Draft version August 17, 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.

Keywords: Astrostatistics techniques (1886), Baryon acoustic oscillations (138), Galaxy counts (588), Two-point correlation function (1951), Large-scale structure of the universe (902)

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 $k^{\rm th}$ bin in separation r is

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

where we take the pair counts to be normalized. 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

2.1. *Standard Clustering Estimation*

We assume we have a data catalog with N_D objects and a random catalog with N_R objects. The pair counts for LS and related estimators are then defined explicitly as

$$DD_k \equiv \frac{2}{N_D(N_D-1)} \sum_{nn'} i(g_k < |x_n - x_{n'}| < h_k) \tag{2}$$

$$DR_k \equiv \frac{1}{N_D N_R} \sum_{nm} i(g_k < |x_n - x_m| < h_k) \tag{3}$$

$$RR_k \equiv \frac{2}{N_R(N_R-1)} \sum_{mm'} i(g_k < |x_m - x_{m'}| < h_k), \tag{4}$$

where DD_k is the count of data–data tracer pairs 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, x is the tracer position, the n and n' indices index data positions, the m and m' indices index random catalog positions, DR_k is the count of data–random pairs, and RR_k is the count of random–random pairs. KSF says: added this next line, thoughts? The tracer position can be 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. KSF says: Generalize to D1 D2 R1 R2? I think I'd prefer to keep as-is so it looks more familiar. But this requires the clunky normalization factor.

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 is 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, it is clear that it is suboptimal 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

$$X = [A^\mathsf{T} C^{-1} A]^{-1} [A^\mathsf{T} C^{-1} Y] \tag{5}$$

where X is the vector of best-fit parameters, A is a matrix with zeroth and first order terms of x data, C is the covariance matrix, and Y is a column vector of y data. The second bracketed term contains the observed data; the first bracketed term weights the data by the errors. 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, $\xi_k = DD_k/RR_k - 1$ (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*

KSF says: Q about notation; I switched from the sub-k notation to vector notation, because these need to be vectors / tensors to then compute the amplitude. But this might be adding additional confusion. I could write DD, RR etc in k-notation, and then switch to vector notation for the amplitude and evaluation equations? Thoughts?
Inspired by least-squares fitting, we generalize the LS estimator defined above in Equations 2-4. We generalize the indicator function i to any function f. We further generalize the arguments of the function to any properties of the galaxies, rather than just their separation. Writing these now in vector notation, with f returning a vector of length K where K is the number of "basis functions", we have KSF says: I introduced this T notation to be the most general, as f can be a function of any galaxy properties, not just separation (which would be e.g. r_{nm}), but it's a bit clunky - especially when it comes to evaluating xi at the amplitudes. Thoughts?

$$DD \equiv \frac{2}{N_D(N_D-1)} \sum_{nn'} f(T_n, T_{n'}) \tag{6}$$

$$DR \equiv \frac{1}{N_D N_R} \sum_{nm} f(T_n, T_m) \tag{7}$$

$$RR \equiv \frac{2}{N_R(N_R-1)} \sum_{mm'} f(T_m, T_{m'}) \tag{8}$$

$$QQ \equiv \frac{2}{N_R(N_R-1)} \sum_{mm'} f(T_m, T_{m'}) \cdot f^{\mathsf{T}}(T_m, T_{m'}), \tag{9}$$

where DD, DR, and RR are now K-vectors, T refers to the data for the given tracer, and QQ is a K-by-K matrix defined as the outer product of basis function evaluations of the random–random pairs. KSF says: what do we think about K-vector notation? K is number of basis functions / components, as stated above KSF says: should use "matrix" or "tensor"?

Then, we can compute the 2pcf as

$$a \equiv QQ^{-1} \cdot (DD + 2DR - RR) \tag{10}$$

$$\xi(T_l, T_{l'}) \equiv a^{\mathsf{T}} \cdot f(T_l, T_{l'}) \tag{11}$$

where T_l and $T_{l'}$ contain the data values at which to evaluate ξ, and a is a K-vector of the "amplitudes" of the basis functions.

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 the 2pcf. The Continuous-Function Estimator outputs the best-fit linear combination of basis functions to describe the 2pcf. In this sense, it is related to the linear least-squares fitting described above. With this 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.

KSF says: What do you think of the ordering of these paragraphs? We can recover the Landy-Szalay estimator with the proper choice of f and T. From our full galaxy pair data T_n and $T_{n'}$, we can use only their separation, $|x_n - x'_n|$. We can then define a set of basis functions f^{T} as,

$$f_k^{\mathsf{T}}(T_n, T_{n'}) = i(g_k < |x_n - x_{n'}| < h_k) \tag{12}$$

where k denotes a particular bin in separation. This is the top-hat or rectangular function. In this case the DD, DR and RR vectors simply become binned pair counts, with diagonal elementsm with g_k and h_k the bin edges as before. The QQ matrix becomes diagonal, with its diagonal elements equal to the RR vector elements. Then the evalution of the amplitudes and ξ results in the equivalent of the LS estimator–just displayed in a continuous form.

With this generalized two-point estimator, the basis functions f need not have hard edges like the top-hat function. They can instead be smooth functions of the pair

separation, chosen to suit the science use case. Further, the bases can make use of other information about the tracers, such as galaxy luminosity. KSF says: should I note here that we don't explore this further in this paper? The estimator has the property that it is invariant under affine transformations; we show this in Appendix A. KSF says: What more should I say about this? Why is it important? "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 can also write down the form of the estimator when we don't need to worry about the survey window; i.e. when we are working with a periodic box. In this case, we can analytically compute the RR term, as well as the QQ 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

We demonstrate the application of the Continuous-Function Estimator on artifical data. We generate lognormal mock catalogs (Coles & Jones 1991) using the `nbodykit` package (Hand et al. 2018). We use an input power spectrum with the Planck cosmology KSF says: CITE. The true correlation function is then known: it is the Fourier transform of the input power spectrum, computed numerically. Our test catalogs have size $(750\ h^{-1}\mathrm{Mpc})^3$ and a galaxy number density of 3×10^{-4}, to match that of the Brightest Cluster Galaxies KSF says: CITE. We construct 100 realizations of this box. We also generate a uniformly sampled random catalog with 10 times the number of galaxies as in the data catalogs.

3.1. *Demonstration using Spline Basis Functions*

KSF says: Should i have a figure showing the basis functions? Hogg says: yes

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: CITE. 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,

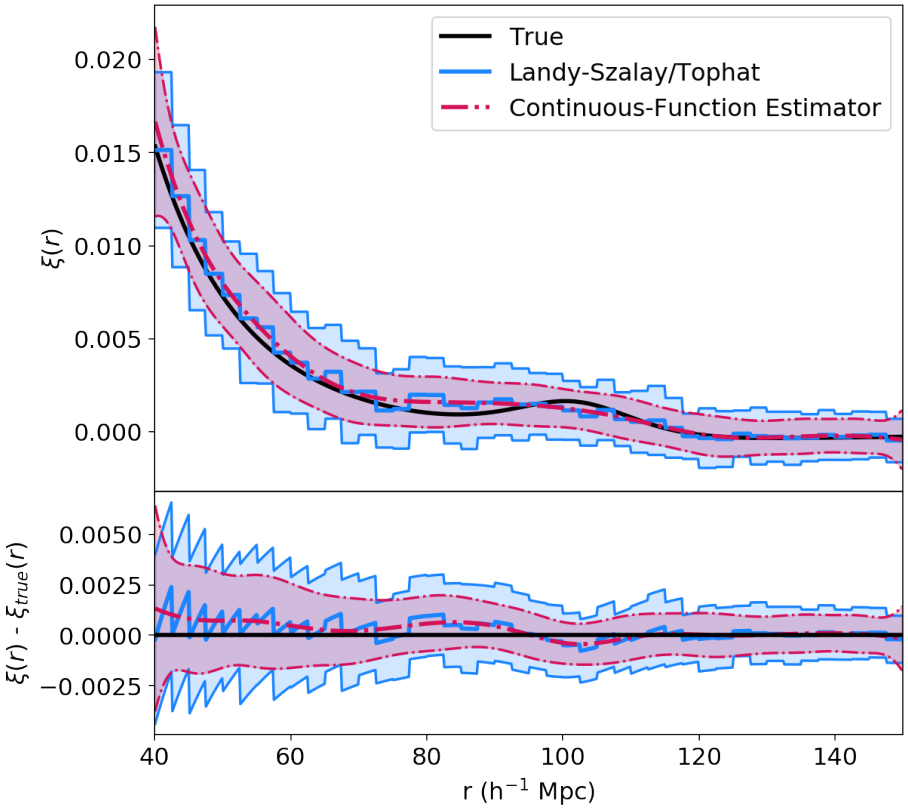

Figure 1. A comparison between the Continuous-Function Estimator with a cubic spline basis function (red dotted) and the standard tophat estimator (blue solid). The shaded region is the 1σ variation in the 100 mock catalogs. The cubic spline estimator has lower bias and lower variance with fewer bins. KSF says: quote rmse in caption? otherwise not clear from figure its less biased. Hogg says: yes KSF says: does tophat need to be a different linestyle?

KSF says: fix to dotted linestyle

which 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 this standard estimator, reformulated as continuous functions using a tophat basis. For the tophat basis we use 44 basis functions in the range $40 < r < 150$ h^{-1}Mpc, each with a width of 2.5 h^{-1}Mpc. For the cubic spline basis, we use the same r range, but with only 11 basis functions, and knots chosen on a grid of 10 h^{-1}Mpc. KSF says: need to specify more about knots? is this accurate (not quite, the knots repeat values at the edge i think...) and should i mention control points?

The results are shown in Figure 3.1, compared to the true input 2pcf. The spline basis results in a 2pcf estimate that has a root mean square error (RMSE) with respect to the truth of 4.38×10^{-4}, compared to 5.31×10^{-4} for the tophat basis. This holds true when we average over the bin KSF says: need to do weighted average based on which r-values contribute? supposed to hear from tinker, as in standard practice, and then compute the RMSE. The spline basis also results in lower variance across the

100 mocks, compared to the tophat basis. Thus, the Continuous-Function Estimator allows for a more accurate and precise estimate of the 2pcf with fewer components.

We note that this is fundamentally different than a kernel-based estimator. In a kernel formulation, each data point is smoothed by a particular kernel, smoothing out features in the resulting function. With our estimator, the basis functions are fixed and the data is projected onto them. 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. KSF says: please add/correct here

3.2. *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 measure the spherically averaged correlation function, $\xi(s)$, where s is the separation between pairs. 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:

$$\alpha = \left(\frac{D_A(z)}{D_A^{\rm mod}(z)}\right)^{2/3}\left(\frac{H^{\rm mod}(z)}{H(z)}\right)^{1/3}\left(\frac{r_s^{\rm mod}}{r_s}\right), \tag{13}$$

where D_A is the angular diameter distance, H is the Hubble constant, r_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

$$\xi^{\rm fit}(s) = B^2\xi^{\rm mod}(\alpha s) + \frac{a_1}{s^2} + \frac{a_2}{s} + a_3 \tag{14}$$

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. 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 take the numerical partial derivative of the model with respect to α, using a change in α of size $d\alpha$. Our fitting function is then

$$\xi^{fit}(s) = B^2\xi^{mod}(s) + Cd\alpha\frac{d\xi_{mod}(\alpha s)}{d\alpha} + \frac{a_1}{s^2} + \frac{a_2}{s} + a_3, \tag{15}$$

where C describes the contribution of α. We input these five terms (with arbitrary scaling) as the basis functions of our estimator. The estimator outputs an amplitude vector a as described in §2.3, which give the contribution of each term—precisely the values of B, C, a_1, a_2, and a_3. From the value of C, we can directly compute α, as $\alpha_{\rm recovered} = \alpha + Cd\alpha$ (where α is our initial guess, typically starting with α=1). That is, a value of $C = 0$ indicates that the model is the best fit to the data and no scale dilation is needed, while nonzero values give the magnitude and direction of the scale dilation parameter.

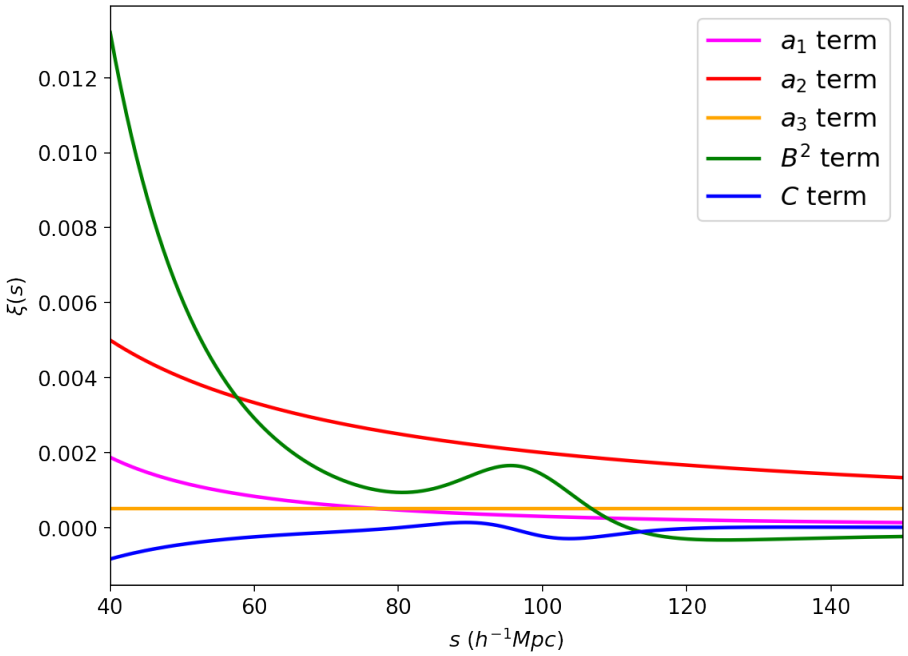

Figure 2. The set of basis functions used to fit for the BAO scale using our estimator. The B^2 term (green) is the fiducial model used to determine the scale dilation parameter α. The C term is the derivative of this model with respect to α, allowing for the estimation of this parameter. The a_1, a_2,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

We demonstrate this method using the same set of lognormal mocks as in §3.1. 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 cosmology used for our mock catalogs, we can compute the true value of the scale dilation parameter, $\alpha_{\rm true} = 0.9574$ (a fairly extreme value but useful for testing

purposes). With this fiducial model, we can construct the basis functions for our estimator; these are shown (with $\alpha = 1$ and the free parameters set to arbitrary values) in Figure 3.2.

We perform an iterative procedure to estimate α. After the first estimation with the $\alpha = 1$, model, we use the value of C to compute the recovered α, as described above. We then take that value as the new α, and re-run the estimator. (In practice, this often jumps over the true value, so we tune the choice of the next α with a parameter η, as $\alpha_{\rm recovered} = \alpha + \eta C d\alpha$, where η is typically 0.1-0.5.) We stop when the percent change between α and $\alpha_{\rm recovered}$ is less than 0.1%; it typically takes fewer than 10 iterations to converge. KSF says: update this with true values We apply this procedure to each of the 100 mocks; the resulting estimate for the correlation function is shown in Figure 2. 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 $\alpha = 0.9572 \pm 0.0315$, very close to the true value $\alpha_{\rm true} = 0.9574$.

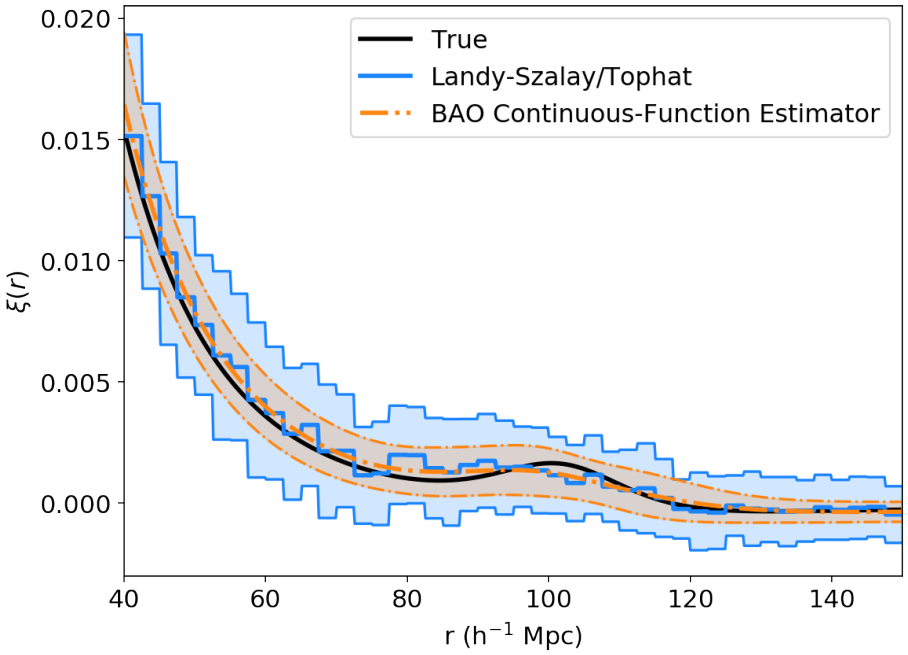

Figure 3. Estimation of the correlation function using our estimator with basis functions based on the BAO fitting function (orange dot-dashed). The line is the mean of the final estimate from the iteration procedure for 100 mocks, and the shaded region is the 1σ variation. We also show the standard 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?

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!

4. DISCUSSION

4.1. *Relationship to Other Existing Estimators*

The 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.

KSF says: Paragraph on KDEs

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. The formulation currently requires a normalization term (i.e. denominator) of the RR counts, as we replace this with our QQ term. This is the case for the Peebles & Hauser (1974) estimator and the Hewett (1982) estimator:

$$\xi_{PH}(r) = \frac{DD - RR}{RR} \tag{16}$$

$$\xi_{Hew}(r) = \frac{DD - DR}{RR}. \tag{17}$$

KSF says: dropped k and (r) notation here, do i need? We could also generalize estimators which have a DR term as the denominator, such as the Davis & Peebles (1983) estimator,

$$\xi_{DP}(r) = \frac{DD - DR}{DR} \tag{18}$$

by defining

$$DQ = \frac{2}{N_D N_R} \sum_{nm} f(T_n, T_m) \cdot f^{\mathsf{T}}(T_n, T_m). \tag{19}$$

This could be extended to almost 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: Mention other work with Barnett about LS estimator?

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

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)

$$\hat{C}_{ij}^{ML} = \frac{1}{N_{mocks}-1}\sum_{k=1}^{N_{mocks}}\Big(x_i^k - \bar{x}_i\Big)\Big(x_j^k - \bar{x}_j\Big), \tag{20}$$

where k denotes the index of the mock, i and j denote the index of the bin or component, x denotes the estimate in that bin for that mock, and $\bar{x}$ denotes the mean value of the estimate in that bin across the mocks. 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)

$$\hat{C}^{-1} = \frac{N_{mocks}-N_{bins}-2}{N_{mocks}-1}\Big(\hat{C}^{ML}\Big)^{-1}. \tag{21}$$

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 $N_{mocks} >> N_{bins}$ (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 $(1 + N_{bins}/N_{mocks})$ (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_{mocks} becomes possible. This will significantly reduce requirements on mocks, which will be particularly important for upcoming large surveys.

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

$$f_k(T_n, T_m) = f_k(r_{nm}, z_n, z_m), \tag{22}$$

where z is the redshift of tracer n or m. 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

It is a pleasure to thank... 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. We represent this by a matrix M, such that

$$f' \leftarrow Mf \tag{A1}$$

Then in the primed basis, the pair counts become

$$DD' = \sum_{nn'} M f_{nn'} = M\, DD \tag{A2}$$

$$DR' = \sum_{nm} M f_{nm} = M\,DR \tag{A3}$$

$$RR' = \sum_{mm'} M f_{mm'} = M\,RR \tag{A4}$$

where $f_{nn'} = f_k(T_n, T_{n'})$. We have factored M out of the summation. For the QQ matrix we have

$$QQ' = \sum_{mm'} (M f_{mm'}) \cdot (M f_{mm'})^{\mathsf{T}} \tag{A5}$$

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

$$= MQQM^{\mathsf{T}} \tag{A7}$$

Then the amplitudes in the primed basis become

$$a' = [MQQM^{\mathsf{T}}]^{-1} \cdot [MDD - 2\,MDR + MRR] \tag{A8}$$

$$= (M^{\mathsf{T}})^{-1} QQ^{-1} \cdot [DD - 2\,DR + RR] \tag{A9}$$

$$= (M^{\mathsf{T}})^{-1} a \tag{A10}$$

and the estimator in the primed basis is

$$\xi' = [(M^{\mathsf{T}})^{-1} a]^{\mathsf{T}} \cdot (Mf) \tag{A11}$$

$$= a^{\mathsf{T}} [(M^{-1})^{\mathsf{T}}]^{\mathsf{T}} \cdot (Mf) \tag{A12}$$

$$= a^{\mathsf{T}} M^{-1} \cdot Mf \tag{A13}$$

$$= a^{\mathsf{T}} \cdot f = \xi. \tag{A14}$$

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 M be invertible. However, any two equivalent bases must be related by the inverse of a transformation matrix, so this requirement is already satisfied.

B. COMPUTING RR AND QQ ANALYTICALLY

The autocorrelation of the random catalog is meant to approximate the window function. When we have a periodic cube, we can compute this RR term analytically. Here we derive this, and then derive the equivalent for our continuous-basis RR and QQ terms.

We consider an annulus around a single galaxy. This annulus has a volume V_{ann}. Taking the box to have average number density $\bar{n}$, the number of galaxies expected in the annulus is $N_{ann} = A\bar{n}$, and thus this galaxy is part of N_{ann} pairs. We do this for each of the $N_D{-}1$ other galaxies, and find a total of $(N_D{-}1)N_{annulus} = \frac{1}{2}(N_D{-}1)V_{ann}\bar{n}$ pairs, where the factor of $\frac{1}{2}$ accounts for the fact that this double-counts pairs. For a cube, $\bar{n} = \frac{N_D}{L^3}$, so we finally count $\frac{1}{2}\frac{N_D}{L^3}(N-1)V_{ann}$ pairs.

For hard-edged radial bins, we can compute A simply as the difference between spherical volumes. We can also represent this as an integral:

$$V_{ann} = \int_{b_1}^{b_2} dV = 4\pi \int_{b_1}^{b_2} r^2 dr \tag{B15}$$

We can generalize this to any basis function $f(r)$ that is a function of r:

$$V_{ann} = 4\pi \int_{b_1}^{b_2} f^2(r) r^2 dr \tag{B16}$$

which we can see reduces to Equation B15 when $f(r)$ is the tophat function (returning 1 or 0 depending on whether r is between b_1 and b_2).

This gives us our full generalized analytic RR term, which has elements

$$RR_{i,\mathrm{ana}} = \frac{1}{2}\frac{N_D}{L^3}(N_D - 1)4\pi \int_0^{r_{\max}} f_i(r) r^2 dr \tag{B17}$$

where i is the index of the basis function vector. Based on the definition of QQ in Equation 9 as the outer product of the basis function vector and its transpose, we can see that the elements analytic QQ term are:

$$QQ_{ij,\mathrm{ana}} = \frac{1}{2}\frac{N_D}{L^3}(N_D - 1)4\pi \int_0^{r_{\max}} f_i(r) f_j(r) r^2 dr \tag{B18}$$

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

When considering a periodic box, the naive estimator is no longer biased, so we can also avoid computing the DR term and calculate the amplitudes as

$$a_{\mathrm{ana}} = QQ^{-1}_{\mathrm{ana}} \cdot DD. \tag{B19}$$

Looking back, it might have seemed strange that we use N_D in calculating the analytical RR term, but we now see that this normalization prefactor cancels out with that of the DD term. Finally, we can compute the correlation function as before in Equation 11.

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

Hand, N., Feng, Y., Beutler, F., et al. 2018, The Astronomical Journal, 156, 160, doi: 10.3847/1538-3881/aadae0

Hartlap, J., Simon, P., & Schneider, P. 2007, Astronomy and Astrophysics, 464, 399, doi: 10.1051/0004-6361:20066170

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

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

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