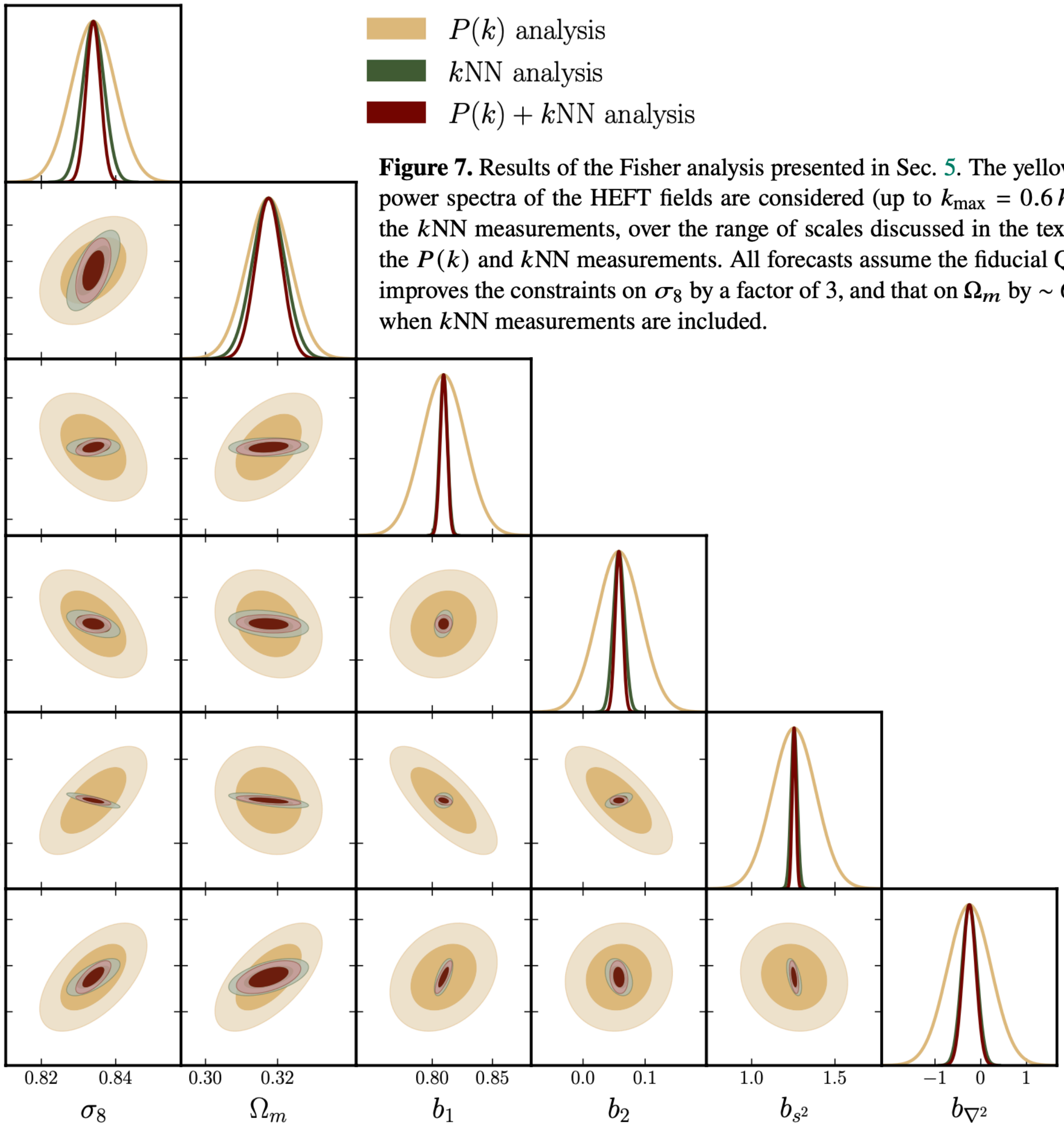


Figure 7. Results of the Fisher analysis presented in Sec. 5. The yellow contours and standard Fisher errors represent the results when only the auto and cross power spectra of the HEFT fields are considered (up to $k_{\rm max} = 0.6\,h{\rm Mpc}^{-1}$). The green contours represent the standard Fisher errors from the analysis of the kNN measurements, over the range of scales discussed in the text. The marooon contours and standard Fisher errors represent the results of combining the $P(k)$ and kNN measurements. All forecasts assume the fiducial Quijote volume of $V = 1(h^{-1}{\rm Gpc})^3$. For this sample, adding in the kNN measurements improves the constraints on σ_8 by a factor of 3, and that on Ω_m by $\sim 60\%$ over the $P(k)$-only analysis. The HEFT bias parameters are also better constrained when kNN measurements are included.

The Lagrangian bias models that are central to HEFT assume that at the initial, Lagrangian, coordinates $\boldsymbol{q}$, the tracer density is related to the large-scale dark matter operators by a linear combination of all contributions allowed by Newtonian symmetries (Vlah et al. 2016)

$$\begin{aligned} \delta_X(\boldsymbol{q}) &= F[\delta(\boldsymbol{q}), s_{ij}(\boldsymbol{q})] \qquad (20)\\ &\approx 1 + b_1\delta(\boldsymbol{q}) + b_2\left(\delta^2(\boldsymbol{q}) - \langle\delta^2\rangle\right) + \\ &\qquad b_{s^2}\left(s^2(\boldsymbol{q}) - \langle s^2\rangle\right) + b_{\nabla^2}\nabla^2\delta(\boldsymbol{q}) + \cdots \\ &\quad + \epsilon(\boldsymbol{q}). \end{aligned}$$

where we have expanded the functional F to second order. Additionally, we have defined the *traceless tidal tensor field* $s_{ij}(\boldsymbol{q}) = \left(\frac{\partial_i\partial_j}{\partial^2} - \frac{1}{3}\delta_{ij}\right)\delta(\boldsymbol{q})$. The field $\epsilon(\boldsymbol{q})$ describes the *stochastic* contribution to the expansion. The stochastic term generally describes the fact the bias expansion describes a statistical relationship, and there can be deviations from this in any given realization. The stochastic fields are, by construction, uncorrelated with the bias operators $\langle\epsilon(\boldsymbol{q})\mathcal{O}_i\rangle = 0$. In the limit of $\epsilon(\boldsymbol{q})$ being distributed as Gaussian white noise, its auto-correlation is constant even after advection to late-times.

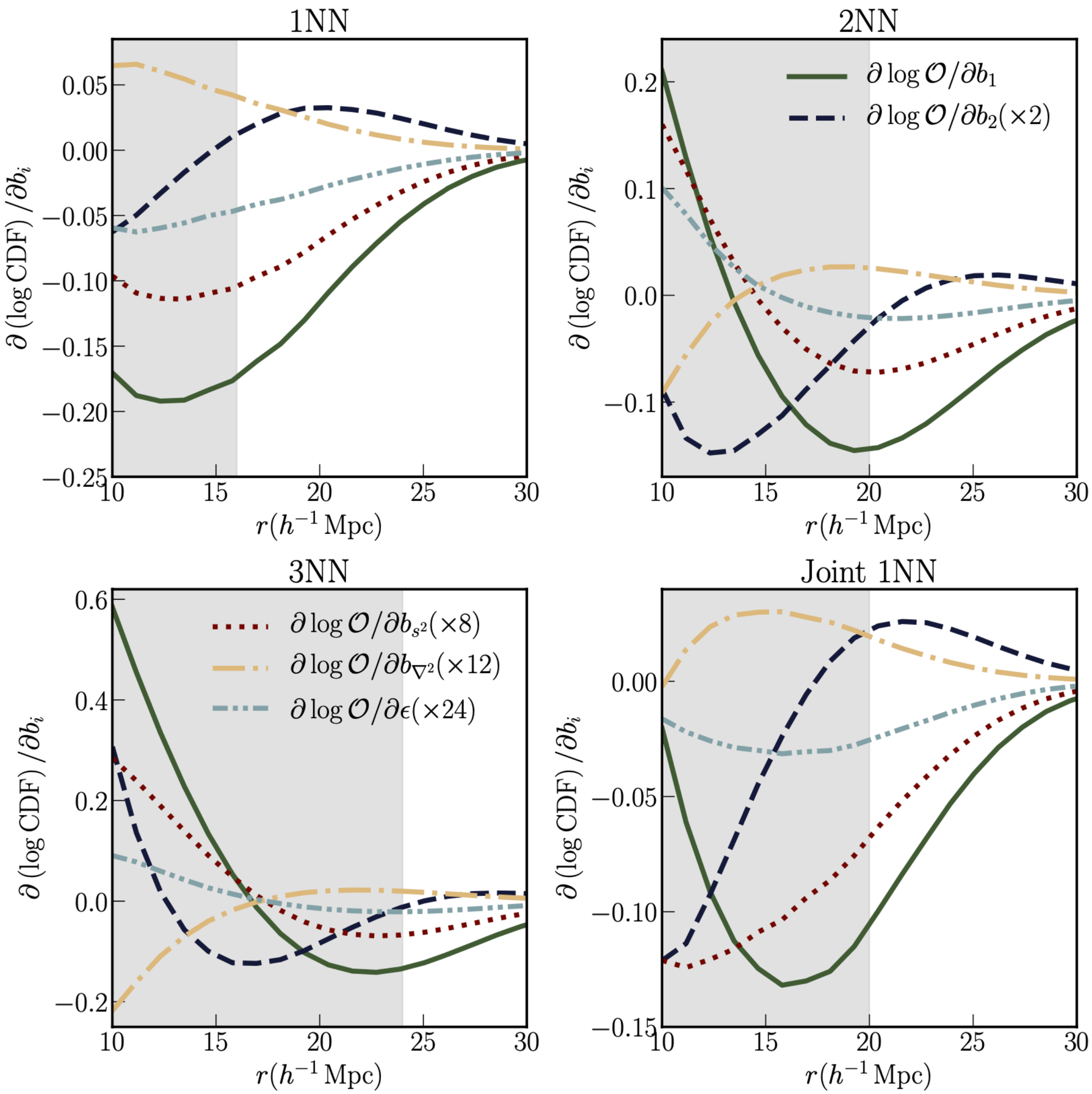


Figure 4. Logarithmic derivatives (around the best-fit values) of different *k*NN-CDFs with respect to the second order Lagrangian bias parameters $b1$, $b2$, b_S2 and $b\nabla2$. Each panel shows the response of a given *k*NN distribution, while lines of the same color represent the same bias parameter across the different panels. The grey shaded regions in each panel represent the scales where the HEFT predictions cannot accurately model the measurements. Similar to the 2-point function, the predictions are most sensitive to changes in $b1$.

Muir, Banerjee, Adikhari, Abel et al in prep

Comparing a number of Modified gravity HOD catalogs with matching two point functions with kNN-CDFs shows their potential to break degeneracies.

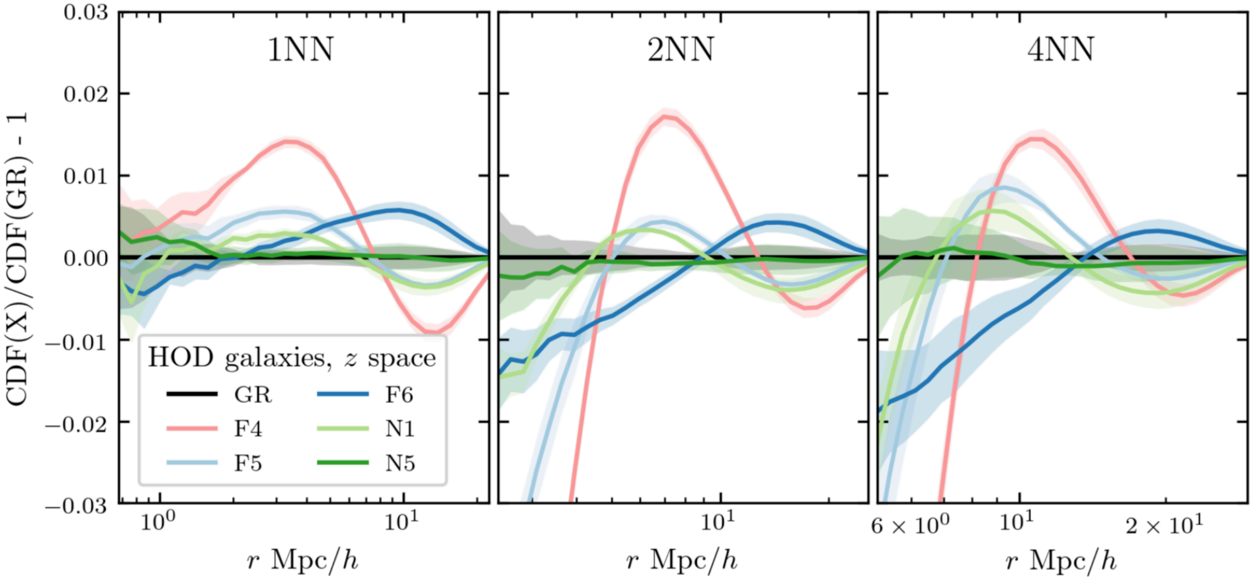
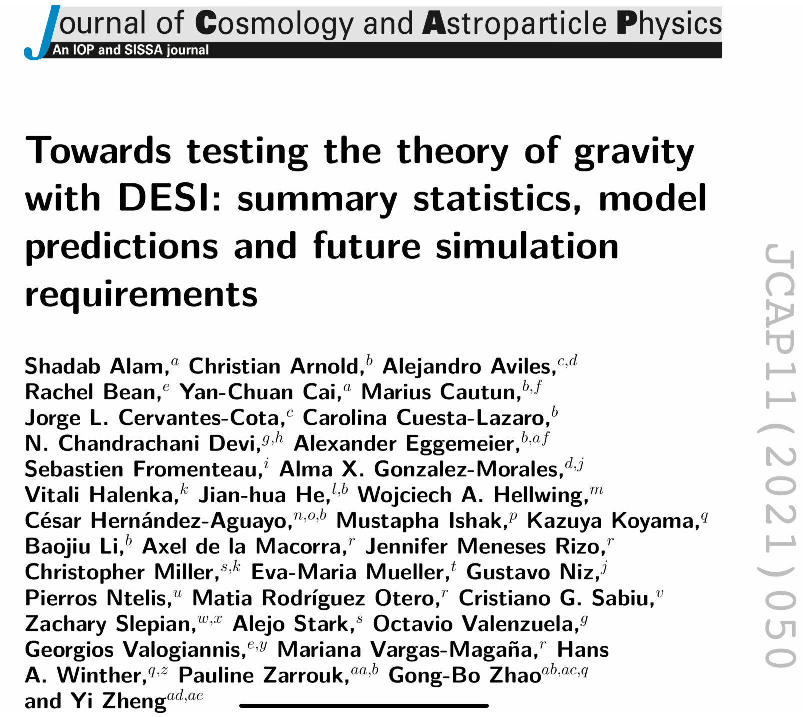


Figure 2. Fractional differences between the kNN CDFS (non-peaked) measured in redshift space between the fiducial MG and GR HOD catalogs. Liness show the average of five realizations for each MG simulated compared to the average of the five GR boxes, and the shaded regions show the standard deviation between the five realizations (also relative to the GR average).

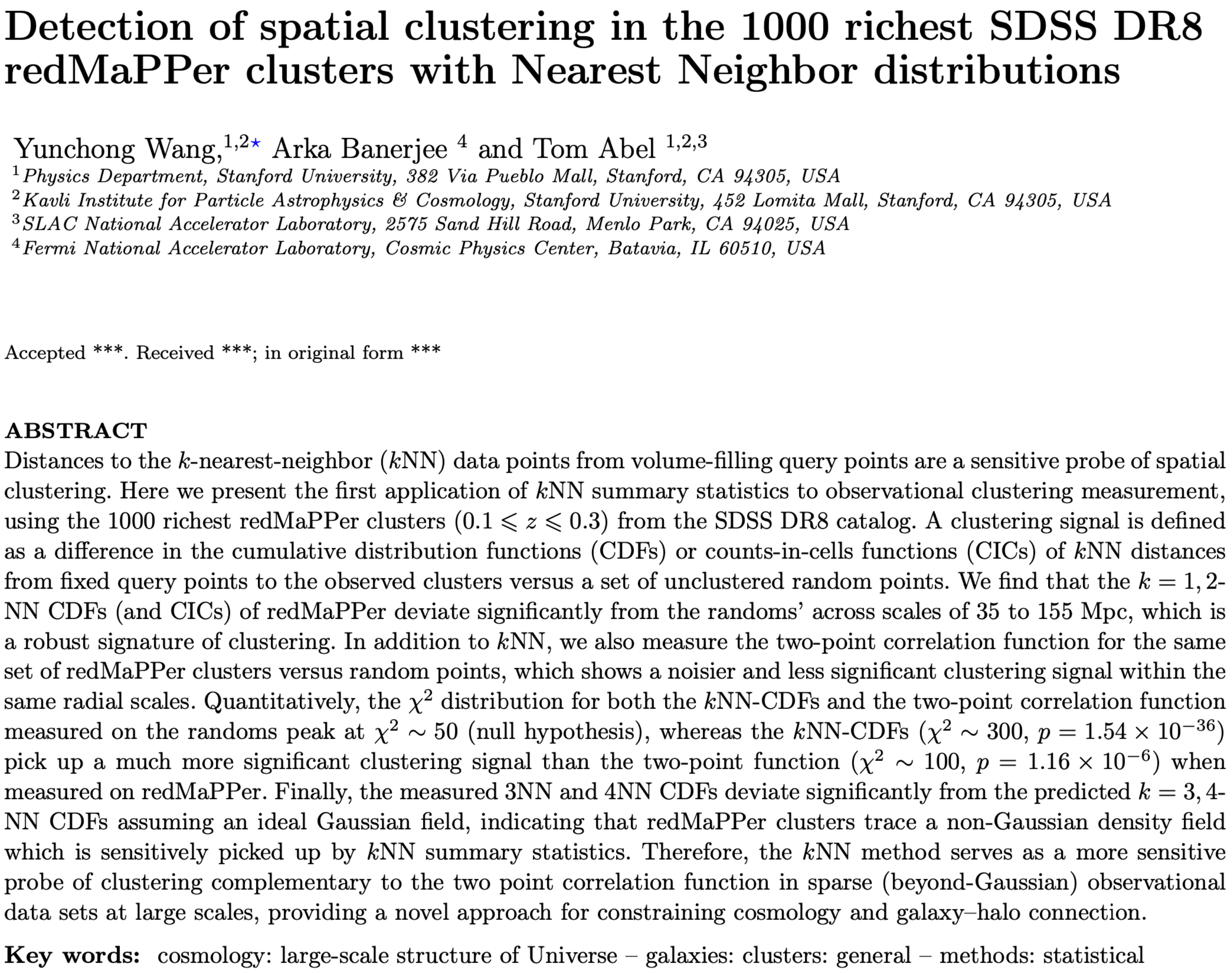

Journal of Cosmology and Astroparticle Physics
An IOP and SISSA journal

Towards testing the theory of gravity with DESI: summary statistics, model predictions and future simulation requirements

Shadab Alam,[a] Christian Arnold,[b] Alejandro Aviles,[c,d] Rachel Bean,[e] Yan-Chuan Cai,[a] Marius Cautun,[b,f] Jorge L. Cervantes-Cota,[c] Carolina Cuesta-Lazaro,[b] N. Chandrachani Devi,[g,h] Alexander Eggemeier,[b,af] Sebastien Fromenteau,[i] Alma X. Gonzalez-Morales,[d,j] Vitali Halenka,[k] Jian-hua He,[l,b] Wojciech A. Hellwing,[m] César Hernández-Aguayo,[n,o,b] Mustapha Ishak,[p] Kazuya Koyama,[q] Baojiu Li,[b] Axel de la Macorra,[r] Jennifer Meneses Rizo,[r] Christopher Miller,[s,k] Eva-Maria Mueller,[t] Gustavo Niz,[j] Pierros Ntelis,[u] Matia Rodríguez Otero,[r] Cristiano G. Sabiu,[v] Zachary Slepian,[w,x] Alejo Stark,[s] Octavio Valenzuela,[g] Georgios Valogiannis,[e,y] Mariana Vargas-Magaña,[r] Hans A. Winther,[q,z] Pauline Zarrouk,[aa,b] Gong-Bo Zhao[ab,ac,q] and Yi Zheng[ad,ae]

JCAP11(2021)050

First Application
to observational Data



Detection of spatial clustering in the 1000 richest SDSS DR8 redMaPPer clusters with Nearest Neighbor distributions

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

Accepted ***. Received ***; in original form ***

ABSTRACT

Distances to the k-nearest-neighbor (kNN) data points from volume-filling query points are a sensitive probe of spatial clustering. Here we present the first application of kNN summary statistics to observational clustering measurement, using the 1000 richest redMaPPer clusters ($0.1 \leqslant z \leqslant 0.3$) from the SDSS DR8 catalog. A clustering signal is defined as a difference in the cumulative distribution functions (CDFs) or counts-in-cells functions (CICs) of kNN distances from fixed query points to the observed clusters versus a set of unclustered random points. We find that the $k = 1, 2$-NN CDFs (and CICs) of redMaPPer deviate significantly from the randoms' across scales of 35 to 155 Mpc, which is a robust signature of clustering. In addition to kNN, we also measure the two-point correlation function for the same set of redMaPPer clusters versus random points, which shows a noisier and less significant clustering signal within the same radial scales. Quantitatively, the χ^2 distribution for both the kNN-CDFs and the two-point correlation function measured on the randoms peak at $\chi^2 \sim 50$ (null hypothesis), whereas the kNN-CDFs ($\chi^2 \sim 300$, $p = 1.54 \times 10^{-36}$) pick up a much more significant clustering signal than the two-point function ($\chi^2 \sim 100$, $p = 1.16 \times 10^{-6}$) when measured on redMaPPer. Finally, the measured 3NN and 4NN CDFs deviate significantly from the predicted $k = 3, 4$-NN CDFs assuming an ideal Gaussian field, indicating that redMaPPer clusters trace a non-Gaussian density field which is sensitively picked up by kNN summary statistics. Therefore, the kNN method serves as a more sensitive probe of clustering complementary to the two point correlation function in sparse (beyond-Gaussian) observational data sets at large scales, providing a novel approach for constraining cosmology and galaxy–halo connection.

Key words: cosmology: large-scale structure of Universe – galaxies: clusters: general – methods: statistical

astro-ph: 2112.04502

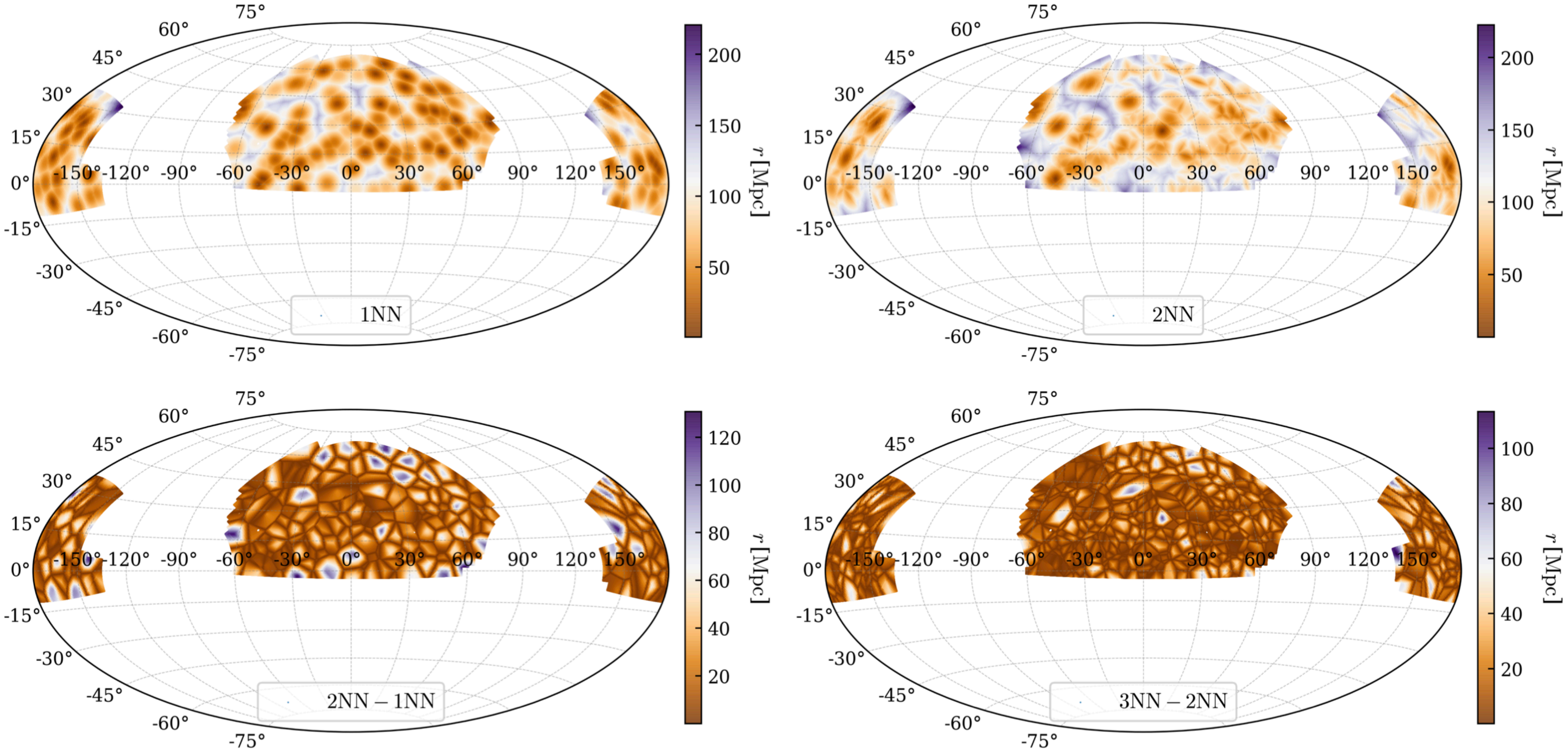


Figure D1. The kNN distances to the 1000 richest redMaPPer clusters measured from 1.3×10^8 query points located at $z = 0.2$ (842.87 Mpc). The query points are all healpy pixels of the SDSS survey footprint with $f_{\rm good} > 0.5$ and $z_{\rm max} \geqslant 0.3$ The color bars denote the kNN distances plotted in each panel. *Top left*: 1NN distance. *Top right*: 2NN distance. *Bottom left*: The difference of the distance to the 2NN and 1NN clusters for every query point. *Bottom right*: The difference of the distance to the 3NN and 2NN clusters for every query point.

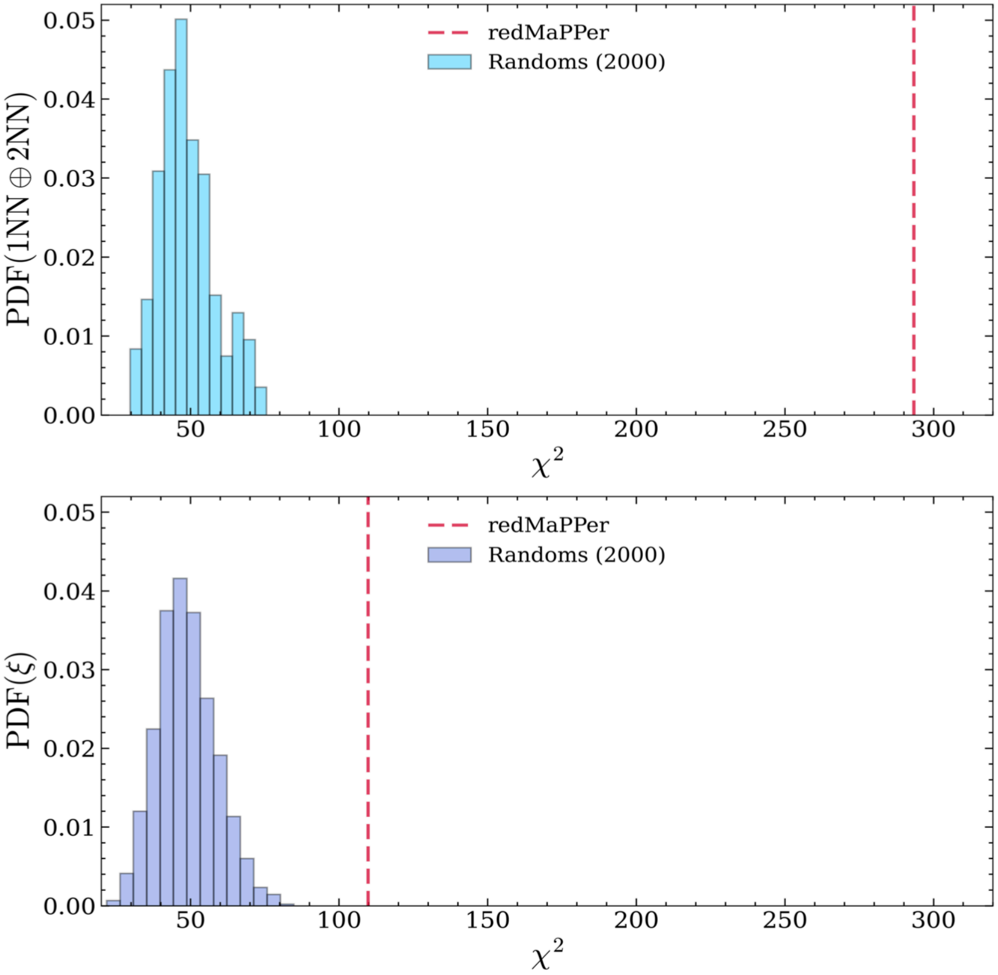
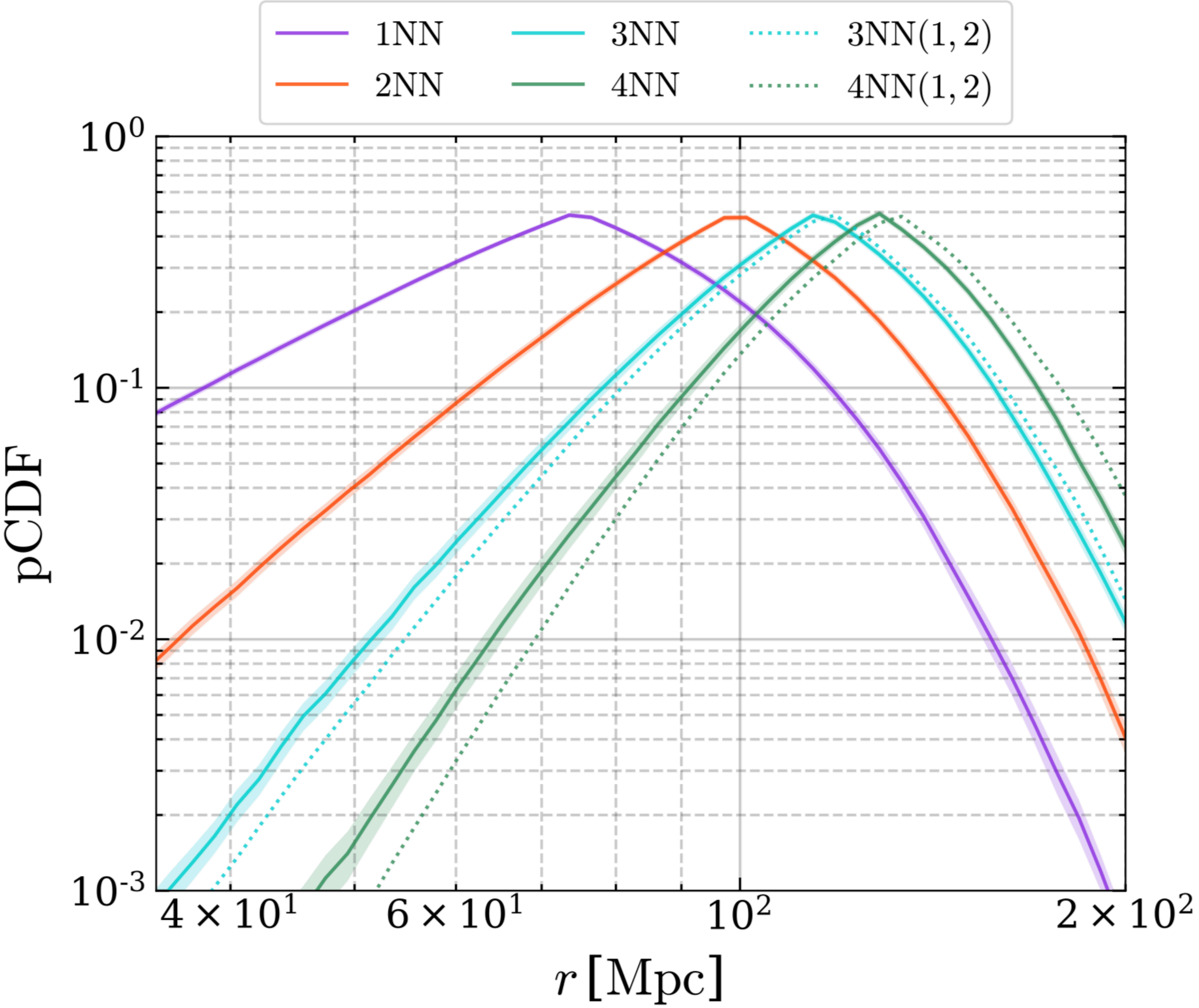


Figure 7. *Upper panel*: χ^2 distribution for the joint $k = 1, 2$ nearest neighbor CDFs, with the blue histograms showing the 2000 random samples and the dashed vertical red line showing the χ^2 of the 1000 richest redMaPPer clusters. *Lower panel*: χ^2 distribution for the two point correlation function $\xi(r)$, with the blue histograms showing the 2000 random samples and the dashed vertical red line denoting the χ^2 of the 1000 richest redMaPPer clusters. The χ^2 distribution for the randoms in the top and bottom panels are almost identical, serving as the null hypotheses for the clustering detection of the two summary statistics.



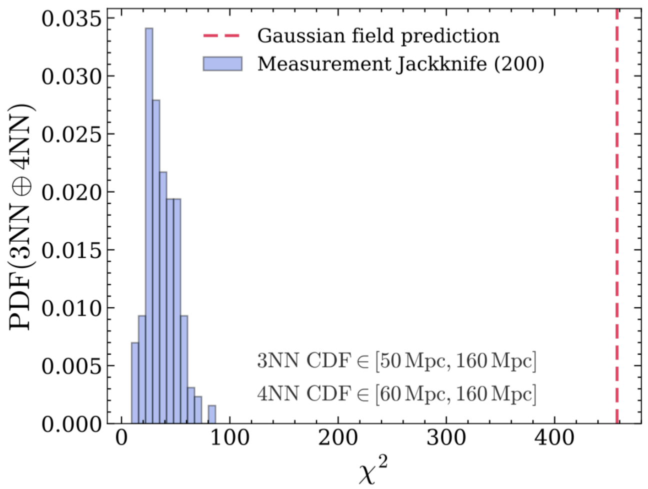


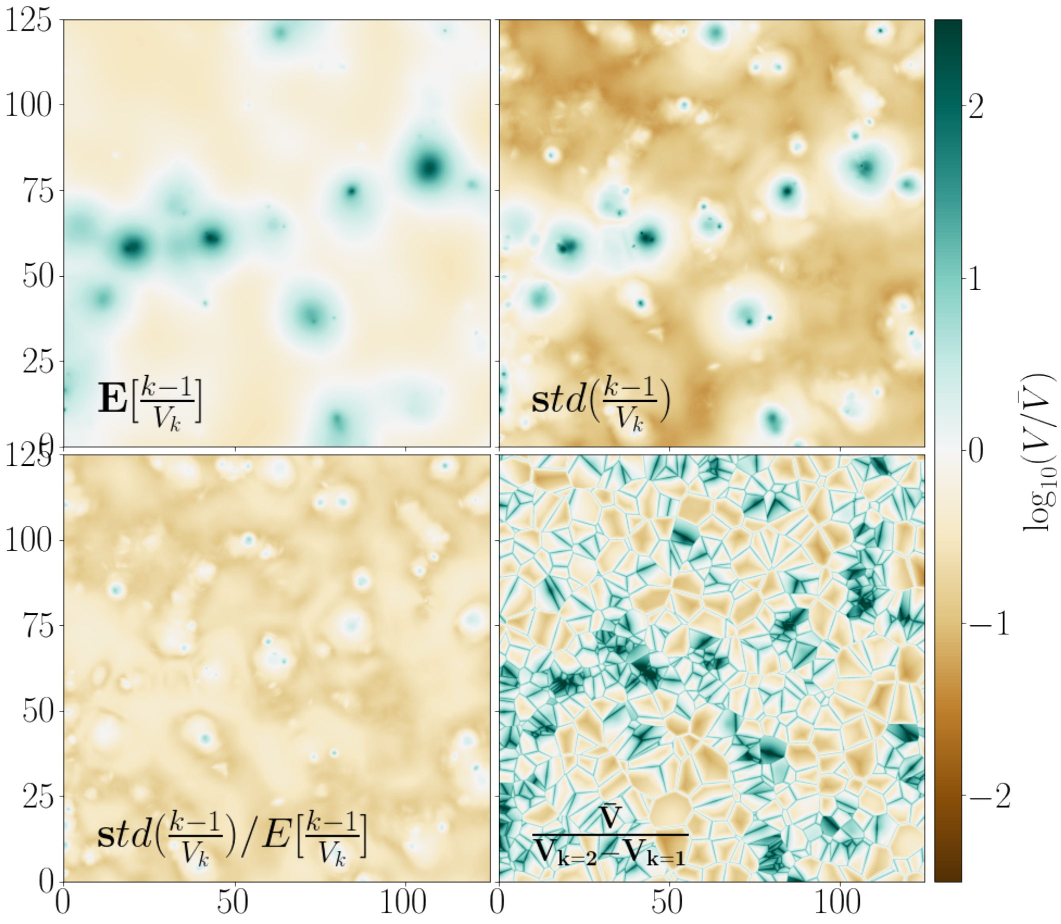
Figure 8. *Upper panel:* The measured $k = 1, 2, 3, 4$-NN CDFs for the 1000 richest redMaPPer clusters are denoted by the solid curves. The shaded regions around them indicate the 1σ Jackknife errors from 200 Jackknife samples. The dotted turquoise and green curves denote the predicted 3NN(1,2) and 4NN(1,2) CDFs assuming a random Gaussian field traced by the clusters, which are derived from the measured 1NN and 2NN CDFs propagated through equations 12 and 13. *Lower panel:* The χ^2 distribution of the joint $k = 3, 4$-NN CDFs. The blue histograms denote the χ^2 of 200 Jackknife it measurements of the 3NN and 4NN CDFs, while the dashed red line marks the χ^2 for the Gaussian field-*prediction* of the 3NN and 4NN CDFs. The predicted CDFs deviate significantly from the measurements, highlighting a robust detection of non-Gaussianity in the redMaPPer clusters.

Testing non-Gausianity

Connection to other statistics

Voronoi cell structure

Suggest map making technique

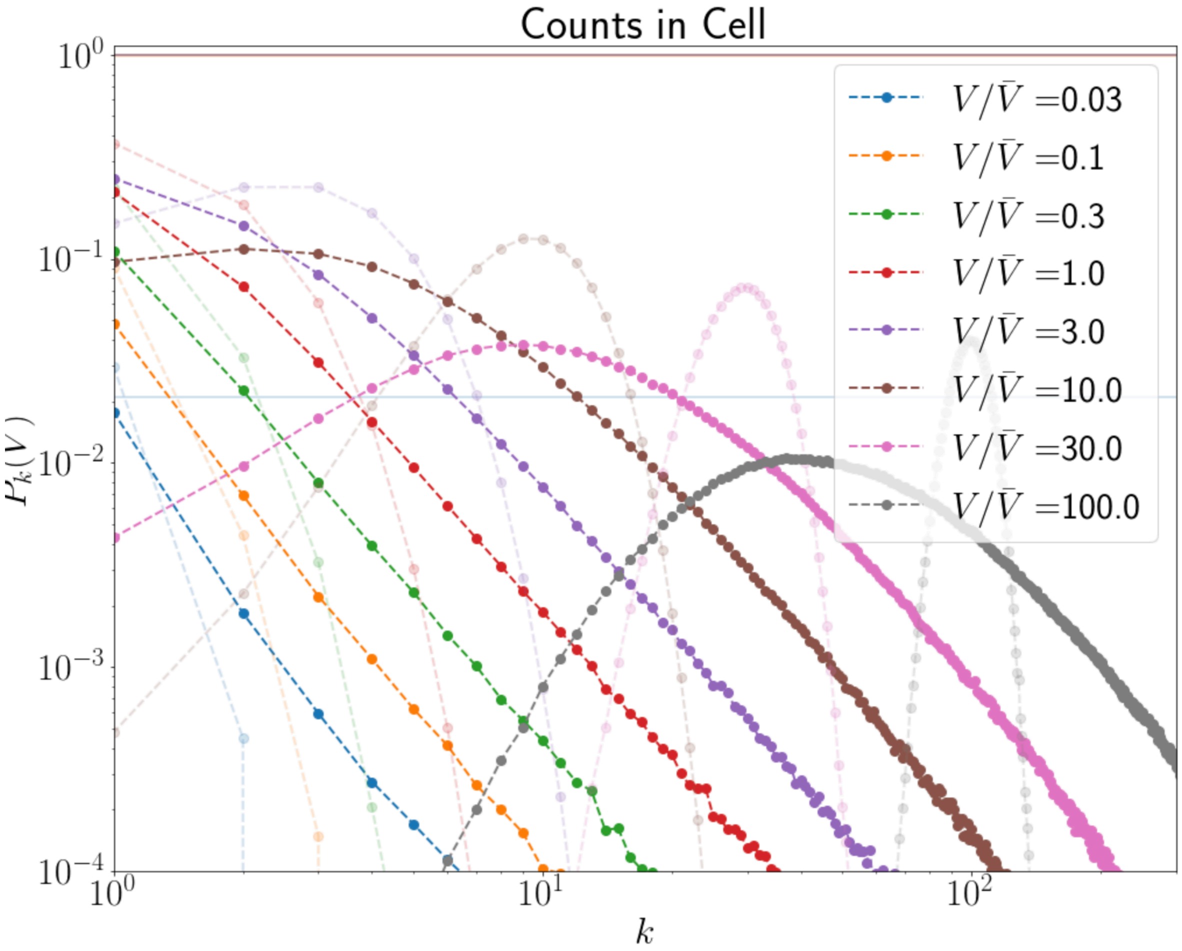


$$P_{k|V} = P_{>k-1|V} - P_{>k|V} \qquad \forall k \geq 1.$$

Counts in Cell/PDF/VPF

From kNN-CDF you can construct counts in cell for any cell volume.
Compare this with standard approach of throwing a large number of spheres at random position and counting how many galaxies are within them.

Void Probability Function (VPF) (White 1978) is $P_{0|V}$ is the special case restricted to empty spheres.

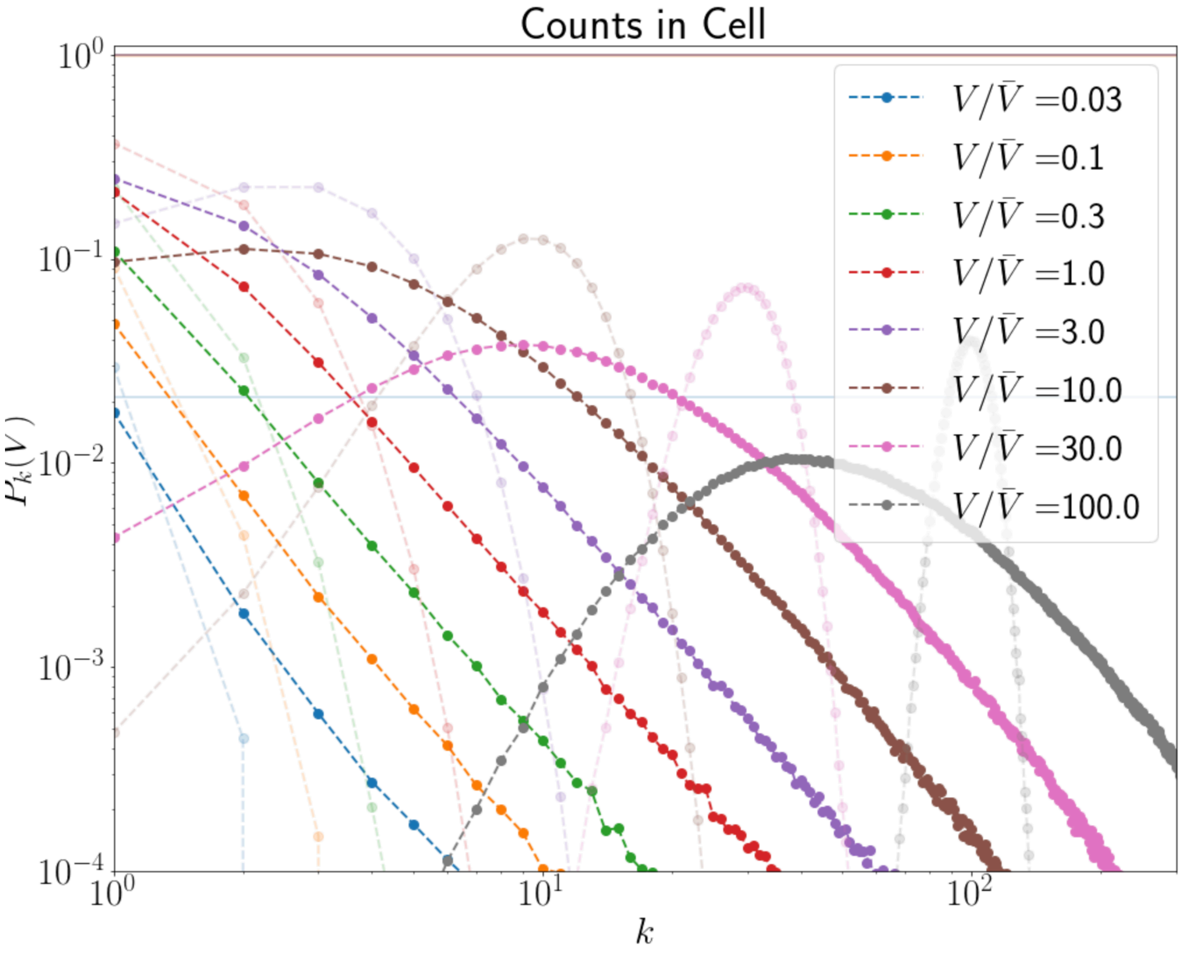


https://github.com/yipihey/kNN-CDFs https://github.com/OliverFHD/CosMomentum https://github.com/mcataneo/pyLDT-cosmo

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

Counts in Cylinders analogous using 2D kNN-CDFs

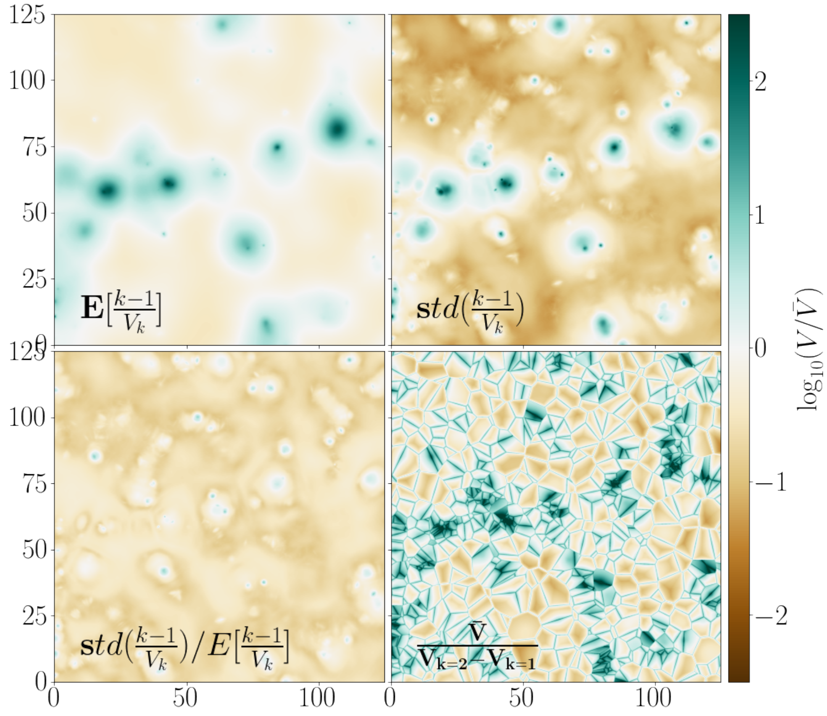
From two-dimensional kNN-CDF you can construct counts in cylinders for any disk area.



https://github.com/yipihey/kNN-CDFs

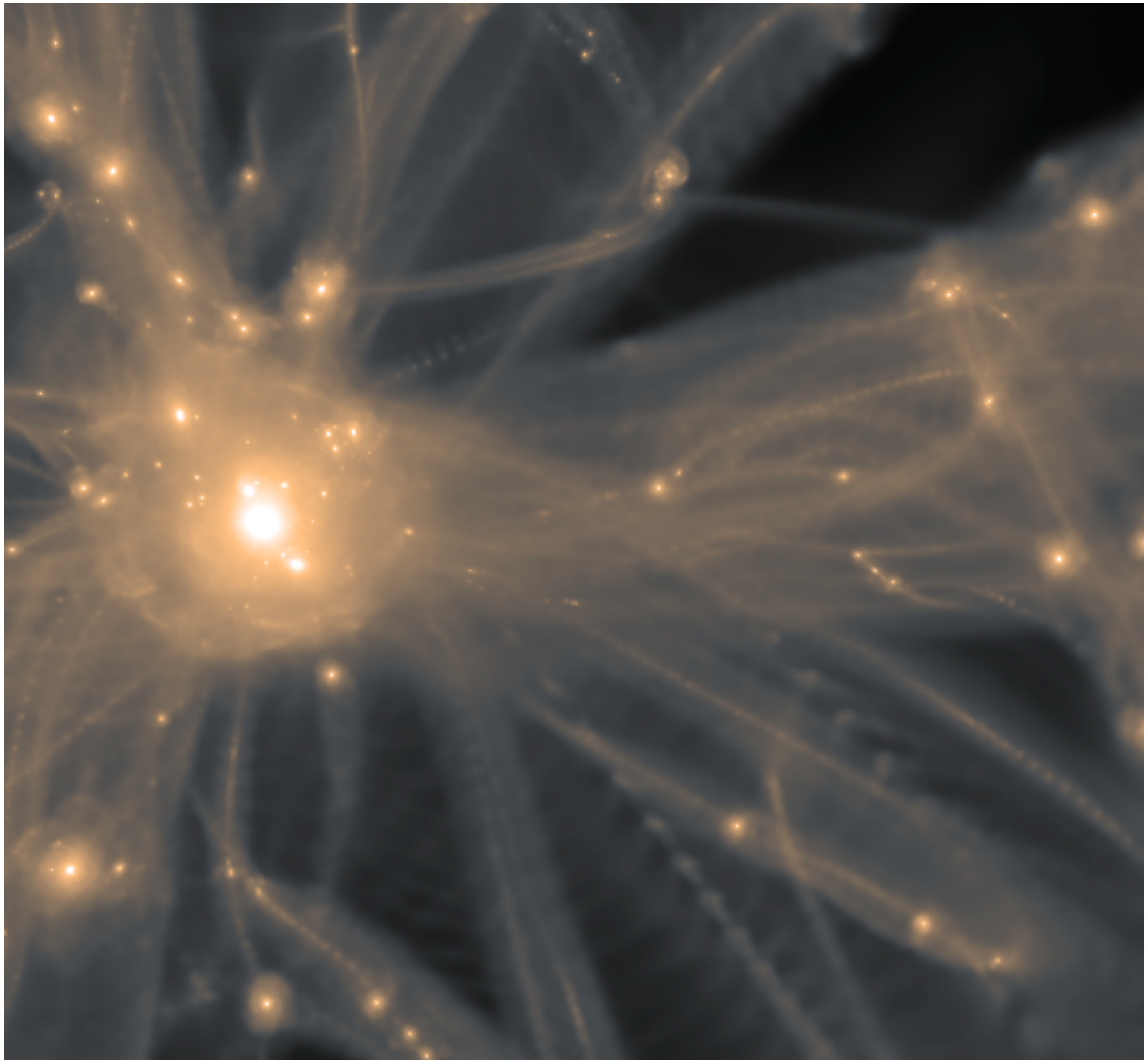
kNN-CDFs Combines estimator and Statistic

Perhaps this more useful generally?

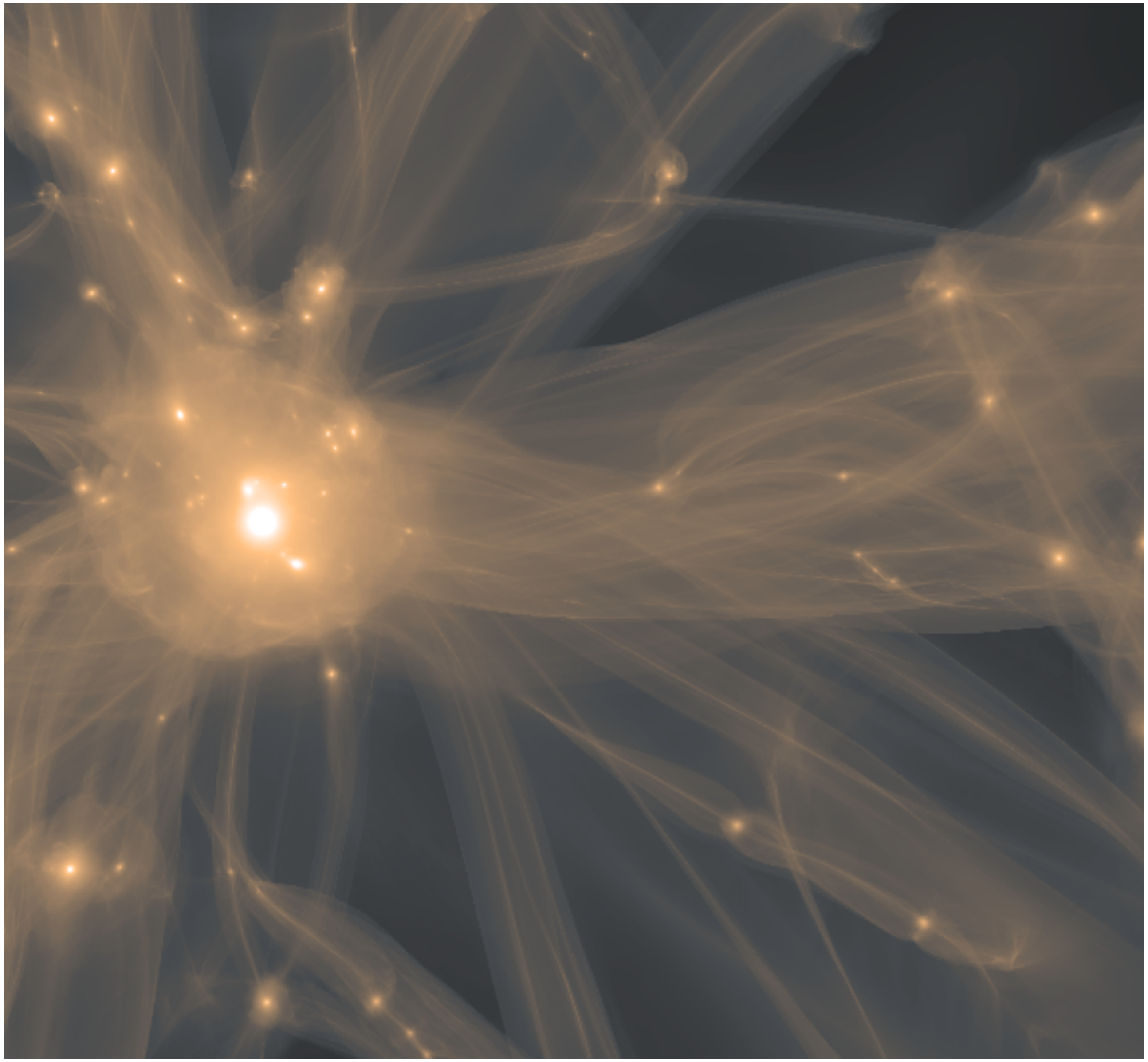


Conclusions

- Developed a novel summary statistic: kNN-CDFs
    - Informative; Interpretable; fast and robust to measure; straightforward to predict; sensitive to all n-point functions and complementary to 2pt, 3pt functions; complete statistic for isotropic fields; no averaging or binning; no nuisance parameters
    - Shot noise is part of the modeling and applications to samples with low statistics is more constraining with kNN-CDFs than with 2pt both for auto and cross-correlation applications
    - Applications suggest this approach is useful
- We provide some example code and implementation guidance and hope you will give it a try and help to further develop this into a ubiquitous approach to characterize clustering.
- Looking forward to working with you this week to combine our approaches.



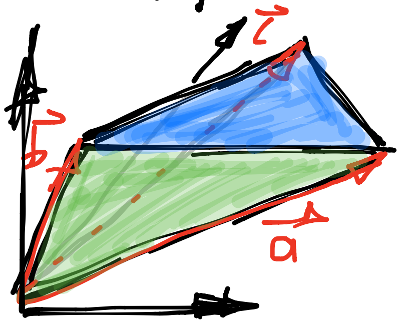
- Abel, Hahn, & Kaehler 2012 MNRAS “Tracing the Dark Matter Sheet in Phase Space”



- Abel, Hahn, & Kaehler 2012 MNRAS “Tracing the Dark Matter Sheet in Phase Space”

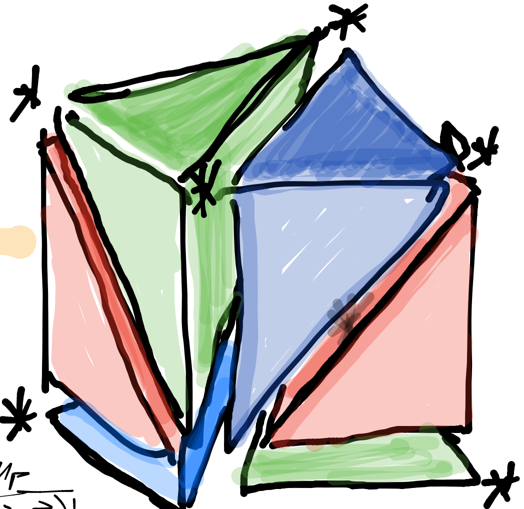
TESSELATE 3D MANIFOLD & TRACK IN 6D PHASE SPACE

- NATURAL TESSELATION SPLITS CUBE INTO 6 EQUAL SIZED TETRAHEDRA
- mass per tetrahedron = 1/6 of DM particle mass.

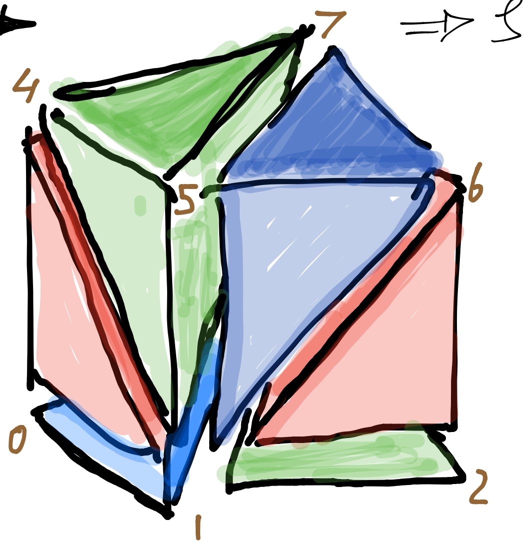


$$V = \frac{|\vec{a}\cdot(\vec{b}\times\vec{c})|}{6}$$

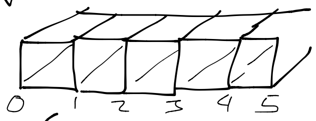
$$\Rightarrow \rho = \frac{m_p}{6V} = \frac{m_p}{|\vec{a}\cdot(\vec{b}\times\vec{c})|}$$



* N-body particle



- Number the edges of the cube
- think of lattice
- Looping over



The initial cartesian (LAGRANGIAN lattice generates the 6N tetrahedra.

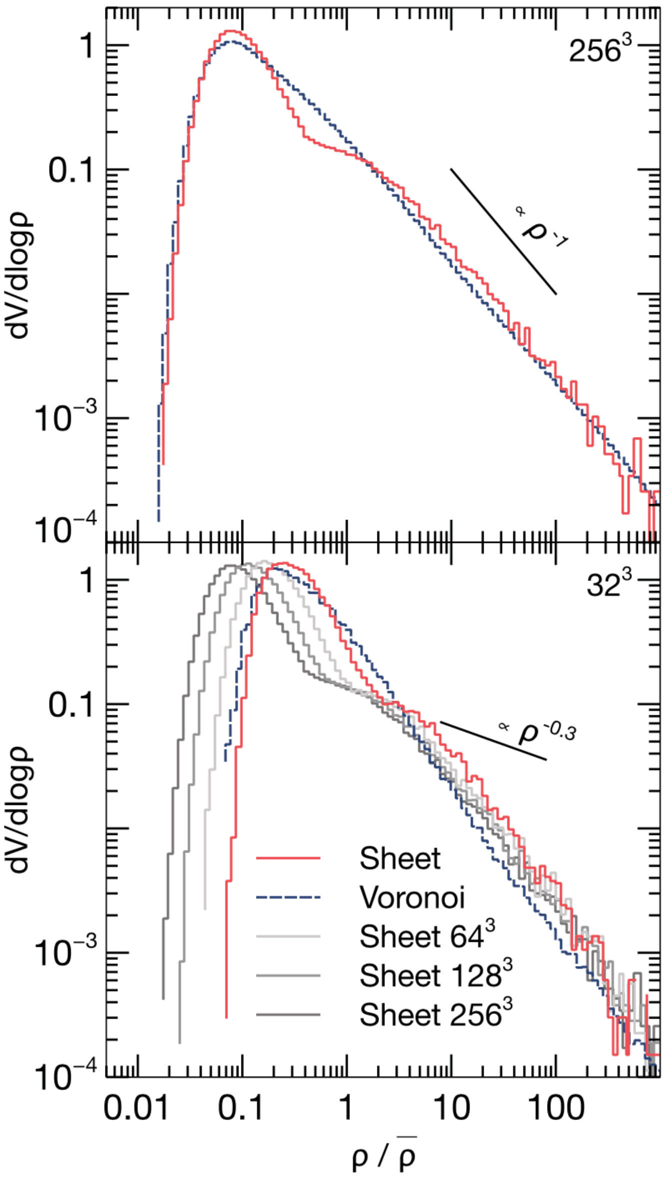


Figure 12. Volume-weighted density distributions. The top panel shows the histograms for the 256^3 run, the bottom panel those for the 32^3 run. The zeroth-order density estimated from a Voronoi tessellation is shown with a dashed line and the total sheet DM density with a solid line. At both resolutions, both the Voronoi and the stream density approach a ρ^{-1} power law at high densities. Also, the two methods produce different estimates at intermediate densities of $\rho/\bar{\rho} \sim 10$. The bottom panel also shows in grey the density histograms from our method for all simulations to aid the comparison.

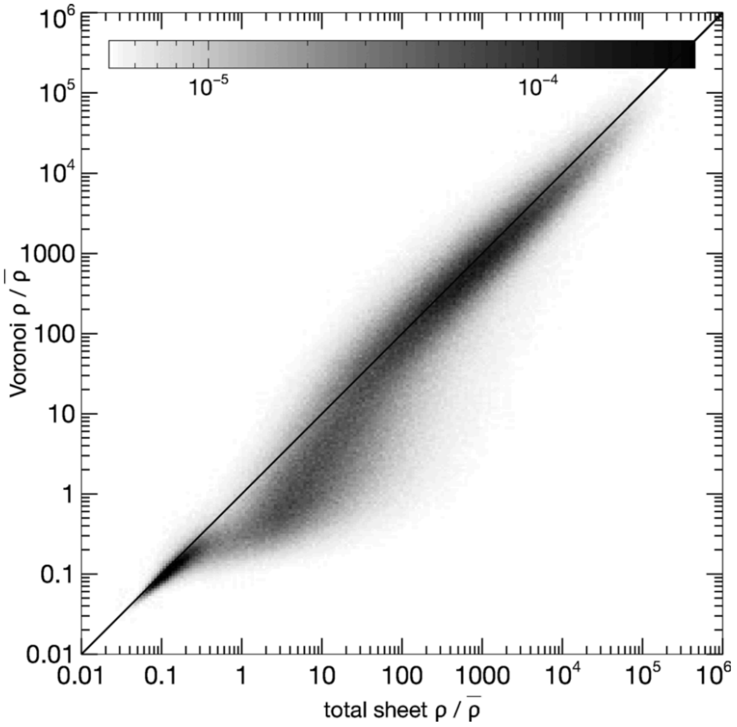


Figure 13. 2D histogram comparing the zeroth-order Voronoi density estimate to the total sheet density. The correspondence is quite good. The largest difference is observed for values between three times and 30 times the mean density of DM. The zeroth-order Voronoi density estimators overestimate the volumes in regions around filaments and sheets.

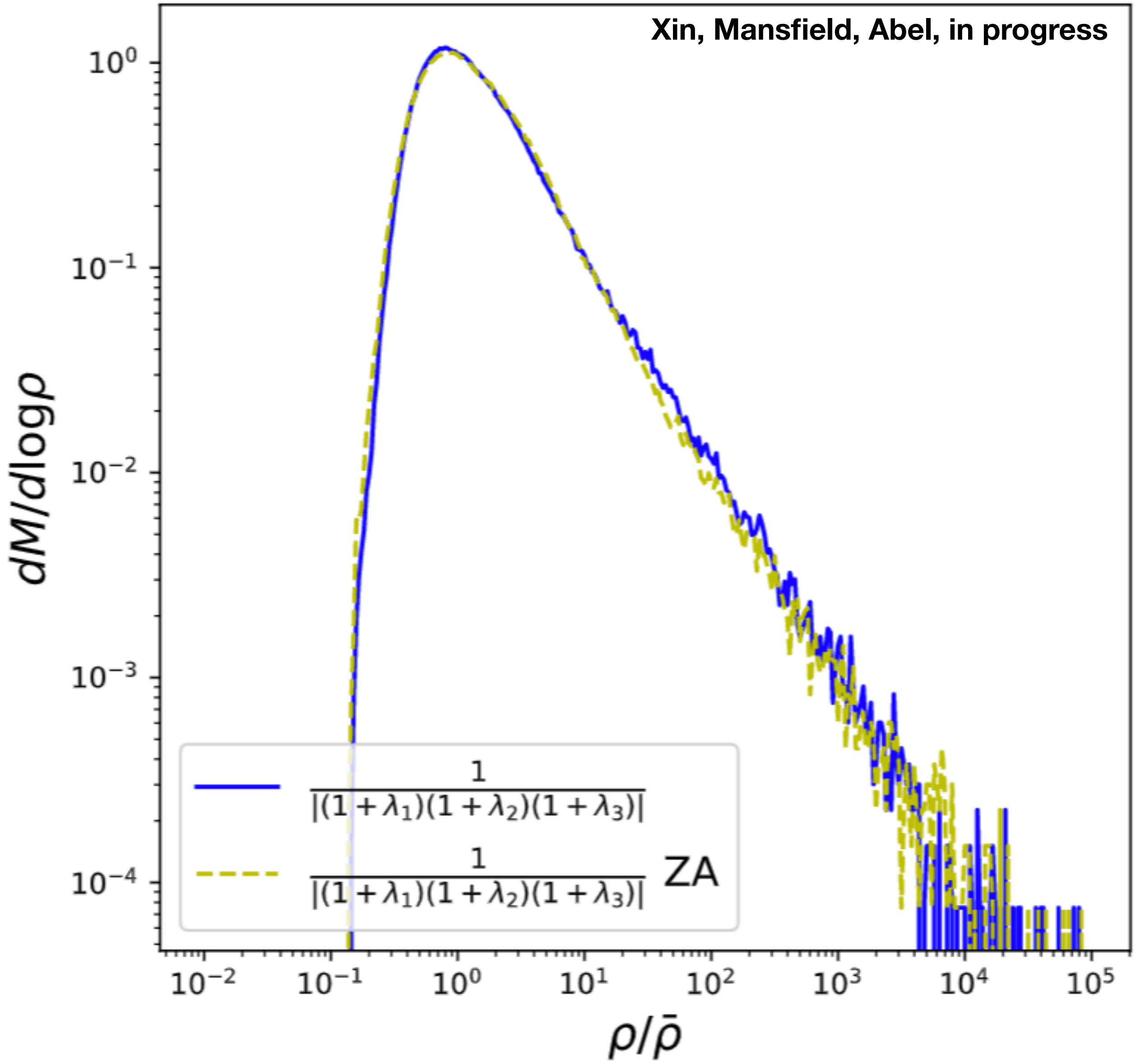
Some things easier to predict than others

- Intriguing:

- 64^3 simulation, 40 Mpc/h

- tetrahedra with initial side length of 10 Mpc/h

- Mass weighted final density extraordinarily well predicted up to over-densities of 1e3 or even 1e4!



Xin, Mansfield, Abel, in progress