Draft version October 23, 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; this 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 with a choice of cubic spline basis functions can perform an estimate of the clustering of artificial data that better represents the smoothness and shape of the 2pcf compared to LS. We further demonstrate that the estimator can be used to directly estimate the Baryon Acoustic Oscillation scale, using a small number of scientifically motivated basis functions. Critically, the reduction in the number of basis functions will lead to a reduction in the number of mock catalogs required for covariance estimation; this is 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 investigating potential inhomogeneities or anisotropies in clustering.

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, Budavari et al. 2003, Li et al. 2006, Abbas & Sheth 2006, Zehavi et al. 2011, Skibba et al. 2014, Durkalec et al. 2018). KSF says: do i not need all these citations here? do i need more elsewhere in the intro? 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). It 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. The 2pcf is the primary tool for extracting cosmological information from galaxy redshift surveys. Such correlation function analyses include Hawkins et al. (2003) for the 2dF Galaxy Redshift Survey (2dFGRS, Colless et al. 2001), Alam et al. (2016) for the Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al. 2013) DR12 analysis, and Elvin-Poole et al. (2017) for the Dark Energey Survey (DES, DES Collaboration 2005).

Estimators of the 2pcf have been studied extensively (e.g. Peebles & Hauser 1974; Davis & Peebles 1983; Hamilton 1993). Traditionally, the 2pcf is 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. Additionally, the error on the inverse covariance matrix estimate depends on the number of bins, with a larger number of

bins resulting in a larger error that propogates to the estimated parameters (Hartlap et al. 2007; Percival et al. 2014). This can be balanced by using a large number of mock galaxy catalogs, but these are exceedingly expensive to generate. This is currently the limiting step in LSS analyses, with on the order of 1000 mock catalogs tailored to the survey needed to achieve the desired precision on the parameters. The requirements on the covariance matrix will get even more stringent as survey size increases and we push towards higher precision; the connection of this liming step with bin choice merits scrutiny of binning in 2pcf analyses.

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 the effects of bin width are small but non-negligible; they showed that there is an optimal bin width given the analysis method that balances the statistical error with biasing the derived BAO peak location. From this literature, it is clear that, when analyzing smooth quantities such as LSS statistics, binning is sinning.

One of the difficulties in performing a two-point estimate is that nontrivial survey boundaries would bias a direct summation of pair counts. To account for the boundaries as well as regions corrupted by issues such as bright foreground stars, typically a large 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. The current standard estimator, proposed by Landy & Szalay (1993) (hereafter LS), takes this approach. It involves a summation of the data–data pairs DD in each separation bin, normalized by random–random pairs RR as well as the data–random pairs DR to improve the bias properties of the estimator. The LS estimator of the correlation function $\hat{\xi}_k$ for the k^{th} bin in separation r is defined as

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

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 traditional 2pcf estimation 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 *Article*, 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 directly 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 tophat basis functions, the estimator exactly reduces to the LS estimator. The Continuous-Function Estimator removes the need for binning and produces a more representative estimate of the 2pcf with fewer basis functions, reducing requirements on mock catalogs for covariance matrix computation. 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 compared to standard analyses.

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

2. MOTIVATION AND FORMULATION

In this *Article*, 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.

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 1. 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 $\boldsymbol{v}_{\rm DD}$. We can then write the LS estimator as

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

The components of the pair-count vectors are defined explicitly as

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

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

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

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. We assume here that the sums are over unique pairs, and that for auto-correlations they exclude self-pairs; the normalization prefactors then account for the total number of possible pairs, explaining the difference between the auto- and cross-correlation factors. 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 *Article* we consider the isotropic real-space 2pcf for simplicity, but the estimators detailed here apply equally well to these alternative configurations. The estimator is also easily applicable to cross-correlations of two data sets.

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*

Clustering estimation 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

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

where $\boldsymbol{x}$ is the vector of best-fit parameters, $\boldsymbol{A}$ is a design matrix containing 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 factor $[\boldsymbol{A}^\mathsf{T}\boldsymbol{C}^{-1}\boldsymbol{y}]$ projects the data onto the features (as in a matched filter). The first bracketed factor $[\boldsymbol{A}^\mathsf{T}\boldsymbol{C}^{-1}\boldsymbol{A}]^{-1}$ rescales the projected features into the space of the parameters.

Standard two-point correlation function estimators are effectively performing such a projection: each bin is the projection of the data pair counts onto the radial separation annulus, and the random-random term rescales this feature. The analogy is clear in the so-called natural estimator of the 2pcf, $\hat{\boldsymbol{\xi}} = \boldsymbol{v}_\mathrm{DD}/\boldsymbol{v}_\mathrm{RR} - 1$ (e.g. Kerscher et al. 2000), with $\boldsymbol{v}_\mathrm{DD}$ aligning with the second bracketed factor and $\boldsymbol{v}_\mathrm{RR}$ aligning with the first (the division can be written as an inverse factor). The Continuous-Function Estimator was inspired by this connection with least-squares fitting; we detail its formulation in the following section.

2.3. *The Continuous-Function Estimator*

We generalize the LS estimator defined above in Equations 3-5 by analogy with least-squares fitting. 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 projections of the basis functions, $\boldsymbol{v}$. These projection vectors are defined as

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

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

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

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

where the summation notation and prefactors are the same as explained for the standard LS estimator.

We can now define the Continuous-Function Estimator as

$$\hat{\xi}(\mathsf{G}_l, \mathsf{G}_{l'}) \equiv \boldsymbol{a}^\mathsf{T} \cdot \boldsymbol{f}(\mathsf{G}_l, \mathsf{G}_{l'}) \ , \tag{11}$$

KSF says: on bolding ξ and f (aka whether they are vectors): i think here it should be unbolded, because this formula will give a single value for the correlation function given those galaxy properties. it was bold before. i think when we are just quoting $\hat{\boldsymbol{\xi}}$ it's bold, but when it's a funciton of anything else it's e.g. $\hat{\xi}(r)$. agree? where $\boldsymbol{a}$ is a K-vector of the computed *amplitudes* of the basis functions

$$\boldsymbol{a} \equiv \boldsymbol{T}_{\rm RR}^{-1} \cdot (\boldsymbol{v}_{\rm DD} - 2\,\boldsymbol{v}_{\rm DR} + \boldsymbol{v}_{\rm RR}) \tag{12}$$

and G_l and $\mathsf{G}_{l'}$ contain the data values at which to evaluate $\hat{\boldsymbol{\xi}}$. 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 evaluate $\hat{\boldsymbol{\xi}}$. In the rest of this *Article*, however, we will only take into account the separation between pairs, and we will write $\hat{\xi}(r)$.

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

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

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 ($\boldsymbol{v}_{\rm D_1R_2}$, $\boldsymbol{v}_{\rm D_2R_1}$, $\boldsymbol{v}_{\rm R_1R_2}$, $\boldsymbol{T}_{\rm R_1R_2}$) generalize as one would expect. The amplitudes then become

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

and we use this to compute the estimator as in Equation 11.

If we consider only the pair separation, and make a proper choice of $\boldsymbol{f}$, the Continuous-Function Estimator reduces to the LS 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 tophat (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 $\boldsymbol{v}_{\rm DD}$, $\boldsymbol{v}_{\rm DR}$ and $\boldsymbol{v}_{\rm RR}$ projection vectors simply become binned pair counts, with bin edges g_k and h_k as before. The $\boldsymbol{T}_{\rm RR}$ tensor becomes diagonal, with its diagonal elements equal to the elements of the $\boldsymbol{v}_{\rm RR}$ vector. 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 tophat 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 $\boldsymbol{v}_{\rm RR}$ term, as well as the $\boldsymbol{T}_{\rm RR}$ term, and use the natural form of the 2pcf estimator. The derivation and formulation of these terms are shown in Appendix B.

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`. Implementation details are discussed in Section 4.4. KSF says: should this whole paragraph (or more of it) be in the implementation section vs here?

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 (Agrawal et al. 2017). We use an input power

spectrum with the Planck cosmology, the same parameters used for the MultiDark–PATCHY simulations (Kitaura et al. 2016) 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 $2 \times 10^{-4}\ h^3\,\mathrm{Mpc}^{-3}$. 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. The results will hold for catalogs with realistic survey windows and random–random terms computed directly with the Continuous-Function Estimator.

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 $36 < r < 156\ h^{-1}\,\mathrm{Mpc}$, each with a width of $8\ h^{-1}\,\mathrm{Mpc}$; 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; the error bars show the standard deviation of the 1000 mocks in each bin. We also show the true input correlation function, 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. 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*

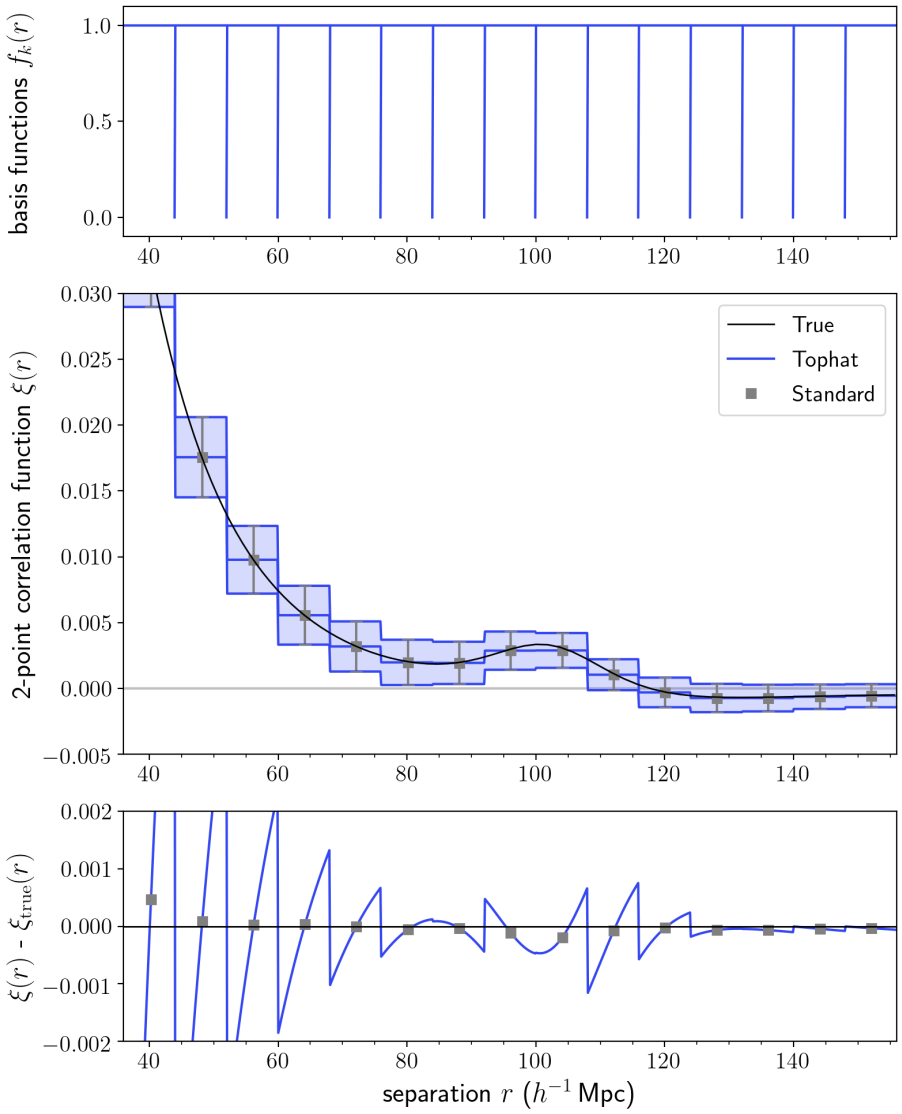

Figure 1. A comparison between the Continuous-Function Estimator with a tophat basis (thin blue lines) and the standard estimator (grey squares). The top panel shows the basis functions used for the tophat estimator. The middle panel shows the mean of the estimated correlation functions for 1000 mock 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.

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 (DeBoor 1987). They have the nice property that the functions and their derivatives can be continuous, 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

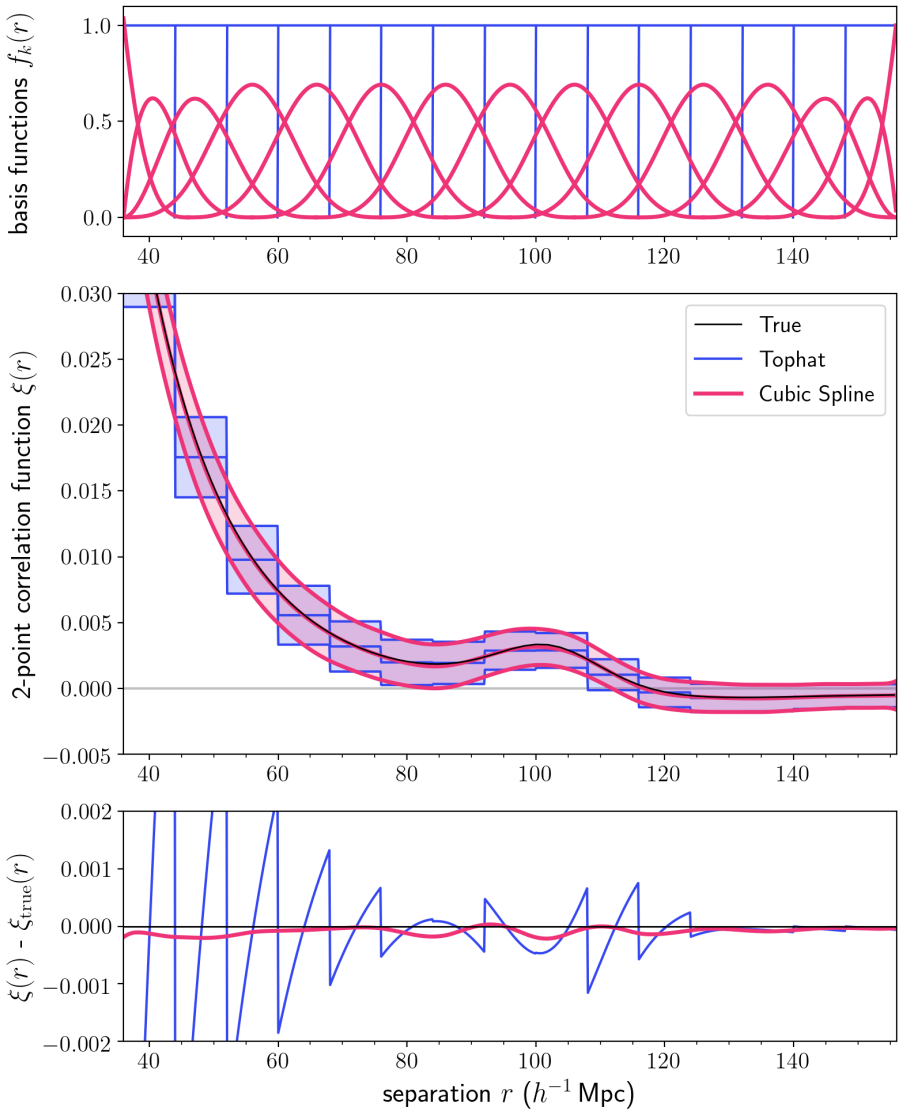

Figure 2. A comparison between the Continuous-Function Estimator with a cubic spline basis function (thick red) and a standard tophat basis (thin blue). The top panel shows the basis functions used for each measurement. The middle panel shows the mean of the estimated correlation functions for each of 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.

constitute the set of basis functions for a cubic spline, as they are the lowest-order spline to have a continuous first derivative.

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. For the cubic spline basis, we use the same r-range and number of basis functions, and knots chosen to evenly span the range. 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 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 mostly localized yet representative estimate of the 2pcf, or 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*

The measurement of the baryon acoustic oscillation (BAO) scale provides an apt use case for our estimator. The BAO feature is a peak in clustering on large scales, $\sim$150 Mpc ($\sim 100h^{-1}$ 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 (Eisenstein et al. 2005; Kazin et al. 2010; Anderson et al. 2011, 2014; Alam et al. 2016).

We base our BAO analysis on the method of the BOSS DR10 and 11 analysis (Anderson et al. 2014). We estimate the spherically averaged 3-dimensional correlation function, $\hat{\xi}(r)$, where r is the separation between pairs. (BAO analyses are typically done in redshift space, estimating $\hat{\xi}(s)$, where s is the redshift-space separation between pairs, but here we are using 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, defined as

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

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

denotes the value for the chosen fiducial model (the non-superscripted parameters are the true values). 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. With isotropic analyses, there is a degeneracy between D_A and H, so typically a combination of these values is reported; the degeneracy can be broken with anisotropic BAO analyses. Our estimator could straightforwardly perform an estimate of the anisotropic correlation function, but for demonstration purposes we perform an isotropic analysis here and focus on the recovered value of α.

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

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

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: KSF is figuring out how this actually done and flesh out this paragraph! 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 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 α_guess, determine the change in this value needed to improve the fit. Our fitting function is then

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

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 magnitude 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. The adopted k_i values are noted in Appendix C.

To use the estimator for a BAO measurement, we input these five terms as the five basis functions of our estimator. The estimator outputs an amplitude vector $\boldsymbol{a}$ as described in Section 2.3, which describes the contribution of each basis function—precisely the values of the free parameters, scaled by k_i. From the value of C, we can determine our estimate of the scale dilation parameter, $\hat{\alpha}$, as $\hat{\alpha} = \alpha_\mathrm{guess} + C\, k_0$, based on the definition of finite derivatives. With this formulation, a value of $C = 0$ indicates that the current α_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 to optimally fit the data. 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.

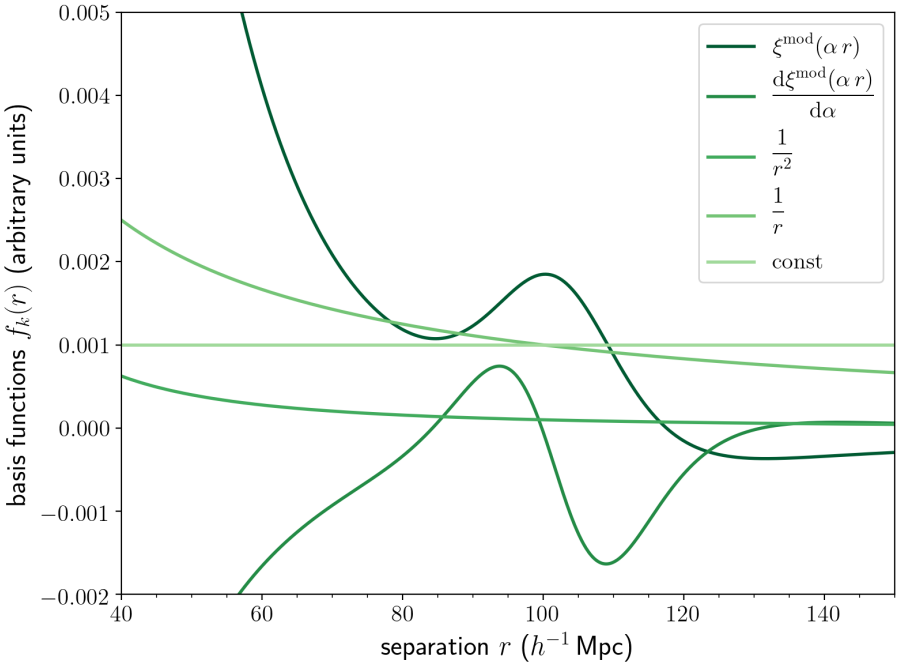

Figure 3. The set of basis functions used to fit for the BAO scale using our estimator. The $\xi^{\rm mod}(\alpha\, r)$ term (darkest green) is the correlation function computed using fiducial model, with some scale dilation α. The derivative term (second-to-darkest green) is the derivative of this model with respect to α, which allows for the direct estimation of this parameter. The other three terms (lighter greens) are nuisance parameters to fit the broadband shape.

We demonstrate this method using our set of lognormal mock catalogs. We construct a recovery test following that in Hinton et al. (2019). We assume the fiducial cosmological model used in Beutler et al. (2017): $\Omega_{\rm m} = 0.31$, $h = 0.676$, $\Omega_{\rm b} = 0.04814$, $n_s = 0.97$. As we know the 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 arbitrary scaling) in Figure 3.

We apply our iterative estimation procedure to each of the 1000 mocks; the mean of the resulting estimates for the correlation function is shown in Figure 3. We show the BAO basis functions in the top panel, as in Figure 3. We compare the Continuous-Function Estimator using the BAO bases with that using tophat bases, as in the previous sections. The correlation function estimated with the BAO bases clearly produces a more representative estimate at all scales. The estimate is also smoother than that

produced using the cubic spline basis functions (Figure 2). More importantly, it is scientifically motivated: the estimator directly gives us the relative contributions of the terms of the BAO fitting functions. Further, the BAO-based estimate requires only five components, while the tophat basis requires 15 components (or bins, in the standard approach) in the same scale range. This is critical for the efficient computation of a precise covariance matrix, as the errors depend on the number of components used for the estimate, as described in Section 4.5. The Continuous-Function Estimator with the BAO basis functions could reduce the number of mocks needed to achieve the same precision by a factor of a few to an order of magnitude; as these expensive cosmological simulations are currently the limiting step in two-point analyses, this could be highly impactful.

We note that these basis functions are significantly different than the tophat or B-spline bases previously explored. One important difference is that they are not localized. This means that data at all scales could contribute to all basis functions. It is then critical for the investigator to ensure that the range of scales chosen is reasonable for the scientific problem at hand and that the final estimate of the parameter of interest does not depend on the details of this choice. The nonlocality also means that the covariance matrix between the components will have a very different structure than typical binned covariance matrices, as all of the components may be highly covariant. One option is to apply a decorrelation transformation to the set of bases in order to suppress the off-diagonal terms, and use these as the basis functions for the estimator. This issue is not unique to our estimator, though, as uncertainties with the standard estimator are already highly correlated. We address this further in Section 4.5. KSF says: anything more reassuring to say here? something about how it's fine if theyre covariant as long as this is appropriately taken into account in analyses (how would one do this)?

KSF says: I will flesh out the following paragraph once I perform the traditional method! The Continuous-Function Estimator with the BAO bases gives us an estimate for α for each of the 1000 mock catalogs. These are shown in comparison with the values estimated using the traditional fitting approach in Figure ZZZ. KSF says: update The methods produce similar values of α ... The mean estimate of the recovered scale dilation parameter is $\alpha = XXX \pm YYY$, very close to the true value of $\alpha = 0.9987$. The mean estimate using the traditional approach is $\alpha = XXX \pm YYY$. Our estimate is more accurate and precise ...

4. DISCUSSION

4.1. *Summary of Results and Limitations*

This *Article* presented a method to estimate the two-point correlation function without the need for binning. It generalizes the standard 2pcf estimator in a manner inspired by least-squares fitting: the Continuous-Function Estimator projects the data onto a set of user-chosen basis functions, and applies a normalization based

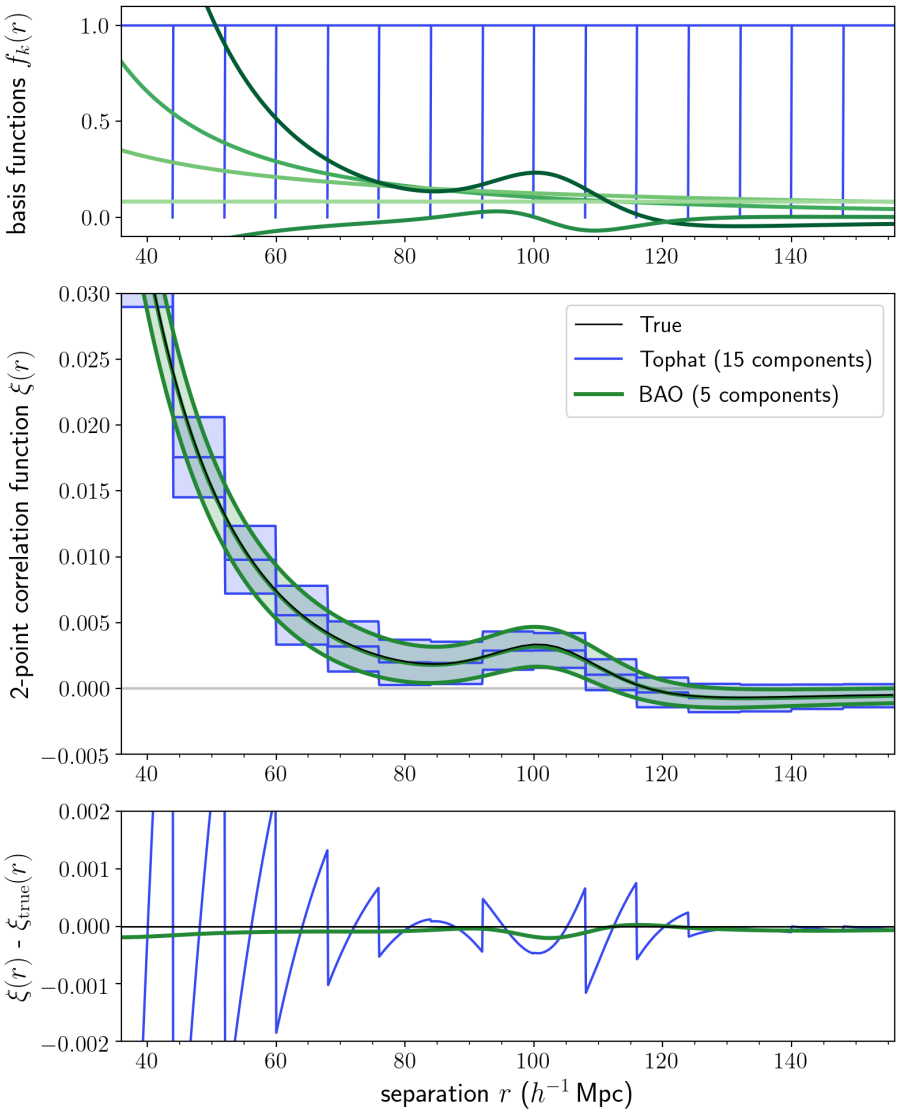

Figure 4. Estimation of the correlation function using our estimator with basis functions based on the BAO fitting function (orange dot-dashed). The line is the mean of the final estimate from the iteration procedure for 1000 mocks, and the shaded region is the 1σ variation. We also show the standard Landy–Szalay estimator, displayed as a tophat function (blue), as well as the true input correlation function (black). The basis functions are the same as those shown in Figure 3, with arbitrary scaling.

on a random catalog. Using a set of artificial mock catalogs, we showed that our method exactly reproduces the results of the standard approach with the choice of tophat basis functions, but in a way that demonstrates what the estimator is really measuring, namely a constant amplitude of clustering at every point within a radial separation bin. The Continuous-Function Estimator, however, has the capacity to be much more expressive than the standard estimator with a different choice of basis functions. We demonstrated its use with cubic spline basis functions, which results in a correlation function estimate that ios much more representative of the expected

shape and smoothness of the 2pcf. Further, the Continuous-Function Estimator can be tailored to the scientific use case; we apply to it a toy baryon acoustic feature analysis, choosing basis functions to be the terms of a modified BAO fitting function. This produced an estimate of the 2pcf that inherently reflects our beliefs about its form, and allowed us to directly estimate the scale dilation parameter with improved accuracy and precision compared to the standard estimator. KSF says: UPDATE this last line when analysis is done.

KSF says: this paragraph requires Hogg eyes! While these are just a few example applications of our estimator (for others see Section 4.6), they demonstrate the problems solved and opportunities created with the Continuous-Function Estimator. In getting rid of the need for binning, the estimator can produce a much more *representative* correlation function. For the 2pcf, binning is a very poor representation of the data; this extends to nearly all science cases. We almost always expect some sort of continuity in the functions we work with; nature does not bin. In many cases, the appropriate approach would be to switch to modeling the process. KSF says: I am not clear how this directly solves the problem at hand; how do you compare the models to data in this case? However, this is not always feasible; another "correct" approach is to work in the limit of infinitesimal bins, though how to do this is not intuitive. In our method, we effectively work in this limit by exploiting the equivalence between infinitesimal bins and continuous basis functions, taking the exact location of each datapoint and projecting it onto the bases. KSF says: is this true? how to phrase properly? This allows us to preserve our expectations about the data and still produce a workable estimate that is more informative about the science we are interested in.

We note that our approach does not help with the *choice* issue inherent in selecting bins. In fact, the Continuous-Function Estimator expands the choice infinitely, as it is much more expressive across multiple axes. The investigator must now choose the set of basis functions from a much larger space, which involves a choice of which information about the galaxies the bases should depend on. There is more room here for arbitrary choice issues and adversarial selections that produce particular results. That said, with a selection of basis functions designed for the science case at hand, the choices can be better motivated and lead directly to the final estimate one is interested in.

Our approach does have certain limitations. One restriction is the fact that the desired function KSF says: model? must be representable in terms of a linear combination of basis functions; this makes it difficult to use with more complex models. The estimator must also evaluate the basis functions for every pair of objects, so highly complex functions would become intractable. Finally, our estimator inherits many of the limitations of the Landy–Szalay estimator, including the fact that it contains a bias on clustered data and it has non-optimal variance properties. However, the the Continuous-Function Estimator is further generalizable to other forms of the estima-

tor (Section 4.3), so in principle there is a formulation that further improves upon these limitations. KSF says: More limitations here?

4.2. *Relationship to Other Modifications of 2pcf Estimators*

The Continuous-Function Estimator is a particular generalization of standard two-point clustering estimation that is well-motivated and unbiased under certain contexts. KSF says: i think the last part is true as much as LS is, but perhaps i shouldnt claim it here - or i need to explain further / back it up. On the surface, however, it appears similar to other two-point function projects, including kernel density estimators and the marked correlation function. While these formulations look similar, ours solves a different problem than these estimators.

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. 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). 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.

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. More importantly, the Continuous-Function Estimator hinges on the $\boldsymbol{T}_{\mathrm{RR}}^{-1}$ term that rescales the data projection, resulting in a estimate that is closely related to a least-squares fit of the data. This is very different from KDE methods, which in most contexts result in a biased estimate. KSF says: too harsh? i think if i say this i need to discuss the bias properties of mine...

Another method that shares similarities with the Continuous-Function Estimator is the marked correlation function (MCF, Beisbart & Kerscher 2000; Sheth 2005). This estimator weights the two-point statistic by "marks," which are typically properties of the tracers. The MCF is useful for studying the connection between galaxies and their spatial clustering. Skibba et al. (2006) used it to determine that luminosity-dependent clustering is a straightforward consequence of mass dependence. Armijo

et al. (2018) applied the MCF to test a class of modified gravity theories by marking with local density, demonstrating that there is additional information in the environmental dependence of clustering. The MCF has also been shown to break the degeneracy between halo occupation distribution parameters and the cosmological parameter σ_8 (White & Padmanabhan 2009).

The Continuous-Function Estimator can easily incorporate the *idea* behind marks by choosing the basis functions to be functions of the desired properties of the tracer in addition to pair separation. Combined with the choice of tophat basis functions and proper normalization, this would indeed be equivalent to the MCF. However, the Continuous-Function Estimator can generalize this concept even further. Rather than still producing a two-point function that is only a function of separation, weighted by the marks, our estimator can elevate the marking properties to another continuous axis. That is, it can estimate a multi-dimensional correlation function as a function of both separation and the given property. This provides a more flexible way to look at the dependence of the 2pcf on the property, and has similar applicability to breaking parameter degeneracies. We elaborate on the use cases for incorporating further tracer information into the choice of bases functions in Section 4.6. KSF says: @hogg, do you still have an issue here with claiming this much relationship between ours and the MCF? it is true that one could select basis functions that just use the properties to compute weights and multiply the pair counts within the basis functions. but this isn't our idea with passing more info, it's projecting into the space of the properties as well, which i think i illustrate

4.3. *Beyond the Landy–Szalay Estimator*

While we have formulated our estimator as a generalization of LS, as this 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) based on the random–random counts; for LS we replace this with our $\boldsymbol{T}_{\mathrm{RR}}$ term (Equation 10). This is also the case for the Peebles & Hauser (1974) (natural) estimator and the Hewett (1982) estimator:

$$\hat{\boldsymbol{\xi}}_{\text{P-H}} = \frac{\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{RR}}}{\boldsymbol{v}_{\mathrm{RR}}} \to \boldsymbol{T}_{\mathrm{RR}}^{-1} \cdot (\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{RR}}) \tag{19}$$

$$\hat{\boldsymbol{\xi}}_{\text{Hew}} = \frac{\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{DR}}}{\boldsymbol{v}_{\mathrm{RR}}} \to \boldsymbol{T}_{\mathrm{RR}}^{-1} \cdot (\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{DR}}) \ . \tag{20}$$

We can also straightforwardly generalize estimators which have a data–random cross-correlation as the normalization term, such as the Davis & Peebles (1983) estimator,

$$\hat{\boldsymbol{\xi}}_{\text{D-P}} = \frac{\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{DR}}}{\boldsymbol{v}_{\mathrm{DR}}} \to \boldsymbol{T}_{\mathrm{DR}}^{-1} \cdot (\boldsymbol{v}_{\mathrm{DD}} - \boldsymbol{v}_{\mathrm{DR}}) \ , \tag{21}$$

where we define

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

The continous form of these estimators can be extended to cross-correlations in a straightforward way as expected. This formulation could also be extended to nearly any linear combination of pair counts. The estimator of Vargas-Magaña et al. (2013), for instance, selects the optimal combination of pair counts; our estimators could be combined to create an even more generalized estimator. However, some estimator formulations use powers of these terms that are nontrivial to reformulate as normalization tensors for our estimation approach; we leave this problem for furture work.

4.4. *Implementation and Computational Performance*

We implement the Continuous-Function Estimator within the widely used correlation function package `Corrfunc` (Sinha & Garrison 2019). The package is written in C with python bindings and utilities. For every computation of the pair separation, we additionally pass the separation and any additional tracer information to the user-defined basis function. We output the traditional pair counts as well as our component vector and, if desired, the component tensor.

The computational scaling for our estimator is by definition the same as the traditional method, as pair-finding remains the limiting factor. However, because the Continuous-Function Estimator must evaluate the set of basis functions for each pair of galaxies, it can take significantly longer. For simple basis functions like splines, this will only marginally decrease performance. For more complicated functions, the Continuous-Function Estimator may incur significant extra computational expense. Basis functions can also be input on a grid (of separation or any other property) and then interpolated; the performance is then similar for all functions, depending on how the interpolation is done, but interpolating each function for each pair does somewhat decrease the performance. Though the performance at the time of estimation may be slower than the traditional estimator, the choice of basis may significantly save computational time in other areas, such as reducing the number of mock catalogs required for covariance matrix estimation; see Section 4.5.

We detail a number of implementation choices here. Our formulation of the Continuous-Function Estimator requires the inverse of the random–random tensor $\boldsymbol{T}_{\mathrm{RR}}$ to compute the amplitudes (Equation 12). However, we don't compute this inverse directly, as can be unstable, and is not in fact the end result we are interested in: we want the dot product between $\boldsymbol{T}_{\mathrm{RR}}^{-1}$ and the numerator $\boldsymbol{v}$ of the estimator. For this reason, we use the "solve" operation which computes the solution $\boldsymbol{a}$ of the well-determined matrix equation $\boldsymbol{T}_{\mathrm{RR}}\,\boldsymbol{a} = \boldsymbol{v}$. We also make sure to report the condition number of the tensor, as numerical precision decreases as the condition number rises. If the condition number large, then a rescaling of the basis functions can improve stability. KSF says: what else to include here?

4.5. *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 analytically, though there is promising progress on this front (e.g. Wadekar et al. 2020). For major analyses, 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)

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

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.

We typically require the inverse covariance matrix for analyses, but its form is nontrivial, as the inverse of an unbiased estimator is not necessarily unbiased. Standard practice applies a correction factor (Hartlap et al. 2007),

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

However, this does not correct for errors in the covariance matrix; these propagate to errors on the estimated 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 $N_{\rm mocks} >> N_{\rm bins}$ (with 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 $(1 + N_{\rm bins}/N_{\rm mocks})$ (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) DR9 analysis, 600 mocks were needed and the two-point correlation function used 41 bins (Sánchez et al. 2012) (though they also perform a restricted 15-bin analysis over the BAO peak scales). For the BOSS DR14 fiducial 2pcf results, 1000 mocks and 18 bins were used Ata et al. (2017). Some surveys have already turned to approximate methods for these mock catalogs instead of performing full cosmological simulations, as the cost is prohibitive. Future surveys will have even more costly requirements on mock catalogs, with larger simulations necessary to cover the larger survey volumes and more realistic mocks required to achieve the desired accuracy and precision on the covariance matrix.

An alternative to increasing $N_{\rm mocks}$ is decreasing $N_{\rm bins}$ 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.

We have shown that we can use the Continuous-Function Estimator to estimate the 2pcf using fewer components and without sacrificing accuracy. This means that we can safely reduce $N_{\rm bins}$, or in our case, the number K of components (or basis functions). The covariance matrix will then express the covariance between these components (rather than bins, as we have obviated binning). 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 mock catalog construction, which will be particularly important for upcoming large surveys. Alternatively, with the same number of mock catalogs, one can achieve increased precision just using the Continuous-Function Estimator as an alternative to the standard estimator.

Another issue with the covariance of the standard estimator is that the uncertainty is highly correlated across bins. Thus the diagonal terms of the covariance matrix are poor representations of the true error on each bin. The errors can be decorrelated by choosing a new estimator that is a linear combination of the original 2pcf bins. Hamilton & Tegmark (2000) proposed a transformation using the symmetric square root of the Fisher matrix, and this was shown in Anderson et al. (2014) to significantly suppress the off-diagonal elements of the covariance matrix. While this decorrelated covariance matrix is not used in the fitting in that analyses, it is useful for visualizing the uncertainty of the 2pcf estimates. The Continuous-Function Estimator could also be used to obtain a decorrelated covariance matrix. One could perform an initial estimation with standard bins or basis functions, and then apply a transformation to decorrelation them. These decorrelated bins could then be passed to the estimator as basis functions, and the analysis run again, in order to obtain a direct estimation that produces representative diagonal errors. This would be particularly important for unlocalized basis functions such as the BAO basis functions, which will have highly correlated errors. KSF says: more to say about the covariance of CFE basis functions here?

4.6. *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).

One could take this idea even further by choosing as basis functions a parametrized model of the 2pcf and the derivatives of this model with respect to the parameters of interest, such as cosmological parameters and even halo occupation distribution parameters. The Continuous-Function Estimator would then output the contribution of the derivative terms, which would be directly translatable to the changes in the model needed to best fit the data. This is analogous to the BAO analysis performed in Section 3.4, but with a higher dimensional space of derivatives of parameters of interest. This approach essentially performs a direct estimation of the parameters, without the need for the intermediate steps of binning and fitting.

Another class of applications involves a choice of basis functions that depend not only on the separation between tracer pairs, but also on the properties of the tracers themselves. One such use case 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'. The Continuous-Function Estimator would then output a 2pcf that is a function of both separation and redshift, providing a continuous way to look at the redshift dependence of clustering.

This approach also lends itself to analyzing the relationship between the LSS and galaxy formation, a connection critical for understanding this astrophysical process. The traditional way of doing this involves binning by galaxy luminosity, and then computing the correlation function of the galaxies in each bin (e.g. Budavari et al. 2003, Zehavi et al. 2011, Durkalec et al. 2018). The Continuous-Function Estimator can remove the need for this extra layer of binning by using basis functions that depend on both the pair separation and on some function of the luminosities of the two galaxies. This would result in a direct way to look at the 2pcf luminosity dependence. This could be extended to other galaxy properties, such as color or Hubble type, as well as environmental properties like the local density (e.g. Li et al. 2006, Abbas & Sheth 2006, Skibba et al. 2014). The Continuous-Function Estimator provides the flexibility to explore such a high-dimensional parameter space, while binned methods become quickly limited by number statistics as one tries to include more parameters.

Beyond these standard use cases, 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, Mukherjee & Wandelt (2018) investigated whether there is a directional dependence in estimated cosmological parameters across the sky, by performing analyses on patches of the Cosmic Microwave Background. 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. The Continuous-Function Estimator would allow for a more precise estimate of this dependence as it doesn't require any sort of patches or spatial binning, instead estimating a multi-dimensional continuous 2pcf. While these effects would be highly degenerate with systematics, our estimator combined with robust systematics mitigation opens investigation channels into the possibility of new physics.

Finally, our estimator can be directly related to a power spectrum analysis. We could choose 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 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 insights, 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 also thanks Roman Scoccimarro, David Grier, Alex Barnett, Lucia Perez, James Rhoads, Sangeeta Malhotra, Drew Jamieson, and Chris Lovell for helpful discussions. All of the code used in this *Article* is available open-source at `github.com/kstoreyf/Corrfunc` and `github.com/kstoreyf/continuous-estimator`.

APPENDIX

A. AFFINE INVARIANCE

The estimate of the 2pcf with the Continuous-Function Estimator should not depend on the scaling of the chosen basis functions. Thus we expect the Continuous-Function Estimator to be invariant under affine transformations of the basis functions, meaning transformations that preserve collinearity and distance ratios; the following demonstrates this affine invariance.

We represent the affine transformation by an invertible 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. We choose $\boldsymbol{M}$ to be invertible to ensure the two bases have the same expressive capacity. Then in the primed basis,

the pair counts become

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

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

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

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

For the random–random tensor we have

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

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

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

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, using the shorthand $\hat{\xi}_{ij} = \hat{\xi}(\mathsf{G}_i, \mathsf{G}_j)$, is

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

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

$$= \boldsymbol{a}^{\mathsf{T}}\,[(M^{-1})^{\mathsf{T}}]^{\mathsf{T}} \cdot (\boldsymbol{M}\,\boldsymbol{f}_{ll'}) \tag{A15}$$

$$= \boldsymbol{a}^{\mathsf{T}}\,\boldsymbol{M}^{-1} \cdot \boldsymbol{M}\,\boldsymbol{f}_{ll'} \tag{A16}$$

$$= \boldsymbol{a}^{\mathsf{T}} \cdot \boldsymbol{f}_{ll'} \tag{A17}$$

$$= \hat{\xi}_{ll'}\ . \tag{A18}$$

Thus after an affine transformation of the basis function, the resulting estimator is equivalent to the estimator in the original basis. The method is shown to be affine invariant.

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 $\boldsymbol{v}_{\mathrm{RR}}$ term analytically in the standard approach to correlation function estimation. Here we derive this, and then derive the equivalent for our continuous-basis $\boldsymbol{v}_{\mathrm{RR}}$ and $\boldsymbol{T}_{\mathrm{RR}}$ terms.

Our goal is to estimate the number of pairs in a periodic cubic volume filled uniformly with tracers, $\boldsymbol{v}_{\mathrm{RR}}^{\mathrm{ana}}$. 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. 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,

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

where we assume spherical symmetry. 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 projection vector $\boldsymbol{v}_{\mathrm{RR}}^{\mathrm{ana}}$, which has elements

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

where we are now integrating over all values of r we are interested in from some r_{min} to r_{max}. (For non-localized basis functions, the fully correct thing would be to integrate from $-\infty$ to ∞, though some bounds must be chosen in practice.)

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

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

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 $\boldsymbol{v}_{\rm DR}$ and calculate the amplitudes as

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

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 11.

This analytic form for the continuous estimator could be extended to basis functions that depend on other tracer properties in addition to pair separation. In this case, one would have to integrate over these axes as well, but the idea is the same.

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 Section 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 magnitude of the projection 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, $c_{\rm thresh}$,

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

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

To achieve convergence, we need to be careful in choosing our next $\alpha_{\mathrm{guess},i}$. 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

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

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 α_{guess}, and the model with a value shifted by a small $\Delta\alpha$,

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

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

We choose the magnitudes of the basis functions k to set them at similar scales, providing improved stability in convergence. We find that the values $k_0 = 0.1$, $k_1 = 10.0$, $k_2 = 0.1$, and $k_3 = 0.001$ provide the fastest convergence, though the results are insensitive to choices near these values.

REFERENCES

Abbas, U., & Sheth, R. K. 2006, Monthly Notices of the Royal Astronomical Society, 372, 1749, doi: 10.1111/j.1365-2966.2006.10987.x

Agrawal, A., Makiya, R., Chiang, C. T., et al. 2017, Journal of Cosmology and Astroparticle Physics, 2017, doi: 10.1088/1475-7516/2017/10/003

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
Armijo, J., Cai, Y. C., Padilla, N., Li, B., & Peacock, J. A. 2018, Monthly Notices of the Royal Astronomical Society, 478, 3627, doi: 10.1093/MNRAS/STY1335
Ata, M., Baumgarten, F., Bautista, J., et al. 2017, MNRAS, 000, 2. https://arxiv.org/abs/arXiv:1705.06373v2
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
Beisbart, C., & Kerscher, M. 2000, The Astrophysical Journal, 545, 6, doi: 10.1086/317788
Beutler, F., Seo, H. J., Saito, S., et al. 2017, Monthly Notices of the Royal Astronomical Society, 466, 2242, doi: 10.1093/mnras/stw3298
Budavari, T., Connolly, A. J., Szalay, A. S., et al. 2003, The Astrophysical Journal, 595, 59, doi: 10.1086/377168
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
Colless, M., Dalton, G., Maddox, S., et al. 2001, Monthly Notices of the Royal Astronomical Society, 328, 1039, doi: 10.1046/j.1365-8711.2001.04902.x
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
DeBoor, C. 1987, A practical Guide to Splines (New York, NY: Springer)
Demina, R., Cheong, S., BenZvi, S., & Hindrichs, O. 2016, MNRAS, 480, 49, doi: 10.1093/mnras/sty1812
DES Collaboration. 2005, The Dark Energy Survey: Status and First results, Tech. rep.
Dodelson, S., & Schneider, M. D. 2013, Physical Review D - Particles, Fields, Gravitation and Cosmology, 88, doi: 10.1103/PhysRevD.88.063537
Durkalec, A., Le Fèvre, O. L., Pollo, A., et al. 2018, Astronomy and Astrophysics, 612, 1, doi: 10.1051/0004-6361/201730734
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
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, The Astrophysical Journal, 633, 560, doi: 10.1086/466512
Elvin-Poole, J., Crocce, M., Ross, A. J., et al. 2017, Physical Review D, 98, 042006, doi: 10.1103/PhysRevD.98.042006
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., & Tegmark, M. 2000, Monthly Notices of the Royal Astronomical Society, 312, 285, doi: 10.1046/j.1365-8711.2000.03074.x
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

Hawkins, E., Maddox, S., Cole, S., et al. 2003, Monthly Notices of the Royal Astronomical Society, 346, 78, doi: 10.1046/j.1365-2966.2003.07063.x
Hewett, P. C. 1982, Monthly Notices of the Royal Astronomical Society, 201, 867, doi: 1982MNRAS.201..867H
Hinton, S. R., Howlett, C., & Davis, T. M. 2019, MNRAS, 000, 1. https://arxiv.org/abs/1912.01175
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
Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C. H., et al. 2016, Monthly Notices of the Royal Astronomical Society, 456, 4156, doi: 10.1093/mnras/stv2826
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, C., Kauffmann, G., Jing, Y. P., et al. 2006, Monthly Notices of the Royal Astronomical Society, 368, 21, doi: 10.1111/j.1365-2966.2006.10066.x
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
Sheth, R. K. 2005, Monthly Notices of the Royal Astronomical Society, 364, 796, doi: 10.1111/j.1365-2966.2005.09609.x
Sinha, M., & Garrison, L. H. 2019, MNRAS, 000, 1. https://arxiv.org/abs/1911.03545
Skibba, R., Sheth, R. K., Connolly, A. J., & Scranton, R. 2006, Monthly Notices of the Royal Astronomical Society, 369, 68, doi: 10.1111/j.1365-2966.2006.10196.x
Skibba, R. A., Smith, M. S. M., Coil, A. L., et al. 2014, Astrophysical Journal, 784, doi: 10.1088/0004-637X/784/2/128
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

Wadekar, D., Ivanov, M. M., & Scoccimarro, R. 2020, 1. https://arxiv.org/abs/2009.00622

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