Draft version October 18, 2020
Typeset using LaTeX **modern** style in AASTeX62

Two-point statistics without bins: A continuous-function generalization of the correlation function estimator for large-scale structure

Kate Storey-Fisher[1] and David W. Hogg[1, 2, 3, 4]

[1] *Center for Cosmology and Particle Physics, Department of Physics, New York University*
[2] *Center for Data Science, New York University*
[3] *Max-Planck-Institut für Astronomie, Heidelberg*
[4] *Flatiron Institute, Simons Foundation*

(Received XXX; Accepted YYY)

ABSTRACT

The two-point correlation function (2pcf) is the most important statistic in structure formation, used to measure the clustering of density field tracers (e.g. galaxies). Current estimators of the 2pcf, including the standard Landy-Szalay (LS) estimator, evaluate the 2pcf in hard-edged bins of separation between objects, which is inappropriate for the science context and results in a loss of information and a poor trade-off between bias and variance. We present a new estimator for the 2pcf, *the Continuous-Function Estimator*, which generalizes LS to a continuous representation and obviates binning in separation or any other pair property. Our estimator replaces the binned pair counts with a linear superposition of basis functions; it outputs the best-fit linear combination of basis functions to describe the 2pcf. It is closely related to the estimator used in linear least-squares fitting. The choice of basis can take into account the expected form of the 2pcf, as well as its dependence on properties other than separation. We show that the Continuous-Function Estimator can estimate the clustering of artificial data in representations that provide more accuracy with fewer basis functions than LS. The Continuous-Function Estimator achieves lower bias and lower variance than LS. We also demonstrate that the estimator can be used to directly estimate the Baryon Acoustic Oscillation scale. Critically, these will permit reductions in the number of mock catalogs required for covariance estimation, currently the limiting step in 2pcf measurements. We discuss applications and limitations of the Continuous-Function Estimator for present and future studies of large-scale structure, including determining the dependence of clustering on galaxy properties and potentially unifying real-space and Fourier-space approaches to clustering measurements.
KSF says: say in abstract that we'll need far fewer mocks; ok to be repetitive

Keywords: Astrostatistics techniques (1886), Baryon acoustic oscillations (138), Cosmology (343), Two-point correlation function (1951), Large-scale structure of the universe (902), Redshift surveys (1378)

1. INTRODUCTION

The large-scale structure (LSS) of the Universe is critical to our understanding of fundamental cosmology. It encodes information about the physics of the early Universe and the subsequent expansion history. In particular, LSS measures the Baryon Acoustic Oscillation (BAO) scale, which results from density fluctuations in the baryon-photon fluid. The distance traveled by these density waves before recombination imprints a feature on the statistical description of the LSS, which can be used to determine the characteristic BAO length scale (Eisenstein & Hu 1997). The LSS also contains the signature of redshift-space distortions caused by the peculiar velocities of galaxies, which are used to measure the growth rate of structure (Kaiser 2014). Additionally, the LSS can be used to constrain galaxy formation in conjunction with models of galaxy bias (e.g. Hamilton 1988). With current observations, the LSS is well-described by a cold dark matter model with a cosmological constant, the standard ΛCDM model. Upcoming galaxy surveys will observe larger volumes with improved measurements, allowing us to test ΛCDM to even higher precision.

The most important statistic for characterizing the LSS is the two-point correlation function (2pcf). The 2pcf measures the excess frequency at which any two galaxies are separated by a given distance, compared to a uniform distribution; effectively, it characterizes the strength of clustering at a given spatial scale. In calculating the 2pcf, the nontrivial survey boundaries of the surveys prevent us from directly summing pair counts. To account for the survey boundaries as well as regions corrupted by issues such as bright foreground stars, a set of random points are Poisson-distributed within the acceptable survey window. The pairwise correlations of these unclustered points are used to normalize out the survey window when estimating the 2pcf of the clustered data.

The 2pcf is traditionally estimated in bins of radial separation. This binning introduces inherent limitations. First, the choice of bins requires a trade-off between bias and variance: fewer bins may bias the result, while more bins increases the variance of measurement. Finite-width bins also result in a loss of information about the property in which one is binning. As we work towards extreme precision in large-scale structure, maximizing the information we extract with our analyses will become increasingly important. Finally, the error on the covariance matrix scales with the number of bins; a larger number of bins reduces bias, but means we must estimate a very large covariance matrix to achieve the required precision. This is currently the limiting step in LSS analyses, requiring many mock catalogs tailored to the survey, and

covariance matrix computation will further limit cosmological constraints as survey size increases.

More generally, binning adds arbitrary boundaries between continuous data; results should not depend on bin choice, yet they sometimes do. Lanzuisi et al. (2017) noted that the choice of binning axis impacts the detected correlation between the luminosity of active galactic nuclei and their host galaxies; Grimmett et al. (2020) devised a method to investigate this correlation in a continuous manner using a hierarchical Bayesian model, eliminating the need for binning. Bailoni et al. (2016) explored the dependence of clustering analyses on the number of redshift bins, finding a non-negligible difference in cosmological parameter uncertainties. The implications for BAO analyses were explored by Percival et al. (2014), who found that there is an optimal bin width given the analysis method. This balances the increasing statistical error with small bins and the offset in the derived BAO peak location with large bins; the effects are small but non-negligible. It is clear that, when analyzing smooth quantities such as LSS statistics, binning is sinning.

Estimators of the 2pcf have been studied extensively (Peebles & Hauser 1974; Davis & Peebles 1983; Hamilton 1993). The current standard estimator was proposed by Landy & Szalay (1993), hereafter LS. It is based on summing all data pairs DD with a given separation and using data-random pairs DR and random pairs RR to correct for the survey boundary. The LS estimator of the correlation function $\hat{\xi}_k$ for the k^{th} bin in separation r is

$$\hat{\xi}_k = \frac{DD_k - 2DR_k + RR_k}{RR_k}. \tag{1}$$

Compared with other estimators based on simple combinations of DD, DR and RR, LS has been shown to have the lowest bias and variance (Kerscher et al. 2000). Estimators of the 2pcf must also take into account the imperfect nature of the survey, including systematic effects, the target completeness, and fiber collisions. To account for these, each galaxy pair is sometimes assigned a weight, and pair counts are replaced by the sum of pair weights.

Variations on the random catalog pair count method have been proposed in recent years. Demina et al. (2016) replaced the DR and RR terms with an integral over the probability map, reducing computation time and increasing precision. An estimator proposed by Vargas-Magaña et al. (2013) iterates over sets of mock catalogs to find an optimal linear combination of data and random pair counts, reducing the bias and variance. An alternative estimator, the marked correlation function (e.g. White & Padmanabhan 2009), avoids the use of a random catalog altogether: it considers the ratio between the 2pcf and a weighted correlation function in which weights are assigned based on galaxy properties, such as the local density. These estimators have all taken probabilistic approaches; others have taken a likelihood approach. Baxter & Rozo (2013) introduced a maximum likelihood estimator for the 2pcf, which achieves

lower variance compared to the LS estimator, enabling finer binning and requiring a smaller random catalog for the same precision.

These estimators present improvements to LS, but they are still limited to estimates in separation bins. Some require additional computational costs or layers of complexity, so the standard formulation of LS continues to be the default estimator used in most analyses.

In this paper, we present a new estimator for the correlation function, the Continuous-Function Estimator, which generalizes the LS estimator to produce a continuous estimation of the 2pcf. The Continuous-Function Estimator projects the galaxy pairs onto a set of continuous basis functions and computes the best-fit linear combination of these functions. The basis representation can depend on the pair separation as well as other desired properties, and can also utilize the known form of the 2pcf. For top-hat basis functions, the Continuous-Function Estimator exactly reduces to the LS estimator. This estimator removes the need for binning and allows for the 2pcf to be represented by fewer basis functions, requiring fewer mock catalogs to compute the covariance matrix. It is particularly well-suited to the analysis of LSS features such as the BAO peak; we find that we can accurately locate the peak with fewer components.

This paper is organized as follows. In §2, we motivate our estimator and explain its formulation. We demonstrate its application on a simulated data set, including a toy BAO analysis, in §3. We discuss the implications and other possible applications in §4.

2. MOTIVATION AND FORMULATION

In this paper, we use the following notation. We write vectors in bold and lowercase, e.g. $\boldsymbol{v}$; tensors in bold and uppercase, e.g. $\boldsymbol{T}$; and unstructured data blobs in sans serif, e.g. G. A hat above a symbol, e.g. $\hat{\boldsymbol{\xi}}$, indicates an estimate of the value. KSF says: where should this go?

2.1. *Standard Two-Point Correlation Function Estimation*

The standard approach to estimating the two-point correlation function involves counting pairs of tracers within a survey volume as a function of separation scale. Let's assume we have a data catalog with N_D objects within a sky volume. We also require a random catalog with N_R objects distributed uniformly throughout the same volume. We can define a set of separation bins which we will use to estimate the 2pcf at various scales. We are then ready to sum in each bin the relevant pairs of objects within and across our catalogs. In standard notation, these pair counts are written as DD, DR, and RR, as in Equation 2. To clarify that these are in fact vectors, with length K where K is the number of bins, we use the symbol $\boldsymbol{v}$; then, for example, the data-data pair counts DD become $\boldsymbol{v}_{\mathrm{DD}}$. We can then write the LS estimator as

$$\hat{\boldsymbol{\xi}} = \frac{\boldsymbol{v}_{\mathrm{DD}} - 2\,\boldsymbol{v}_{\mathrm{DR}} + \boldsymbol{v}_{\mathrm{RR}}}{\boldsymbol{v}_{\mathrm{RR}}}. \tag{2}$$

The pair counts are defined explicitly as KSF says: should i define something like $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ so i don't have to keep writing it out? or better to be explicit throughout?

$$[\boldsymbol{v}_{\rm DD}]_k \equiv \frac{2}{N_{\rm D}(N_{\rm D}-1)} \sum_{nn'} i(g_k < |\boldsymbol{r}_n - \boldsymbol{r}_{n'}| < h_k) \tag{3}$$

$$[\boldsymbol{v}_{\rm DR}]_k \equiv \frac{1}{N_{\rm D}N_{\rm R}} \sum_{nm} i(g_k < |\boldsymbol{r}_n - \boldsymbol{r}_{m}| < h_k) \tag{4}$$

$$[\boldsymbol{v}_{\rm RR}]_k \equiv \frac{2}{N_{\rm R}(N_{\rm R}-1)} \sum_{mm'} i(g_k < |\boldsymbol{r}_m - \boldsymbol{r}_{m'}| < h_k), \tag{5}$$

where $[\boldsymbol{v}]_k$ is the pair counts in bin k (which has bin edges g_k and h_k), i is an indicator function that returns 1 if the the condition is true and otherwise returns 0, $\boldsymbol{r}$ is the tracer position, the n and n' indices index data positions, and the m and m' indices index random catalog positions. The tracer position can be in real or redshift space, or broken down into the transverse and line-of-sight directions in the anisotropic correlation function; in this paper we consider the isotropic real-space 2pcf for simplicity, but the estimators detailed here apply equally well to these alternative configurations. Our results here also apply in a straightforward way to the cross-correlation; for details see §4.1.

The LS estimator is known to be optimal (i.e. it is unbiased and has minimum variance) under a particular set of conditions: in the limit of unclustered data, for a data volume much larger than the scales of interest, and an infinitely large random catalog. In practice the latter two limits are sufficiently satisfied, but the data we are interested in are clustered. Vargas-Magaña et al. (2013) show that for clustered data, the LS estimator has lower variance than other estimators, but does not reach the Poisson noise limit. When applied to clustered data, LS does show a bias on very large scales (>130 h^{-1} Mpc), but the bias is significantly smaller than that of most other estimators (Kerscher 1999, Vargas-Magaña et al. 2013). LS is also less sensitive to the number of random points than other estimators (Kerscher et al. 2000). While LS has been sufficient for past analyses, its persisting bias and suboptimal variance under imperfect conditions mean that improvement is possible, and will be necessary for realistic large-scale structure measurements on modern datasets.

2.2. *Least Squares Fitting*

Estimating clustering is closely related to least-squares fitting. We are essentially trying to find the best representation of spatial data in the space of two-point radial separation. Recall that the linear least-squares fit to a set of data is

$$\boldsymbol{x} = [\boldsymbol{A}^{\rm T}\boldsymbol{C}^{-1}\boldsymbol{A}]^{-1}[\boldsymbol{A}^{\rm T}\boldsymbol{C}^{-1}\boldsymbol{y}] \tag{6}$$

where $\boldsymbol{x}$ is the vector of best-fit parameters, $\boldsymbol{A}$ is a design matrix with zeroth and first order terms (and possibly higher order) functions of fitting features, $\boldsymbol{C}$ is the covariance matrix, and $\boldsymbol{y}$ is a column vector of $\boldsymbol{y}$ data to be fit. The second bracketed

term projects the data onto the features; the first bracketed term renormalizes this quantity. KSF says: renormalizes what?? In the case of the 2pcf, the observed data is the pair counts at a given separation, and the weights are provided by the pair counts of the random catalog. Indeed, this is reminiscent of the so-called natural estimator of the 2pcf, $\boldsymbol{\xi} = \boldsymbol{v}_{\rm DD}/\boldsymbol{v}_{\rm RR} - 1$ (e.g. Kerscher et al. 2000).

From this connection, we can infer the form of the estimator. KSF says: What else to say in this section?? How to expand on what's here?

2.3. *The Continuous-Function Estimator*

Inspired by least-squares fitting, we generalize the LS estimator defined above in Equations 3-5. We generalize the indicator function i to any function $\boldsymbol{f}$, which returns a vector of length K where K is the number of basis functions. We further generalize the arguments of the function to any properties of the galaxies, rather than just the separation between pairs; we call G the data payload for a single galaxy. This gives us, instead of pair counts, a vector of amplitudes of the basis functions, $\boldsymbol{v}$. We then define the Continuous-Function Estimator as KSF says: STILL FIGURING OUT we should double count so that we could be sure to be symmetric; change 2-¿1 and say that we explicitly double count here. tho in an implementation we might not, and then prefactor changes. also say that the notation, sum over n n', does not self add. if want to be explicit, use double sum. think i should do this... then need to do double sum for LS def. for my version, write it as parallel as possible as LS, and call out changes

$$\boldsymbol{v}_{\rm DD} \equiv \frac{2}{N_{\rm D}\,(N_{\rm D}-1)} \sum_n \sum_{n'} \boldsymbol{f}(\mathsf{G}_n, \mathsf{G}_{n'}) \tag{7}$$

$$\boldsymbol{v}_{\rm DR} \equiv \frac{1}{N_{\rm D}\,N_{\rm R}} \sum_n \sum_m \boldsymbol{f}(\mathsf{G}_n, \mathsf{G}_m) \tag{8}$$

$$\boldsymbol{v}_{\rm RR} \equiv \frac{2}{N_{\rm R}\,(N_{\rm R}-1)} \sum_m \sum_{m'} \boldsymbol{f}(\mathsf{G}_m, \mathsf{G}_{m'}) \tag{9}$$

$$\boldsymbol{T}_{\rm RR} \equiv \frac{2}{N_{\rm R}\,(N_{\rm R}-1)} \sum_m \sum_{m'} \boldsymbol{f}(\mathsf{G}_m, \mathsf{G}_{m'}) \cdot \boldsymbol{f}^{\mathsf{T}}(\mathsf{G}_m, \mathsf{G}_{m'}). \tag{10}$$

Then, we can compute the 2pcf as

$$\boldsymbol{a} \equiv \boldsymbol{T}_{\rm RR}^{-1} \cdot (\boldsymbol{v}_{\rm DD} - 2\,\boldsymbol{v}_{\rm DR} + \boldsymbol{v}_{\rm RR}) \tag{11}$$

$$\hat{\boldsymbol{\xi}}(\mathsf{G}_l, \mathsf{G}_{l'}) \equiv \boldsymbol{a}^{\mathsf{T}} \cdot \boldsymbol{f}(\mathsf{G}_l, \mathsf{G}_{l'}) \tag{12}$$

where a K-vector of the computed *amplitudes* of the basis functions, and G_l and $\mathsf{G}_{l'}$ contain the data values at which to evaluate ξ. KSF says: TOWRITE: discuss index / sum notation, sums over number of points in dataset KSF says: say that cross-correlation is easy to see from this, point to section in discussion. could also put this right below the definition. We emphasize that these are not real datapoints, but

instead allow us to evaluate the 2pcf at any set of parameters. In the standard case, G_l and $\mathsf{G}_{l'}$ would effectively be an imaginary pair of galaxies that has a separate r at which we want to evaluate ξ, and we would compute $\hat{\boldsymbol{\xi}}$ for such a pair at every separation we are interested in. With our general formulation, we could choose basis functions that depend on other galaxy properties, to investigate the effect of these on the 2pcf; then, we would also choose each G_l and $\mathsf{G}_{l'}$ pair to have values of properties at which we want to evlauate $\hat{\boldsymbol{\xi}}$. In the rest of this paper, however, we will only take into account the separation between pairs, and we will write $\hat{\boldsymbol{\xi}}(r)$.

The Continuous-Function Estimator can be straightforwardly generalized to cross-correlations between two datasets. In this case, we consider datasets D_1 and D_2, and associated random catalogs R_1 and R_2. We then have cross-correlations rather than auto-correlations for the data-data and random-random terms, and two different data-random terms, crossing each dataset with the opposite random catalog. The data-data term becomes

$$\boldsymbol{v}_{\mathrm{D_1D_2}} \equiv \frac{1}{N_{\mathrm{D_1}}\, N_{\mathrm{D_2}}} \sum_{n_1} \sum_{n_2} \boldsymbol{f}(\mathsf{G}_{n_1}, \mathsf{G}_{n_2}) \tag{13}$$

where n_1 and n_2 index the data points in each catalog, and the normalization factor is now simply the product of catalog sizes as we are no longer concerned with double-counting. The other terms ($\boldsymbol{v}_{\mathrm{D_1R_2}}$, $\boldsymbol{v}_{\mathrm{D_2R_1}}$, $\boldsymbol{v}_{\mathrm{R_1R_2}}$, $\boldsymbol{T}_{\mathrm{R_1R_2}}$) generalize as one would expect. The amplitudes then becomes

$$\boldsymbol{a} \equiv \boldsymbol{T}^{-1}_{\mathrm{R_1R_2}} \cdot (\boldsymbol{v}_{\mathrm{D_1D_2}} - \boldsymbol{v}_{\mathrm{D_1R_2}} - \boldsymbol{v}_{\mathrm{D_2R_1}} + \boldsymbol{v}_{\mathrm{R_1R_2}}) \tag{14}$$

and we use this to compute the estimator as in 12.

If we consider only the pair separation, and make a proper choice of $\boldsymbol{f}$, the Continuous-Function Estimator reduces to the Landy-Szalay estimator. Explicitly, from our full galaxy pair data G_n and $\mathsf{G}_{n'}$, we can use only their separation, $|\boldsymbol{r}_n - \boldsymbol{r}_{n'}|$. We can then define a set of K basis functions $\boldsymbol{f}$ as

$$\boldsymbol{f}_k(\mathsf{G}_n, \mathsf{G}_{n'}) = i(g_k < |\boldsymbol{r}_n - \boldsymbol{r}_{n'}| < h_k). \tag{15}$$

This is the common top-hat (or rectangular) function; the index k denotes a particular bin in separation, and here also indexes the basis functions, as each top-hat is a separate basis function. In this case the $\boldsymbol{v}_{\mathrm{DD}}$, $\boldsymbol{v}_{\mathrm{DR}}$ and $\boldsymbol{v}_{\mathrm{RR}}$ vectors simply become binned pair counts, with bin edges g_k and h_k as before. The $\boldsymbol{T}_{\mathrm{RR}}$ tensor becomes diagonal, with its diagonal elements equal to the $\boldsymbol{v}_{\mathrm{RR}}$ vector elements. Then the evaluation of the amplitudes $\boldsymbol{a}$ and the correlation function estimate $\hat{\boldsymbol{\xi}}$ results in the equivalent of the LS estimator–just displayed in a continuous form.

We call this generalized 2pcf estimator the Continuous-Function Estimator. It replaces the binned pair counts of LS with any set of basis functions; the linear superposition of these basis functions is our estimate of the 2pcf. Essentially, the Continuous-Function Estimator outputs the best-fit linear combination of basis functions to describe the 2pcf. In this sense, it is deeply related to the linear least-squares

fitting described above. With our formulation, we no longer need to first bin our data and then fit a function; rather, the estimator directly projects the data (the pair counts) onto the desired function. This function can be nearly anything; the only limitation is it must be able to be written as a linear combination of basis functions.

With this generalized two-point estimator, the basis functions need not have hard edges like the top-hat function. They can instead be smooth functions of the pair separation, or chosen to suit the science use case. Further, the bases can make use of other information about the tracers or the survey; they are extremely general. The estimator also has the property that it is invariant under affine transformations, as it should be so that the result does not depend on e.g. the magnitude of the bases; we show this in Appendix A.

We can also write down the form of the Continuous-Function Estimator when we are working with a periodic box and don't need to worry about the survey window. In this case, we can analytically compute the $\boldsymbol{v}_{\mathrm{RR}}$ term, as well as the $\boldsymbol{T}_{\mathrm{RR}}$ term, and use the natural form of the 2pcf estimator. The derivation and formulation of these terms are shown in Appendix B.

KSF says: Discuss the limit of infinitesimal bins? Do we know what this is? Hogg says: Yes, we should discuss this, and show how we are related to that. KSF says: I'm not sure what to say about this.

We implement this estimator based on the correlation function package `Corrfunc` by Sinha & Garrison (2019). `Corrfunc` is the state-of-the-art package for computing correlation functions and other clustering statistics; it is extremely fast and user-friendly, and is used in many published analyses. It is also modular and open-source, making it a natural choice as a base for our implementation. Our implementation of the Continuous-Function Estimator is also open-source and available at `github.com/kstoreyf/Corrfunc`.

3. EXPERIMENTS AND RESULTS

3.1. *Lognormal Mock Catalogs*

We demonstrate the application of the Continuous-Function Estimator on a set of artifical data. We generate lognormal mock catalogs (Coles & Jones 1991) using the `lognormal_galaxies` code by (**?**). We use an input power spectrum with the Planck cosmology, the same parameters used for the MultiDark-PATCHY simulations (**?**) made for the Baryon Oscillation Spectroscopic Survey (Boss, Dawson et al. 2013. This assumes a cold dark matter model with $\Omega_m = 0.307115$, $\Omega_b = 0.048206$, $\sigma_8 = 0.8288$, $n_s = 0.9611$, and $h = 0.6777$. Our fiducial test set is 1000 realizations of periodic cubes with size $(750\ h^{-1}\,\mathrm{Mpc})^3$ and a galaxy number density of 2×10^{-4}. We choose to perform these tests on periodic boxes so that we may compute the random-random term analytically (see Appendix B), significantly cutting down on computation time. We note, though, that we expect our results to hold for catalogs with realistic survey windows and random-random terms computed directly with the Continuous-Function

Estimator. KSF says: I am realizing that it's a bit funny that our estimator hinges on the tensor random term, yet we are computing that analytically in all these tests. I think it's very fair, bc the formulation of the analytic $\boldsymbol{v}_{\rm RR}$ & $\boldsymbol{T}_{\rm RR}$ terms is based on our formulation. I don't think this is a reason to change everything now; but maybe i should include the analytic formulations above, and refer to the appendix for the full derivation?

3.2. *Comparison of Standard Tophat Basis Functions*

We first estimate the correlation function of our mocks using the the standard estimator. We choose 15 separation (r) bins in the range $36 < r < 156\ h^{-1}\,$Mpc, each with a width of $8\ h^{-1}\,$Mpc; this was found to be the optimal bin width by Percival et al. (2014), and is standard for two-point analyses. We apply the estimator to each of our 1000 mock catalogs. The mean of these estimated correlation functions is shown in Figure 1 (grey squares); the error bars show the standard deviation of the 1000 mocks in each bin. KSF says: do i note the legend info here, or just in figure caption? We also show the true input correlation function (thin black line), and the bottom panel shows the absolute error between the estimated and true correlation functions.

There remains an ambiguity in the r-value at which to plot the result of the standard estimator. The volume-weighted average is often used, or a weighted average depending on the pairs in the bin; this choice propagates to differences in comparing the estimate to models (though at the precision of current surveys these differences are not significant). Here we plot the standard estimator with the volume-weighted average.

We demonstrate the Continuous-Function Estimator with a tophat basis function as a check. We choose tophat functions with the same locations and widths as the bins used for the standard estimator; these are shown in the top panel of Figure 1. As this estimator computes the 2pcf in a continuous form, we plot the result as a continuous function at every r value (thin blue line). This results in a step function form for the correlation function, which is in fact what the typical estimator is computing. The values of the correlation function at each step exactly align with the result of the standard estimator. In fact, we emphasize that this step function is exactly what the standard estimator is estimating; we have just made explicit the fact that the each estimate is for the entire bin. When we look at the error with respect to the truth (bottom panel), the error blows up at the edges of each bin, where the continuous estimate deviates most significantly from the truth. This demonstrates that the standard tophat estimator is not a good representation of the true 2pcf.

3.3. *Demonstration using Spline Basis Functions*

A natural extension of tophat basis functions is the B-spline. B-splines of order n are piecewise polynomials of order $n-1$; they constitute the basis functions for spline interpolation. KSF says: do i need a citation here? would be a textbook probs... They have the nice property that the functions and their derivatives can be continuous,

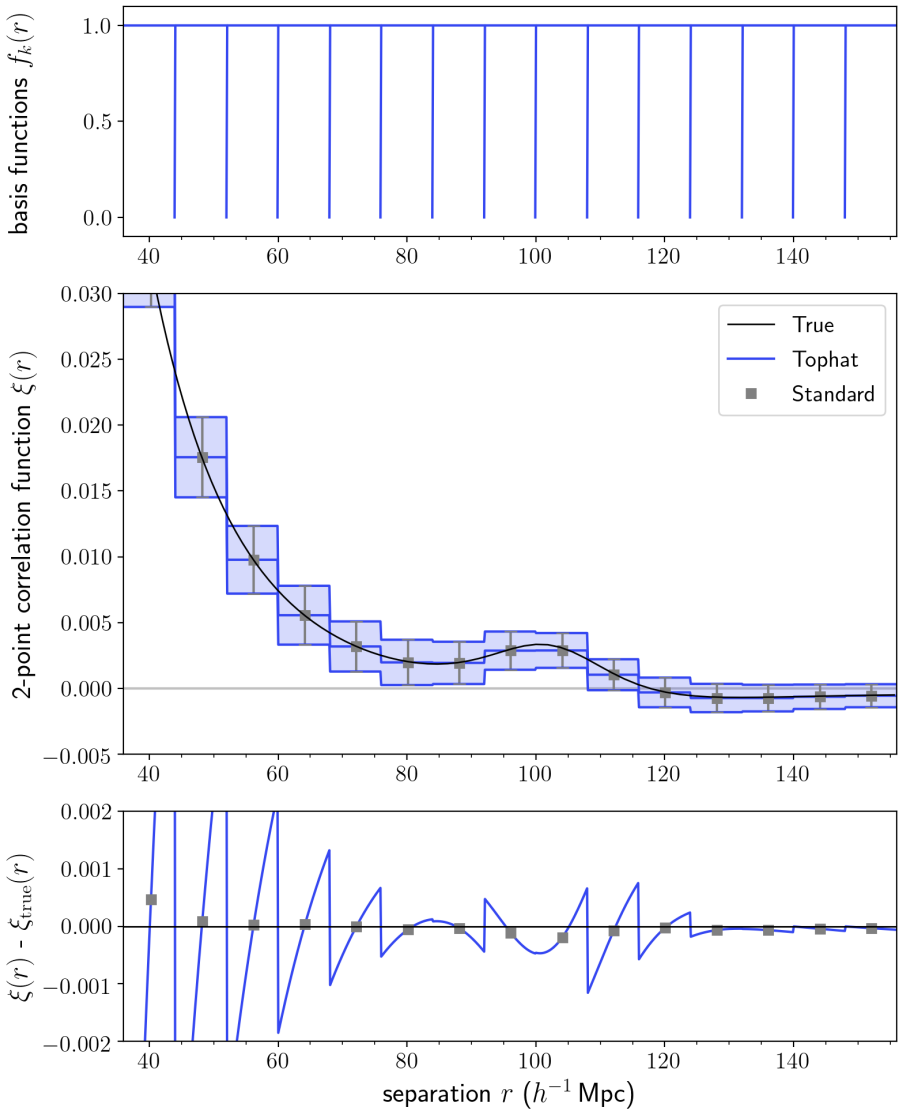

Figure 1. A comparison between the Continuous-Function Estimator with a tophat basis (thin blue lines) and the standard estimator (grey squares). The top panel shows the basis functions used for the tophat estimator. The middle panel shows the mean of the estimated correlation functions for 1000 mock catalogs, compared to the true input 2pcf (thin black line); the shaded region and errorbars are the standard deviation of the 2pcf estimate. The lower panel shows the absolute error between the estimate and true 2pcf. The Continuous-Function Estimator with a tophat basis is exactly equivalent to the standard estimator, but in a continuous form, emphasizing the fact that binning results in a poor representation of the true 2pcf.

depending on the order. Further, B-splines are well-localized, which provides a more direct comparison to the typical tophat basis (which is entirely localized). For this demonstration we use fourth-order B-splines, which constitute the set of basis functions for a cubic spline, as they are the lowest-order spline to have a continuous first derivative.

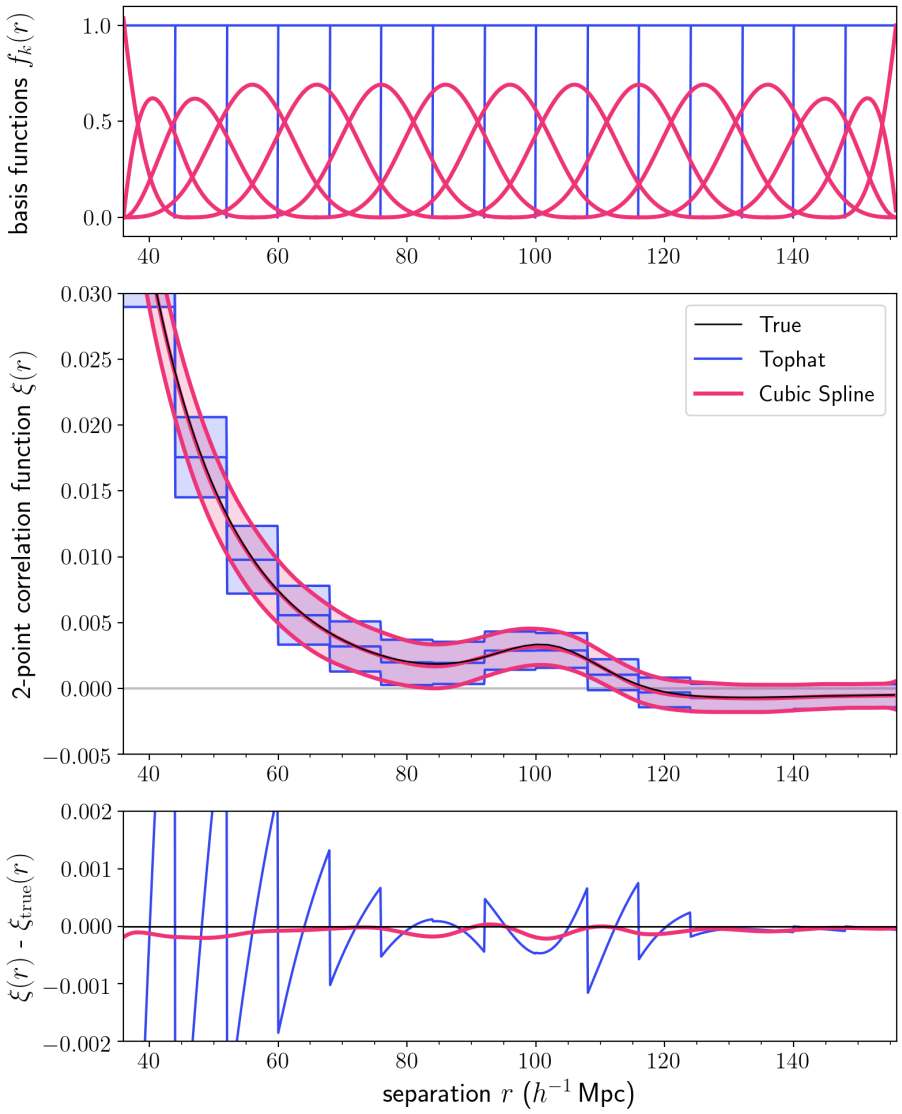

Figure 2. A comparison between the Continuous-Function Estimator with a cubic spline basis function (thick red) and a standard tophat basis (thin blue). The top panel shows the basis functions used for each measurement. The middle panel shows the mean of the estimated correlation functions for each of the 1000 mock catalogs compared to the true input 2pcf (thin black); the shaded region is the standard deviation. The lower panel shows the absolute error between the estimate and true 2pcf. It is clear that the spline basis function results in a correlation function that is a better representation of the true 2pcf in its shape and smoothness.

We compare the estimator with a cubic spline basis to the standard estimator, reformulated as continuous functions using a tophat basis; the results are shown in Figure 2. The basis functions are shown in the top panel of the figure. We use the same tophat basis as above (thin blue). For the cubic spline basis, we use the same r-range and number of basis functions, and knots chosen to evenly span the range (thick red). The cubic spline bases on the edge have different shapes such that they remain normalized; we note that generally, one should choose the basis functions such

that the range of the 2pcf that is of interest does not depend on the range of the basis functions.

The estimator using the cubic spline basis clearly produces a better fit to the true correlation function (thin black) in its shape and smoothness at every point across the scale range, compared to the estimator using the tophat basis. The bottom panel shows the error with respect to the truth; the cubic spline estimator is more generally more accurate, and straightforward to compare to the truth (or model) at every scale. On the other hand, in order to compare the binned correlation to a model, one must integrate the model over the bin range, though in practice the model is often just evaluated at the effective r of each bin. This comparison demonstrates that there exist other sets of basis functions that produce better representations of the data compared to the standard tophat/binned estimator. The choice of a high-order spline may be useful for cases in which one wants a simple representative estimate of the 2pcf, or wants smooth derivatives. Generally, the choice of basis functions should be tailored to the scientific use case; in the next section we explore the case of a BAO analysis.

3.4. *BAO Scale Estimation Test*

Measurement of the BAO scale provides a good use case for our estimator. The BAO feature is a peak in clustering on large scales, ∼150 Mpc, making it less sensitive to small-scale astrophysical effects. It is one of the best tools for constraining cosmological models, in particular the distance-redshift relation (Kazin et al. 2010; Anderson et al. 2011, 2014; Alam et al. 2016).

KSF says: what else do I need to cite for BAO?

We base our BAO estimation on the method of the BOSS DR10 and 11 analysis (Anderson et al. 2014). We estimate the spherically averaged 3-dimensional correlation function, $\hat{\boldsymbol{\xi}}(r)$, where r is the separation between pairs. (BAO analyses is typically done in redshift space, estimating $\hat{\boldsymbol{\xi}}(s)$, where s is the redshift-space separation between pairs, but here we are using a toy model a periodic box in which we know the true galaxy positions so we just use the real-space distance r.) In order to extract information about the baryon acoustic feature from galaxy clustering, we must choose a fiducial cosmological model to convert redshifts to distances. If we choose an incorrect model, the scales in the power spectrum will be dilated, so the oscillation wavelength—and thus the BAO peak position—will be shifted. We can model this shift as a scale dilation parameter, α, which is a function of the relevant distance scales in the true and fiducial cosmologies:

$$\alpha = \left(\frac{D_{\rm A}(z)}{D_{\rm A}^{\rm mod}(z)}\right)^{2/3} \left(\frac{H^{\rm mod}(z)}{H(z)}\right)^{1/3} \left(\frac{r_{\rm s}^{\rm mod}}{r_{\rm s}}\right), \tag{16}$$

where $D_{\rm A}$ is the angular diameter distance, H is the Hubble constant, $r_{\rm s}$ is the sound horizon scale at the drag epoch, and the superscript "mod" denotes the value for the

fiducial model. Qualitatively, if the fit prefers $\alpha > 1$, this suggests the true position of the BAO peak is at a smaller scale than in the fiducial model, whereas if $\alpha < 1$, the peak is at a larger scale.

In standard practice, the fitting function used to determine the value of α is

$$\xi^{\rm fit}(r) = B^2 \xi^{\rm mod}(\alpha r) + \frac{a_1}{r^2} + \frac{a_2}{r} + a_3 \tag{17}$$

where B is a constant that allows for a large-scale bias, and a_1, a_2, and a_3 are nuisance parameters to account for the broadband shape. A χ^2 fit is performed with five free parameters: α, B, a_1, a_2, and a_3. KSF says: TODO: check how this is actually done! The resulting value for α is used to derive the actual values of the distance scales of interest. Typically, density-field reconstruction is performed before applying the estimator to correct for nonlinear growth around the BAO scale (Eisenstein et al. 2007); for our toy example, we will omit this step.

The form of the standard fitting function is well-suited to our estimator, as it is a few-parameter model with a linear combination of terms. To use our estimator to estimate α, we add a term that includes the partial derivative of the model with respect to α. This allows us to have fixed basis functions, and for an initial choice of $\alpha_{\rm guess}$, determine the change in this value needed to improve the fit. Our fitting function is then

$$\xi^{\rm fit}(r) = B^2\, \xi^{\rm mod}(\alpha_{\rm guess}\, r) + C\, k_0\, \frac{{\rm d}\xi^{\rm mod}(\alpha_{\rm guess}\, r)}{{\rm d}\alpha} + a_1\, \frac{k_1}{r^2} + a_2\, \frac{k_2}{r} + a_3\, k_3, \tag{18}$$

where C is an additional coefficient that describes the contribution of the derivative term, and k_0, k_1, k_2, and k_3 are constants that determine the initial amplitude of the basis functions. In this case, the free parameters are B^2, C, a_1, a_2, and a_3. Note that in theory the choice of k_i values shouldn't matter as the estimator is affine invariant (see Appendix A), but in practice reasonable choices are important for stability.

To use the estimator for a BAO measurement, we input these five terms as the five basis functions of our estimator. KSF says: update this given the above list of free params The estimator outputs an amplitude vector $\boldsymbol{a}$ as described in §2.3, which describes the contribution of each basis function—precisely the values of the free parameters. From the value of C, we can determine our estimate of the scale dilation parameter, $\hat{\alpha}$, as $\hat{\alpha} = \alpha_{\rm guess} + C\, k_0$. With this formulation, a value of $C = 0$ indicates that the current $\alpha_{\rm guess}$ gives the best fit to the data (given the chosen cosmological model), while nonzero values give the magnitude and direction of the necessary change in the scale dilation parameter. In practice, we apply an iterative procedure to converge at our best estimate $\hat{\alpha}$; this procedure and other implementation details are described in Appendix C.

We demonstrate this method using the same set of lognormal mocks as in §3.3. We construct a recovery test following that in **?**. We assume the fiducial cosmological model used in **?**: $\Omega_{\rm m} = 0.31$, $h = 0.676$, $\Omega_{\rm b} = 0.04814$, $n_s = 0.97$. As we know the

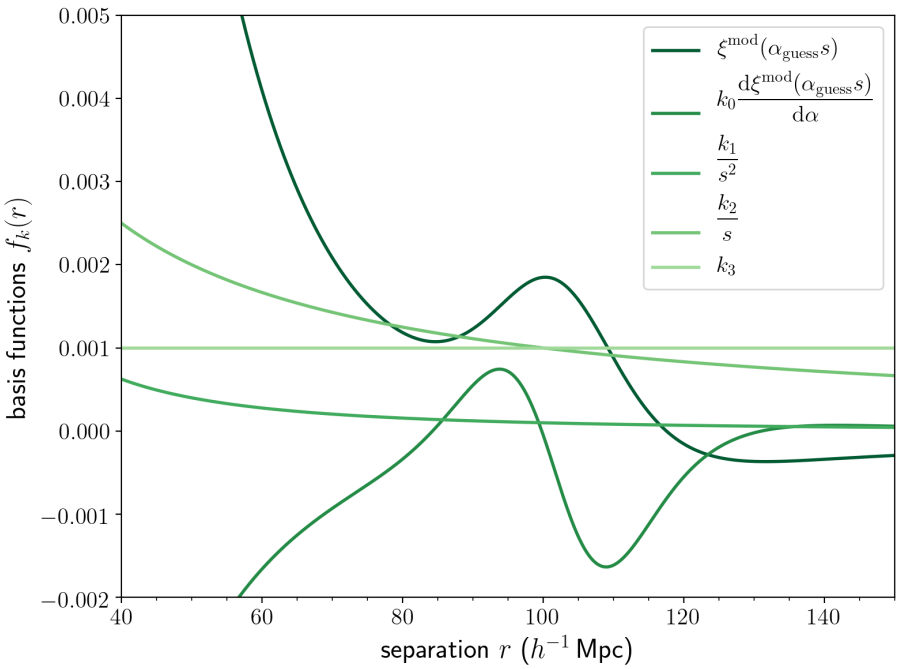

Figure 3. The set of basis functions used to fit for the BAO scale using our estimator. The B^2 term (green) is the fiducial model used to determine the scale dilation parameter α. The C term is the derivative of this model with respect to α, allowing for the estimation of this parameter. The a_1, a_2, and a_3 terms are nuisance parameters to fit the broadband shape. KSF says: Make colorblind friendly! try diff greens (change alpha?)

KSF says: fix latex; displaystyle KSF says: call to basis function legend in fig 4 caption

cosmology used for our mock catalogs, we can compute the true value of the scale dilation parameter, $\alpha_{\rm true} = 0.9987$. (Here our choice of fiducial model happened to be close to the true model, so our $\alpha_{\rm true}$ is very close to 1; this is typical, as our cosmological model is fairly well-constrained.) With this fiducial model, we can construct the basis functions for our estimator; these are shown (with $\alpha = 1$ and reasonable choices for the scaling parameters k) in Figure 3.4.

We apply our iterative estimation procedure to each of the 1000 mocks; the resulting estimate for the correlation function is shown in Figure 3. The mean BAO estimate is shown in orange, and the mean tophat estimate is in blue; the truth is in black. Our estimator clearly better represents the shape of the known 2pcf. The mean value of the final recovered scale dilation parameter is $\alpha = XXX \pm YYY$, very close to the true value.

We note that these basis functions are significantly different than the tophat or B-spline bases previously explored, mainly because they are not localized. This means that data at all scales could contribute to all basis functions. It is then critical to ensure that the final parameter estimate does not rely on the range of scales chosen. We have confirmed that in this application, the result is robust to the chosen range as long as the scales cover the range $40 < r < 200\ h^{-1}\,$Mpc, the typical range used in BAO analyses. KSF says: Actually have to check this and update numbers!

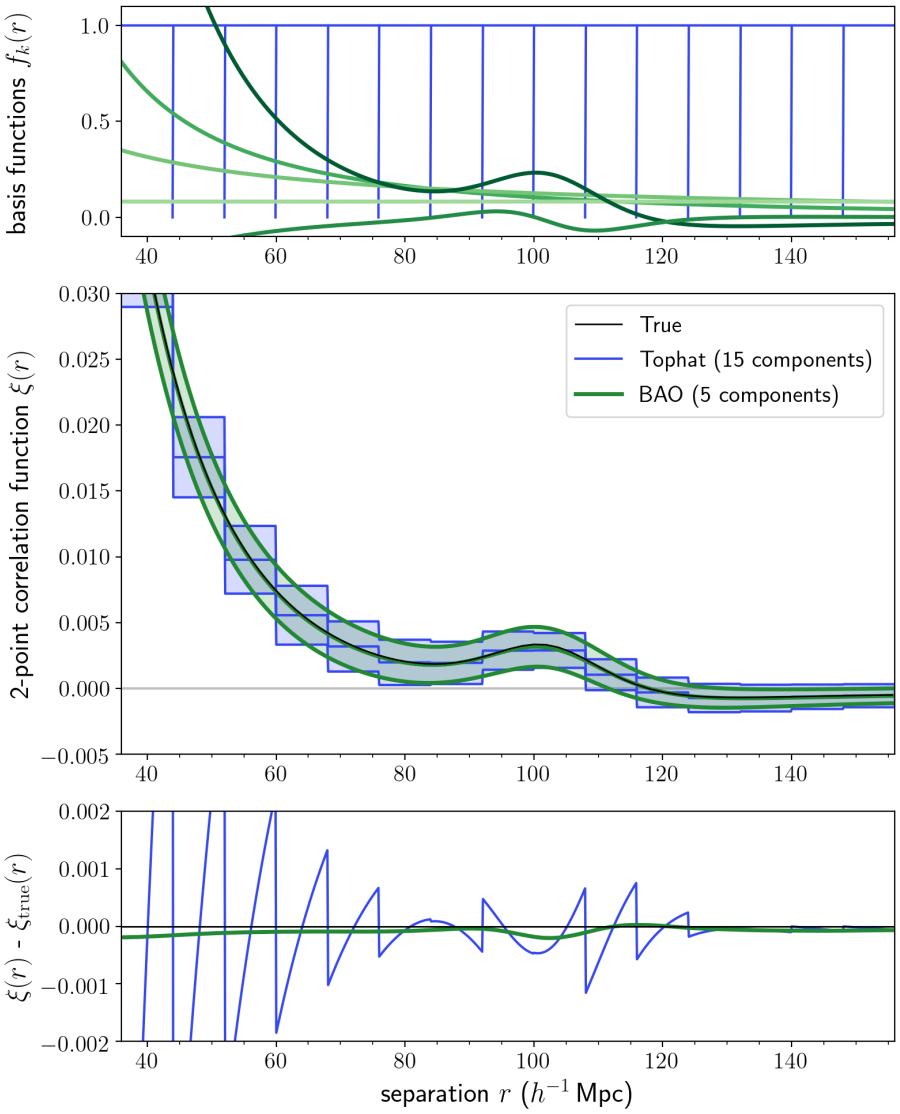

Figure 4. Estimation of the correlation function using our estimator with basis functions based on the BAO fitting function (orange dot-dashed). The line is the mean of the final estimate from the iteration procedure for 1000 mocks, and the shaded region is the 1σ variation. We also show the standard Landy-Szalay estimator, displayed as a tophat function (blue), as well as the true input correlation function (black). KSF says: This is for the 1e-4 box, need to run bigger box. KSF says: Should this and the basis set figure be a single figure with two panels? KSF says: Should i have a figure showing the sum of the basis functions and how that gets you the final cf?

4. DISCUSSION

4.1. *Relationship to Other Existing Estimators*

The Continuous-Function Estimator has properties similar to existing estimators, including kernel density estimators and the marked correlation function. With the proper choice of basis functions, the Continuous-Function Estimator can indeed produce both of these estimators; it is more general than either of them.

Kernel density estimation (KDE) is a class of methods for estimating a probability density function from a set of data. KDE methods essentially smooth the data with a given kernel, often a Gaussian. KSF says: should i cite the original KDE papers from the 50s and 60s? or not necessary bc not the focus? This is useful when we want to reconstruct a distribution without making many assumptions about the data, as is required in parametric methods. KDEs have found use in many areas of astrophyics, for example to measure the 21cm power spectrum with reduced foreground contamination (Trott et al. 2019), and to estimate luminosity functions with superior performance compared to binned methods (Yuan et al. 2020). KSF says: there are others, should i list more citations? without describing? Hatfield et al. (2016) uses a KDE approach to estimate the angular correlation function, in order to address the issues of information loss and arbitrary bin choice inherent to binning; they optimize for the kernel choice, and find a correlation function consistant with that of the binned method. KSF says: I couldnt find any other papers that use KDEs on correlation functions, but i found this one by chance. more to say here? the paper doesn't draw strong conclusions from this.

Specifically, kernel density estimators take the contribution of each data point to be a kernel function centered on that value, and sum these to determine the full distribution. In contrast, the Continuous-Function Estimator projects each data point onto fixed basis functions, which are distinct from the typical understanding of kernels. As such, our estimator is not smearing out the data, as KDEs do; it is using the data to directly infer the contribution of each basis function. This preserves the information in the data to the degree given by the chosen set of basis functions, which can in fact enhance features rather than smooth them. That said, the formulation of the Continuous-Function Estimator is general enough that it can perform a kernel density estimate of the correlation function, by choosing $f(r)$ to be a kernel centered on r. However, our uses here of the estimator are fundamentally different from KDE methods, as they use use fixed basis functions that can take advantage of the science use case and preserve maximal information from the data. KSF says: this last sentence is a bit repetitive / are there more differences im missing?

KSF says: Paragraph on marked cf

4.2. *Beyond the Landy-Szalay Estimator*

While we have formulated our estimator as a generalization of LS, as it is the standard used in 2pcf analyses and has optimal properties under certain conditions, we can also reformulate it for other estimators. Our formulation currently requires a normalization term (i.e. denominator) of the $\boldsymbol{v}_{\mathrm{RR}}$ counts, as we replace this with our $\boldsymbol{T}_{\mathrm{RR}}$ term. This is the case for the Peebles & Hauser (1974) (natural) estimator and the Hewett (1982) estimator:

$$\xi_{PH} = \frac{\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{RR}}}{\boldsymbol{v}_{\mathrm{RR}}} \to \boldsymbol{T}_{\mathrm{RR}}^{-1} \cdot (\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{RR}}) \tag{19}$$

$$\xi_{Hew} = \frac{\boldsymbol{v}_{\rm DD} - \boldsymbol{v}_{\rm DR}}{\boldsymbol{v}_{\rm RR}} \rightarrow \boldsymbol{T}_{\rm RR}^{-1} \cdot (\boldsymbol{v}_{\rm DD} - \boldsymbol{v}_{\rm DR}) \,. \tag{20}$$

We can also generalize estimators which have a $\boldsymbol{v}_{\rm DR}$ cross-correlation term as the denominator, such as the Davis & Peebles (1983) estimator,

$$\xi_{DP} = \frac{\boldsymbol{v}_{\rm DD} - \boldsymbol{v}_{\rm DR}}{\boldsymbol{v}_{\rm DR}} \rightarrow \boldsymbol{T}_{\rm DR}^{-1} \cdot (\boldsymbol{v}_{\rm DD} - \boldsymbol{v}_{\rm DR}) \tag{21}$$

by defining

$$\boldsymbol{T}_{\rm DR} = \frac{2}{N_{\rm D}\, N_{\rm R}} \sum_n \sum_m \boldsymbol{f}(\mathsf{G}_n, \mathsf{G}_m) \cdot \boldsymbol{f}^{\mathsf{T}}(\mathsf{G}_n, \mathsf{G}_m). \tag{22}$$

This formulation could be extended to nearly any linear combination of pair counts. The estimator of Vargas-Magaña et al. (2013) selects the optimal combination of pair counts; our estimators could be combined to create an even more generalized estimator. KSF says: some of the terms in V-M include DD in the denom - need to mention this? KSF says: what about estimator with DRsquared in the denom (hamilton estimator?) as trivial as naive guess? It can also be extended to cross-correlations in a straightforward way as expected.

4.3. *Computational Performance*

The computational scaling is by definition the same as traditional estimators, because pair-finding remains the limiting factor. The Continuous-Function Estimator has the additional need for evaluating the function f for each pair of galaxies. For simple basis functions like splines, this will only marginally decrease performance. For more complicated functions such as evaluating a cosmological model, the Continuous-Function Estimator may incur extra computational expense. Basis functions can also be input on a grid and then interpolated; the performance is then the same for all functions, but the interpolation for each function for each pair does somewhat decrease the performance.

KSF says: Is there a need to say much more? If not, maybe doesn't deserve its own section - could this go in the implementation section? Or maybe tacked onto another small section if we have one that makes sense KSF says: where to put other implemenation notes? e.g. using DD(s,mu) for realspace cf, with 1 big mu bin.

We don't actually take the inverse of the T matrix ...

4.4. *Effect on Covariance Matrix Estimation*

We have shown that the Continuous-Function Estimator results in 2pcf estimates that are just as accurate with fewer components. This is critical when estimating the covariance matrix, which is necessary for parameter inference. The covariance matrix is difficult to compute theoretically;instead, it is usually estimated by evaluating the 2pcf on a large number of mock catalogs and computing the covariance between the bins (e.g. Reid et al. 2010; Anderson et al. 2014). The unbiased estimator for the

sample covariance matrix is (e.g. Anderson 2003) KSF says: index notation getting clunky, thoughts?

$$\hat{\boldsymbol{C}}_{ij}^{\mathrm{ML}} = \frac{1}{N_{\mathrm{mocks}} - 1} \sum_{q=1}^{N_{\mathrm{mocks}}} \left([\boldsymbol{\xi}_q]_i - \bar{\boldsymbol{\xi}}_i \right) \left([\boldsymbol{\xi}_q]_j - \bar{\boldsymbol{\xi}}_j \right)^{\mathsf{T}}, \tag{23}$$

where q denotes the index of the mock, i and j denote the index of the bin or component, $\boldsymbol{\xi}$ denotes the estimate in that bin for that mock, and $\bar{\boldsymbol{\xi}}$ denotes the mean value of the estimate in that bin across the mocks, where we have omitted the hat for clarity. KSF says: i removed the hats here because they didn't play nicely w the average bar; thoughts? To get an unbiased estimate of the inverse covariance matrix, we require a correction factor, as the inverse of an unbiased estimator is not necessarily unbiased. The unbiased estimator for the sample inverse covariance matrix can be shown to be (Hartlap et al. 2007)

$$\hat{\boldsymbol{C}}^{-1} = \frac{N_{\mathrm{mocks}} - N_{\mathrm{bins}} - 2}{N_{\mathrm{mocks}} - 1} \left(\hat{\boldsymbol{C}}^{\mathrm{ML}} \right)^{-1}. \tag{24}$$

The variance in the elements of this estimator then have a dependence on N_{mocks} and N_{bins}. This propagates to the derived cosmological parameters, resulting in an overestimation of the error bars (Hartlap et al. 2007; Dodelson & Schneider 2013 Percival et al. 2014; Taylor & Joachimi 2014). Assuming that $N_{\mathrm{mocks}} >> N_{\mathrm{bins}}$ (and both much larger than the number of parameters to be estimated), and that the measurements are Gaussian distribued, the error bars are inflated by a factor of $(1 + N_{\mathrm{bins}}/N_{\mathrm{mocks}})$ (i.e., the true constraints are tighter than the derived ones). This factor becomes critical at the precision of cosmological parameter estimation (Percival et al. 2014).

Typically, this is dealt with by generating a very large number of mocks. For the Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al. 2013), ∼600 mocks were needed and the analysis used 41 bins (Sánchez et al. 2012). Future surveys will have more costly requirements on mock catalogs, with larger simulations necessary to cover the larger survey volumes.

An alternative to increasing N_{mocks} is decreasing N_{bins} to achieve the same error on precision. In the standard method, this is shown to *increase* the statistical error, albeit only slightly Percival et al. (2014). A substantial increase in bin width would prevent capturing information in finer clustering features; even the relatively broad BAO peak requires a bin size on the order of its width of $\sim 10h^{-1}$ Mpc. In fact, in the standard method more bins would typically be desireable, but the number is limited by the available number of mocks for covariance matrix computation.

With our estimator, we have shown that we can reduce the variance by using fewer components, without sacrificing accuracy. This means that we can safely reduce N_{bins}, or in our replacement of bins with continuous functions, the number of basis functions K. The covariance matrix will be the covariance between these basis

functions. KSF says: worth noting that the structure of this covmat will be significantly different, esp if non-orthogonal? To then achieve the same precision on the error on the cosmological parameters, a lower value of $N_{\rm mocks}$ becomes possible. This will significantly reduce requirements on mocks, which will be particularly important for upcoming large surveys. KSF says: cite extensively lit on enormous number of mocks needed to get covmat right. KSF says: I think the result of discussions was that there wasn't a good way of showing this without propagating all the way to cosmological parameters. Would love a figure showing lower covariance errors but not sure how without full propagation

4.5. *Further Applications*

The formulation of the Continuous-Function Estimator opens up many possibilities for extracting information from the correlation function. The most straightforward applications are standard basis functions or linearizeable astrophysical models, as we have shown here. Other applications for the direct estimation of cosmological parameters could include the growth rate of cosmic structure f (Satpathy et al. 2016; Reid et al. 2018) and primordial non-Gaussianity in the local density field f_{NL}^{local} (Karagiannis et al. 2014). KSF says: mention idea of doing full cosmo model analysis by taking derivs wrt cosmological params? cool but less connected to citeable papers perhaps

We can take our estimator a step further by choosing basis functions that depend not only on the separation between tracer pairs, but also on the properties of the tracers themselves. One such application is the redshift dependence of the Alcock-Paczynski effect Alcock & Paczynski (1979), which can be used to constrain the matter density Ω_m and the dark energy equation of state parameter w (Li et al. 2016). The basis functions f in this case would take the form

$$\boldsymbol{f}_k(\mathsf{G}_n, \mathsf{G}_{n'}) = \boldsymbol{f}_k(|\boldsymbol{r}_n - \boldsymbol{r}_{n'}|, z_n, z_{n'}), \tag{25}$$

where z is the redshift of tracer n or n'. Another potential use case is the luminosity and color dependence of galaxy clustering, which can be used to understand the relationship between galaxy formation and the LSS (Zehavi et al. 2011). This could be extended to other galaxy properties.

The estimator gives us the opportunity to investigate more subtle or exotic signals which are anomalous with respect to our conventional models. Anomalies could appear as inhomogeneities or anisotropies in the data. For example, cosmological parameters could vary across the sky, which has previously investigated in patches across the Cosmic Microwave Background (Mukherjee & Wandelt 2018). Another possibility is anisotropy in the cosmic acceleration, which could leave signatures in measurements made using various phenomena including baryon acoustic oscillations (Faltenbacher et al. 2012) and Type Ia supernovae (Colin et al. 2019). With our estimator, we could introduce a dependence on location or direction into our basis functions, and

constrain the potential deviation from homogeneity or isotropy. While these effects would be highly degenerate with systematics, our estimator combined with robust systematics mitigation allows us to investigate the possibility of new physics.

Finally, our estimator can be directly related to a power spectrum analysis. We could use a Fourier basis as our set of continuous functions. This would allow us to directly project the data onto Fourier modes. This represents a step towards unifying the correlation function and the power spectrum. KSF says: there's more to say here but I'm not sure what

5. SUMMARY

KSF says: TODO: write short summary

KSF was supported by the NASA FINESST grant [grant number] during the completion of this work. The authors thank Jeremy Tinker and Michael Blanton for helpful discussions, Roman Scoccimarro for insightful conversation, and the members of the Flatiron Astronomical Data Group for useful feedback. KSF would like to acknowledge significant code feedback and support from Manodeep Sinha, as well as Lehman Garrison. KSF thanks Drew Jamieson, Chris Lovell for helpful discussion... All of the code used in this paper is available open-source at `github.com/kstoreyf/Corrfunc` and `github.com/kstoreyf/continuous-estimator`.

APPENDIX

A. AFFINE INVARIANCE

The estimator is invariant under an affine transformation, meaning that the estimate of the 2pcf will be equivalent under rescalings or stretches. KSF says: check/reword this We represent the affine transformation by a transformation matrix $\boldsymbol{M}$ that modifies the basis functions $\boldsymbol{f}$, such that

$$\boldsymbol{f}' \leftarrow \boldsymbol{M}\,\boldsymbol{f} \tag{A1}$$

where the prime indicates our affine-transformed basis. KSF says: im also using primes for indices, looks a bit confusing; another notation for this? Then in the primed basis, the pair counts become KSF says: double sum? align w choice in sec 2 KSF says: equal or equivalent? (2 or 3 lines)

$$\boldsymbol{v}'_{\mathrm{DD}} = \sum_{nn'} \boldsymbol{f}'_{nn'} = \sum_{nn'} \boldsymbol{M}\,\boldsymbol{f}_{nn'} = \boldsymbol{M}\,\boldsymbol{v}_{\mathrm{DD}} \tag{A2}$$

$$\boldsymbol{v}'_{\mathrm{DR}} = \sum_{nm} \boldsymbol{f}'_{nm} = \sum_{nm} \boldsymbol{M}\,\boldsymbol{f}_{nm} = \boldsymbol{M}\,\boldsymbol{v}_{\mathrm{DR}} \tag{A3}$$

$$\boldsymbol{v}'_{\mathrm{RR}} = \sum_{mm'} \boldsymbol{f}'_{mm'} = \sum_{mm'} \boldsymbol{M}\,\boldsymbol{f}_{mm'} = \boldsymbol{M}\,\boldsymbol{v}_{\mathrm{RR}} \tag{A4}$$

where we use the shorthand $\boldsymbol{f}_{ij} = \boldsymbol{f}(\mathsf{G}_i, \mathsf{G}_j)$ and we have omitted the normalization factors for simplicity. In the last step, we have factored $\boldsymbol{M}$ out of the summation and written the primed vectors in terms of the unprimed vectors.

For the random-random tensor we have

$$\boldsymbol{T}'_{\rm RR} = \sum_{mm'} (\boldsymbol{M}\,\boldsymbol{f}_{mm'}) \cdot (\boldsymbol{M}\,\boldsymbol{f}_{mm'})^{\mathsf{T}} \tag{A5}$$

$$= \boldsymbol{M} \left[\sum_{mm'} \boldsymbol{f}_{mm'} \cdot \boldsymbol{f}^{\mathsf{T}}_{mm'} \right] \boldsymbol{M}^{\mathsf{T}} \tag{A6}$$

$$= \boldsymbol{M}\,\boldsymbol{T}_{\rm RR}\,\boldsymbol{M}^{\mathsf{T}} \tag{A7}$$

Then the amplitudes in the primed basis become

$$\boldsymbol{a}' = \boldsymbol{T}'^{-1}_{\rm RR} \cdot (\boldsymbol{v}'_{\rm DD} - 2\,\boldsymbol{v}'_{\rm DR} + \boldsymbol{v}'_{\rm RR}) \tag{A8}$$

$$\boldsymbol{a}' = [\boldsymbol{M}\boldsymbol{T}_{\rm RR}\boldsymbol{M}^{\mathsf{T}}]^{-1} \cdot [\boldsymbol{M}\,\boldsymbol{v}_{\rm DD} - 2\,\boldsymbol{M}\,\boldsymbol{v}_{\rm DR} + \boldsymbol{M}\,\boldsymbol{v}_{\rm RR}] \tag{A9}$$

$$= (\boldsymbol{M}^{\mathsf{T}})^{-1}\,\boldsymbol{T}^{-1}_{\rm RR}\,\boldsymbol{M}^{-1} \cdot \boldsymbol{M}\,[\boldsymbol{v}_{\rm DD} - 2\,\boldsymbol{v}_{\rm DR} + \boldsymbol{v}_{\rm RR}] \tag{A10}$$

$$= (\boldsymbol{M}^{\mathsf{T}})^{-1}\,\boldsymbol{T}^{-1}_{\rm RR} \cdot [\boldsymbol{v}_{\rm DD} - 2\,\boldsymbol{v}_{\rm DR} + \boldsymbol{v}_{\rm RR}] \tag{A11}$$

$$= (\boldsymbol{M}^{\mathsf{T}})^{-1}\,\boldsymbol{a} \tag{A12}$$

and the estimator $\hat{\boldsymbol{\xi}}'$ in the primed basis, where we are again using the shorthand $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}(\mathsf{G}_i, \mathsf{G}_j)$, is

$$\hat{\boldsymbol{\xi}}' = \boldsymbol{a}'^{\mathsf{T}} \cdot \boldsymbol{f} \tag{A13}$$

$$\hat{\boldsymbol{\xi}}' = [(\boldsymbol{M}^{\mathsf{T}})^{-1}\,\boldsymbol{a}]^{\mathsf{T}} \cdot (\boldsymbol{M}\,\boldsymbol{f}) \tag{A14}$$

$$= \boldsymbol{a}^{\mathsf{T}}\,[(M^{-1})^{\mathsf{T}}]^{\mathsf{T}} \cdot (\boldsymbol{M}\,\boldsymbol{f}) \tag{A15}$$

$$= \boldsymbol{a}^{\mathsf{T}}\,\boldsymbol{M}^{-1} \cdot \boldsymbol{M}\,\boldsymbol{f} \tag{A16}$$

$$= \boldsymbol{a}^{\mathsf{T}} \cdot \boldsymbol{f} \tag{A17}$$

$$= \hat{\boldsymbol{\xi}}. \tag{A18}$$

Thus after an affine transformation of the basis function, the resulting estimator is equivalent to the estimator in the original basis.

We note that this requires $\boldsymbol{M}$ be invertible. However, any two equivalent bases must be related by the inverse of a transformation matrix, so this requirement is already satisfied. KSF says: I don't understand what i wrote here. what more to say about this?

B. COMPUTING THE RANDOM-RANDOM TERMS ANALYTICALLY

The autocorrelation of the random catalog is meant to approximate the window function. When we have a periodic cube, we can compute this $\boldsymbol{v}_{\rm RR}$ term analytically. Here we derive this, and then derive the equivalent for our continuous-basis $\boldsymbol{v}_{\rm RR}$ and $\boldsymbol{T}_{\rm RR}$ terms.

Our goal is to estimate the number of pairs in a periodic cubic volume filled uniformly with tracers, $\boldsymbol{v}^{\rm ana}_{\rm RR}$. We first consider an annulus indexed by k around a single galaxy, with radial edges g_k and h_k. This annulus has a volume V_k. Taking

the box to have an average number density $\bar{n}$, the number of galaxies expected in the annulus is $N_k = V_k \bar{n}$, and thus our selected galaxy contributes N_k pairs to the count. KSF says: do i put mathspaces between multiplications when there's already a parenthesis? We do this for each of the $N_\mathrm{D} - 1$ other galaxies, and after including a factor of $\frac{1}{2}$ accounts for the fact that this double-counts pairs, we find a total pair count of $[\boldsymbol{v}_\mathrm{RR}^\mathrm{ana}]_k = \frac{1}{2}(N_\mathrm{D} - 1)N_k = \frac{1}{2}(N_\mathrm{D} - 1)V_k \bar{n}$. For a cubic volume, $\bar{n} = N_\mathrm{D}/L^3$, so our final pair count for the annulus is

$$[\boldsymbol{v}_\mathrm{RR}^\mathrm{ana}]_k = \frac{1}{2}\frac{N_\mathrm{D}}{L^3}(N_\mathrm{D} - 1)V_k. \tag{B19}$$

We next need to compute V_k; for hard-edged radial bins, we can compute V_k simply as the difference between spherical volumes. We can represent this more generally as an integral,

$$V_k = \int_{g_k}^{h_k} dV = 4\pi \int_{g_k}^{h_k} r^2 dr. \tag{B20}$$

We can easily generalize this to any basis function $\boldsymbol{f}_k(r)$ that is only a function of r,

$$V_k = 4\pi \int_{g_k}^{h_k} \boldsymbol{f}_k(r) r^2 dr \tag{B21}$$

where k is now the index of the basis functions. We can see that this reduces to Equation B20 when $\boldsymbol{f}(r)$ is the tophat function (returning 1 or 0 depending on whether or not r falls between g_k and h_k).

Combining the above equations gives us our full generalized analytic random-random vector $\boldsymbol{v}_\mathrm{RR}^\mathrm{ana}$, which has elements

$$[\boldsymbol{v}_\mathrm{RR}^\mathrm{ana}]_k = \frac{1}{2}\,\frac{N_\mathrm{D}}{L^3}\,(N_\mathrm{D} - 1)\,4\pi \int_0^{r_\mathrm{max}} \boldsymbol{f}_k(r)\, r^2\, dr \tag{B22}$$

where we are now integrating over all values of r we are interested in (from 0 out to some r_max), to account for non-localized basis functions.

Based on the definition of $\boldsymbol{T}_\mathrm{RR}$ in Equation 10 as the outer product of the basis function vector and its transpose, we can see that the elements of the analytic random-random tensor $\boldsymbol{T}_\mathrm{RR}^\mathrm{ana}$ can be written as

$$[\boldsymbol{T}_\mathrm{RR}^\mathrm{ana}]_{kk'} = \frac{1}{2}\,\frac{N_\mathrm{D}}{L^3}(N_\mathrm{D} - 1)\,4\pi \int_0^{r_\mathrm{max}} \boldsymbol{f}_k(r)\, \boldsymbol{f}_{k'}(r)\, r^2\, dr \tag{B23}$$

This could be further generalized to account for basis functions that take other properties as input.

When considering a periodic box, the natural estimator is no longer biased, so we can also avoid computing the cross-correlation term $\boldsymbol{v}_\mathrm{DR}$ and calculate the amplitudes as

$$\boldsymbol{a}_\mathrm{ana} = [\boldsymbol{T}_\mathrm{RR}^\mathrm{ana}]^{-1} \cdot \boldsymbol{v}_\mathrm{DD}. \tag{B24}$$

Looking back, it might have seemed strange that we use $N_{\rm D}$ in calculating the analytical term $\boldsymbol{v}_{\rm RR}^{\rm ana}$, but we now see that this normalization prefactor cancels out with that of the $\boldsymbol{v}_{\rm DD}$ term. Finally, we use these amplitudes $\boldsymbol{a}_{\rm ana}$ to compute the correlation function $\hat{\boldsymbol{\xi}}_{\rm ana}$ as before in Equation 12.

C. IMPLEMENTATION OF ESTIMATION WITH BAO BASIS FUNCTIONS

C.1. *Iterative Procedure*

The Continuous-Function Estimator can be used to measure the baryon acoustic oscillation (BAO) scale by choosing the basis functions to terms of a BAO fitting function, as described in 3.4. For this application, we need to choose a fiducial cosmology for our bases, which will be offset from the true cosmology. This offset can be encoded by a scale dilation parameter α, which contains the information about the BAO scale; see Equation 16. As our fitting function requires a fiducial model and an initial guess of this parameter, $\alpha_{\rm guess}$, and then determines the change needed, an iterative procedure is needed to converge to the best-fit value.

We start with assuming that we have chosen our fiducial model to match our true cosmology (we in all likelihood have not, but it's not a bad initial guess), giving us an initial $\alpha_{\rm guess} = 1.0$. We then apply the Continuous-Function Estimator to perform the measurement, and obtain the amplitude C for the derivative term in our model as in Equation 18. This gives us our estimate $\hat{\alpha}$ of the scale dilation parameter from this initial model; for the ith iteration, we have

$$\hat{\alpha}_i = \alpha_{{\rm guess},i} + C_i\, k_0 \tag{C25}$$

where k_0 is the chosen scaling parameter for the derivative basis function as in Equation 18.

We choose the convergence criterion to be when the fractional change in $\hat{\alpha}$ between subsequent iterations falls below a threshold, $c_{\rm thresh}$ KSF says: name for this variable?,

$$\left| \frac{\hat{\alpha}_i - \hat{\alpha}_{i-1}}{\hat{\alpha}_i} \right| < c_{\rm thresh}. \tag{C26}$$

For our application we choose $c_{\rm thresh} = 0.00001$.

To achieve convergence, we need to be careful in choosing our next $\alpha_{{\rm guess},i}$. If it is far from the best estimate, C_i will be large, and our resulting estimate $\hat{\alpha}_i$ will be inaccurate. We thus include a damping parameter η between 0 and 1 to improve our convergence. Our next guess is then KSF says: what should the assignment symbol be here?

$$\alpha_{{\rm guess},i+1} \leftarrow \alpha_{{\rm guess},i} + \eta\, C_i\, k_0. \tag{C27}$$

The choice of η is important for stability and speed of convergence; too large a value can lead to a back-and-forth cycle in which the result hops between two values and never converges, and too small a value would make convergence take a very long time. In our application, we start with $\eta = 0.5$. We check if our estimate is jumping over

the true value by checking if the error changes sign; if it does, we reduce η by a factor of 0.75.

C.2. *Implementation Details*

We implement the partial derivative in the fitting function as a finite difference between model with the our chosen value of $\alpha_{\rm guess}$, and the model with a value shifted by a small $\Delta\alpha$, KSF says: what should the assignment symbol be here?

$$\frac{\mathrm{d}\xi^{\rm mod}(\alpha s)}{\mathrm{d}\alpha} \leftarrow \frac{\xi^{\rm mod}(\alpha_{\rm guess}s) - \xi^{\rm mod}((\alpha_{\rm guess}+\Delta\alpha)s)}{\Delta\alpha}. \tag{C28}$$

In our implemenation we take $\Delta\alpha = 0.001$; we check that our results are insensitive to this choice.

We choose the amplitudes of the basis functions k to set them at similar scales. We use $k_0 = 0.1$, $k_1 = 10.0$, $k_2 = 0.1$, and $k_3 = 0.001$.

REFERENCES

Alam, S., Ata, M., Bailey, S., et al. 2016, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample, Tech. rep. https://arxiv.org/abs/1607.03155v1

Alcock, C., & Paczynski, B. 1979, An evolution free test for non-zero cosmological constant, Tech. rep.

Anderson, L., Aubourg, E., Bailey, S., et al. 2011, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations in the Data Release 9 Spectroscopic Galaxy Sample, Tech. rep. https://arxiv.org/abs/1203.6594v1

Anderson, L., Aubourg, É., Bailey, S., et al. 2014, Monthly Notices of the Royal Astronomical Society, 441, 24, doi: 10.1093/mnras/stu523

Anderson, T. 2003, An Introduction to Multivariate Statistical Analysis, doi: 10.1080/00401706.1986.10488123

Bailoni, A., Spurio Mancini, A., Amendola, L., et al. 2016, Improving Fisher matrix forecasts for galaxy surveys: window function, bin cross-correlation, and bin redshift uncertainty, Tech. rep. https://arxiv.org/abs/1608.00458v3

Baxter, E. J., & Rozo, E. 2013, Astrophysical Journal, 779, 15, doi: 10.1088/0004-637X/779/1/62

Coles, P., & Jones, B. 1991, Monthly Notices of the Royal Astronomical Society, 248, 1, doi: 10.1093/mnras/248.1.1

Colin, J., Mohayaee, R., Rameez, M., & Sarkar, S. 2019, Astronomy and Astrophysics, 631, doi: 10.1051/0004-6361/201936373

Davis, M., & Peebles, P. J. E. 1983, The Astrophysical Journal Supplement Series, 267, 465, doi: 10.1086/190860

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, Astronomical Journal, 145, 55, doi: 10.1088/0004-6256/145/1/10

Demina, R., Cheong, S., BenZvi, S., & Hindrichs, O. 2016, MNRAS, 480, 49, doi: 10.1093/mnras/sty1812

Dodelson, S., & Schneider, M. D. 2013, Physical Review D - Particles, Fields, Gravitation and Cosmology, 88, doi: 10.1103/PhysRevD.88.063537
Eisenstein, D. J., & Hu, W. 1997, The Astrophysical Journal, 496, 605, doi: 10.1086/305424
Eisenstein, D. J., Seo, H.-J., Sirko, E., & Spergel, D. N. 2007, The Astrophysical Journal, 664, 675, doi: 10.1086/518712
Faltenbacher, A., Li, C., & Wang, J. 2012, Astrophysical Journal Letters, 751, doi: 10.1088/2041-8205/751/1/L2
Grimmett, L. P., Mullaney, J. R., Bernhard, E. P., et al. 2020, MNRAS, 000, 1. https://arxiv.org/abs/2001.11573
Hamilton, A. J. S. 1988, The Astrophysical Journal, 331, L59, doi: 10.1086/185235
—. 1993, Astrophysical Journal, 417, 19
Hartlap, J., Simon, P., & Schneider, P. 2007, Astronomy and Astrophysics, 464, 399, doi: 10.1051/0004-6361:20066170
Hatfield, P. W., Lindsay, S. N., Jarvis, M. J., et al. 2016, Monthly Notices of the Royal Astronomical Society, 459, 2618, doi: 10.1093/mnras/stw769
Hewett, P. C. 1982, Monthly Notices of the Royal Astronomical Society, 201, 867, doi: 1982MNRAS.201..867H
Kaiser, N. 2014, Monthly Notices of the Royal Astronomical Society, 227, 1, doi: 10.1093/mnras/227.1.1
Karagiannis, D., Shanks, T., & Ross, N. P. 2014, Monthly Notices of the Royal Astronomical Society, 441, 486, doi: 10.1093/mnras/stu590
Kazin, E. A., Blanton, M. R., Scoccimarro, R., et al. 2010, Astrophysical Journal, 710, 1444, doi: 10.1088/0004-637X/710/2/1444
Kerscher, M. 1999, Astronomy and Astrophysics, 343, 18. https://arxiv.org/abs/9811300
Kerscher, M., Szapudi, I., & Szalay, A. 2000, The Astrophysical Journal, 535, L13, doi: 10.1086/312702
Landy, S. D., & Szalay, A. S. 1993, The Astrophysical Journal, 412, 64
Lanzuisi, G., Delvecchio, I., Berta, S., et al. 2017, Astronomy and Astrophysics, 602, doi: 10.1051/0004-6361/201629955
Li, X.-D., Park, C., Sabiu, C. G., et al. 2016, The Astrophysical Journal, 832, 1, doi: 10.3847/0004-637X/832/2/103
Mukherjee, S., & Wandelt, B. D. 2018, Journal of Cosmology and Astroparticle Physics, doi: 10.1088/1475-7516/2018/01/042
Peebles, P. J. E., & Hauser, M. G. 1974, The Astrophysical Journal Supplement Series, 28, 19, doi: 10.1086/190308
Percival, W. J., Ross, A. J., Sánchez, A. G., et al. 2014, Monthly Notices of the Royal Astronomical Society, 439, 2531, doi: 10.1093/mnras/stu112
Reid, B. A., Seo, H.-J., Leauthaud, A., Tinker, J. L., & White, M. 2018, A 2.5% measurement of the growth rate from small-scale redshift space clustering of SDSS-III CMASS galaxies, Tech. Rep. 0000. https://arxiv.org/abs/arXiv:1404.3742v2
Reid, B. A., Percival, W. J., Eisenstein, D. J., et al. 2010, Monthly Notices of the Royal Astronomical Society, 404, 60, doi: 10.1111/j.1365-2966.2010.16276.x
Sánchez, A. G., Scóccola, C. G., Ross, A. J., et al. 2012, Monthly Notices of the Royal Astronomical Society, 425, 415, doi: 10.1111/j.1365-2966.2012.21502.x
Satpathy, S., Alam, S., Ho, S., et al. 2016, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: On the measurement of growth rate using galaxy correlation functions, Tech. rep. https://arxiv.org/abs/1607.03148v2
Sinha, M., & Garrison, L. H. 2019, MNRAS, 000, 1. https://arxiv.org/abs/1911.03545
Taylor, A., & Joachimi, B. 2014, Monthly Notices of the Royal Astronomical Society, 442, 2728, doi: 10.1093/mnras/stu996

Trott, C. M., Fu, S. C., Murray, S. G., et al. 2019, Monthly Notices of the Royal Astronomical Society, 486, 5766, doi: 10.1093/mnras/stz1207

Vargas-Magaña, M., Bautista, J. E., Hamilton, J.-C., et al. 2013, Astronomy & Astrophysics, 554, A131, doi: https://doi.org/10.1051/0004-6361/201220790

White, M., & Padmanabhan, N. 2009, Mon. Not. R. Astron. Soc, 395, 2381, doi: 10.1111/j.1365-2966.2009.14732.x

Yuan, Z., Jarvis, M. J., & Wang, J. 2020, The Astrophysical Journal Supplement Series, 248, 1, doi: 10.3847/1538-4365/ab855b

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, Astrophysical Journal, 736, doi: 10.1088/0004-637X/736/1/59