Clustered Universe: Redshift Tests In Structure

Alex Brinson
Mentor: Charles Steinhardt

ABSTRACT

Motivated by foundational ideas in cosmological structure evolution, we investigate a novel method to determine redshift demographics of high-z galaxy surveys from their clustering statistics. A methodological framework based on pairwise distance measurements, two-point correlation functions, and flat-sky power spectra is developed to quantify and identify structure signals characteristic particular redshifts, and from this we attempt to decompose mixed-z samples into estimates for the galactic abundances of individual redshift bins. CAMB and HaloFit are used to derive theoretical predictions for correlation functions at arbitrary redshift, but telescope-specific sources of systematic error remain a major hurdle which must be overcome before our technique is applicable to future high-redshift surveys.

1. Introduction

Since the days of Hubble, it has been known that the universe expands, causing faraway objects (usually galaxies) to recede faster than nearby objects. This effect can be quantified by recording objects' emission spectra for an object in question, and measuring the extent by which characteristic spectral lines have been redshifted.

$$\frac{\lambda_{observed}}{\lambda_{emitted}} \equiv z + 1 \qquad \textbf{(Eq. 1)}$$

In the business, astrophysicists often characterize galaxies by referring simply to this *cosmological redshift* (often denoted by the letter z). When combined with a cosmological model about the expansion history of the universe (often ΛCDM), a galaxy's redshift encodes a variety of distance measure, as well as an estimate for how old the universe was when the observed light was actually emitted.

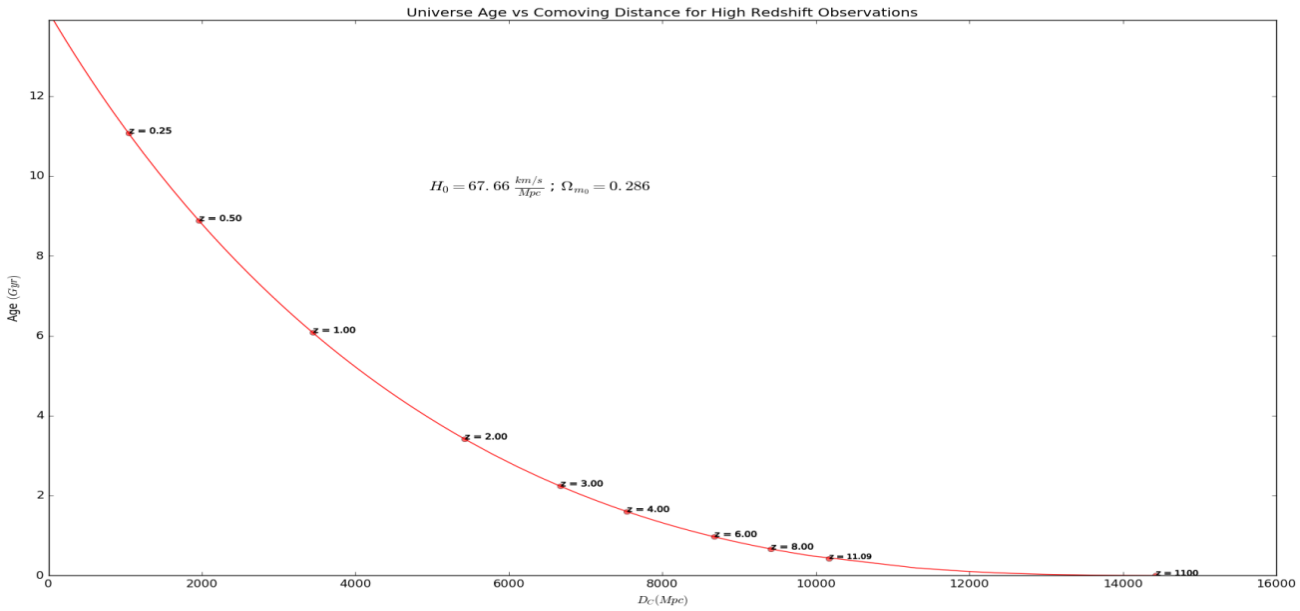

Figure 1: Comoving Distance (Mpc) vs Age of Universe, Labeled at Certain Redshifts

While Equation (1) is how we actually define redshift, it would be impossible to directly compute z from spectral line measurements for every galaxy. For the vast majority of galaxy observations – especially at high redshift – we lack the necessary spectral resolution to identify individual emission lines and calculate *spectroscopic redshifts*. Instead, astrophysicists often resort to a number of inductive inference methods known collectively as photometry, which convert flux measurements over broad wavelength bands into *photometric redshifts*. While

photometric techniques serve an invaluable role, they aren't infallible, and their error sources aren't independent. As telescopes continue to advance and astrophysicists expand their search to increasingly distant regimes, the importance of developing independent techniques for redshift estimation should not be overlooked. In this paper, we consider a few (particularly ambitious) cluster-based approaches to redshift estimation. Section **2** provides an introductory overview of the relevant concepts in theoretical cosmology which underpin our investigations. Section **3** explains our methodology, and section **4** details our results and provides a short conclusion.

2. Cosmic Structure

While our picture of the universe preceding the epoch of recombination is still the subject of debate among cosmologists (e.g. Inflation, Big Bounce, Conformal Cyclic Cosmology), it is known empirically that the origins of structure in the universe were already present by the time of last scattering. Approximately 380,00 years after the Big Bang, the universe had finally spread and cooled enough for photons to decouple from the primordial plasma. This decoupling (as well as the related recombination of protons, neutrons, and electrons into neutral atoms) finally allowed light to travel freely – all 13.8 billion light years from its last interaction in the hot cosmic soup until its eventual detection by a radio telescope's antennae. Collectively, these photons (constantly arriving at the earth from all directions) form the *cosmic microwave background*, and the COBE, WMAP, and Planck satellites have thoroughly investigated its structure on a large range of length scales.

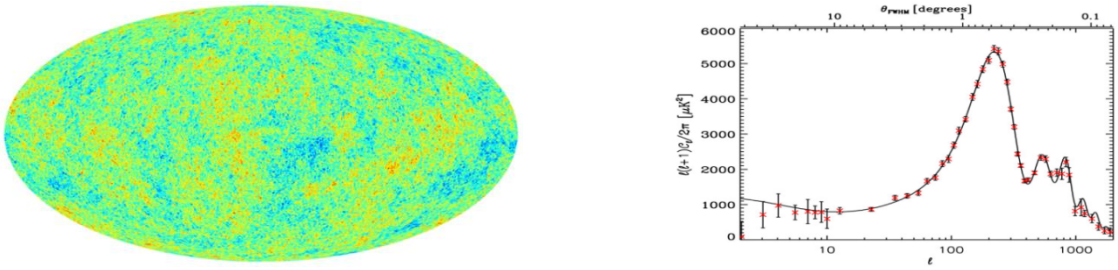

Figure 2: Structure of Temperature Anisotropies in the CMB[i]

While the aforementioned satellites most directly measured fluctuations in CMB temperature, these temperature inhomogeneities are strongly tied to inhomogeneities in the baryonic and (mostly) dark matter mass densities (e.g. gravitationally-induced acoustic oscillations, Sachs-Wolfe red/blue-shifting)[ii], and so we can surmise that by this earliest observable period (gravitational wave-facilitated observations not withstanding) in the history of our universe, the seeds of future galaxies were already sown.

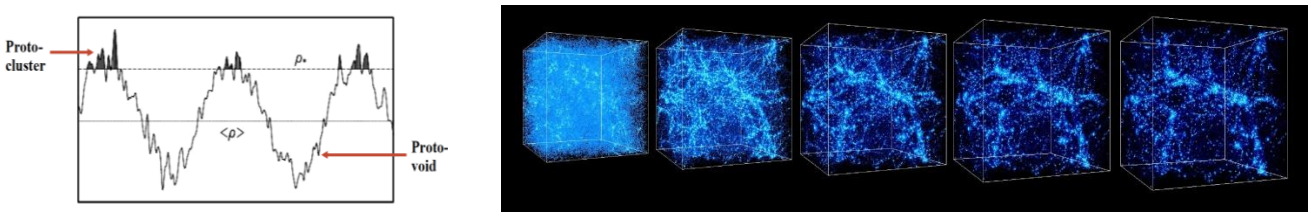

Figure 3: Evolution of Structure in the Universe
Left: Initial Density Fluctuations On Many Scales[iii] **Right: Formation of the large-scale structure in the Universe**[iv]

From this initial mass distribution, the universe evolves primarily due to two competing effects: gravitational attraction and the expansion of spacetime. The seeds of the earliest galaxies correspond to the locations of proto-clusters, local maxima in the matter density field which surpass a threshold level necessary for gravitational collapse to overcome the large scale expansion of the universe. Due to the "scale-invariant" power spectrum of the initial fluctuations, the over-dense and under-dense regions of the universe are woven into

each other with a sort of "Sponge-like" topology[v], and as time passes the under-dense regions empty out and expand, while the over-dense regions further concentrate into dense filaments, often referred to as "the Cosmic Web".

3. Methods/Definitions

3.1. Data Set

All of our methods were developed and tested using data from the COSMOS[vi] catalogue. COSMOS is a deep beam survey covering about 2 square degrees on the sky and containing data on over 1 million galaxies. We filtered for galaxies that fell in an angular window which had minimal masking effects, and which had photometric redshifts (PhotZ) between 0 and 6. Some galaxies also had spectroscopic redshifts (SpecZ), but since spectroscopic followup is a limited resource, these galaxies were the exception and not the rule. Unless otherwise stated, assume PhotZ data was used for everything.

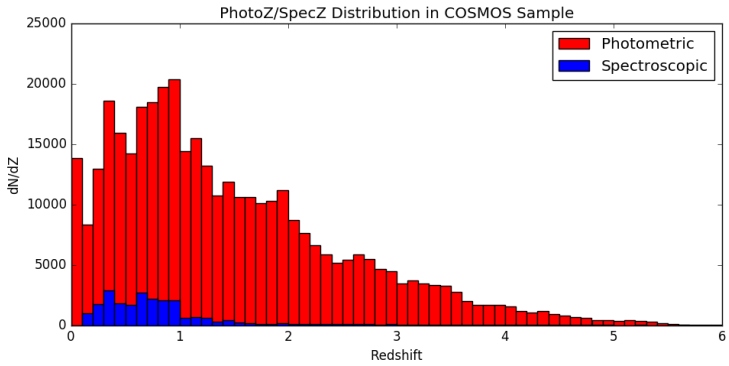

Figure 4: Histogram showing counts of galaxies in our dataset with photmetric(red)/spectroscopic(blue) redshift, binned by redshift

3.2. Pair-wise Distance Distributions

To quantify structure, we looked at the angular distances between every distinct pair of galaxies in a sample. In a collection of N distinct objects, there are $\binom{N}{2} = \frac{N\,(N-1)}{2}$ pairs, and for each of these pairs we compute the Pythagorean distance between the objects, then use all $\binom{N}{2}$ distance measurements to construct a histogram. Upon normalizing the histograms based on the total number of pairs sampled, we obtain an empirical probability density functions for the distributions of pairwise distances, which I refer to as PwD PDFs for concision.

The above procedure characterizes the spacing between objects within a single population. We call these auto-pairs, because the pairs are composed of two objects from the same population. If we consider multiple overlaid populations (e.g. galaxies at different redshifts, viewed in the same angular window) however, then we must also account for cross-pairs. For example, if we construct a sample of N_A objects from population A and N_B objects from population B, then the composite sample of $N = N_A + N_B$ objects will once again have a Pairwise distance histogram constructed from $\binom{N}{2}$ distinct pairs, but this histogram can be decomposed into 3 ingredients: the pairwise distance distributions from **(1)** the $\binom{N_A}{2}$ auto-pairs in population A, **(2)** the $\binom{N_B}{2}$ auto-pairs in population B, and **(3)** the $N_A * N_B$ cross-pairs. Indeed, our decomposition accounts for all of the pairs in the combined sample

$$\binom{N}{2} = \frac{N*(N-1)}{2} = \frac{(N_A+N_B)*(N_A+N_B-1)}{2} = \frac{N_A^2-N_A}{2} + \frac{N_B^2-N_B}{2} + \frac{2*N_A*N_B}{2}$$
$$= \binom{N_A}{2} + \binom{N_B}{2} + N_A * N_B \quad \blacksquare$$

Note that this decomposition generalizes to mixtures of more than 2 sub-populations.

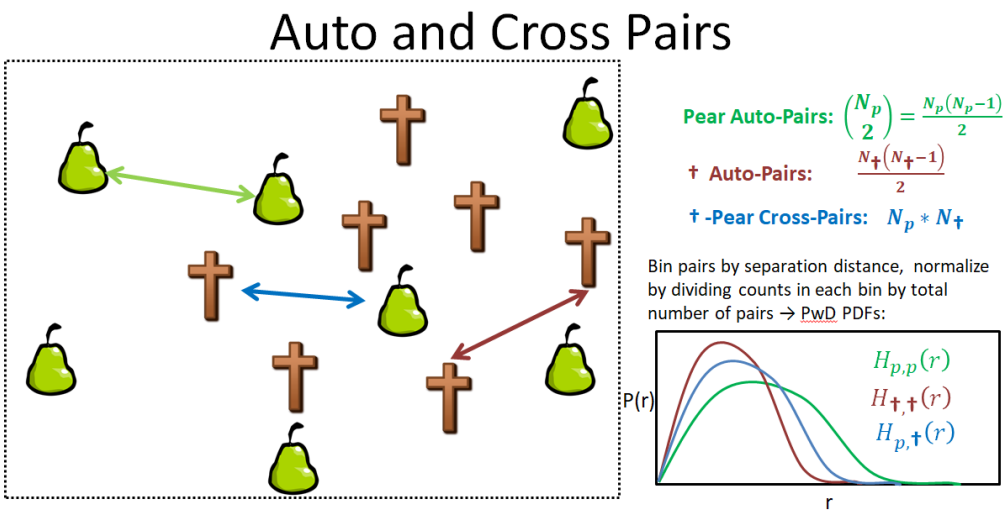

Figure 5: Pictograph showing auto-pairs and cross-pairs (left), as well as the resulting PwD distribution functions (right)

3.3. Correlation Functions

While Pairwise Distance PDFs do give a sense of the typical (angular) spacings between objects in a sample, we can do more. Specifically, we can get a sense of how much "structure" there is in a given sample by comparing our empirically measured distributions to a PwD PDF which corresponds to zero structure, and seeing how they stack up at different (angular) distance scales.

We define our "zero point" of structure as the PwD distribution resulting from (i.i.d.) randomly drawn points within our considered angular window. As it turns out, an analytic solution exists for the resulting PwD PDF[vii]. While the derivation is relatively straightforward, the result is a three-part piecewise function. I'll leave the grungy expression out of this report and instead just show graphically that the function does what it's supposed to do.

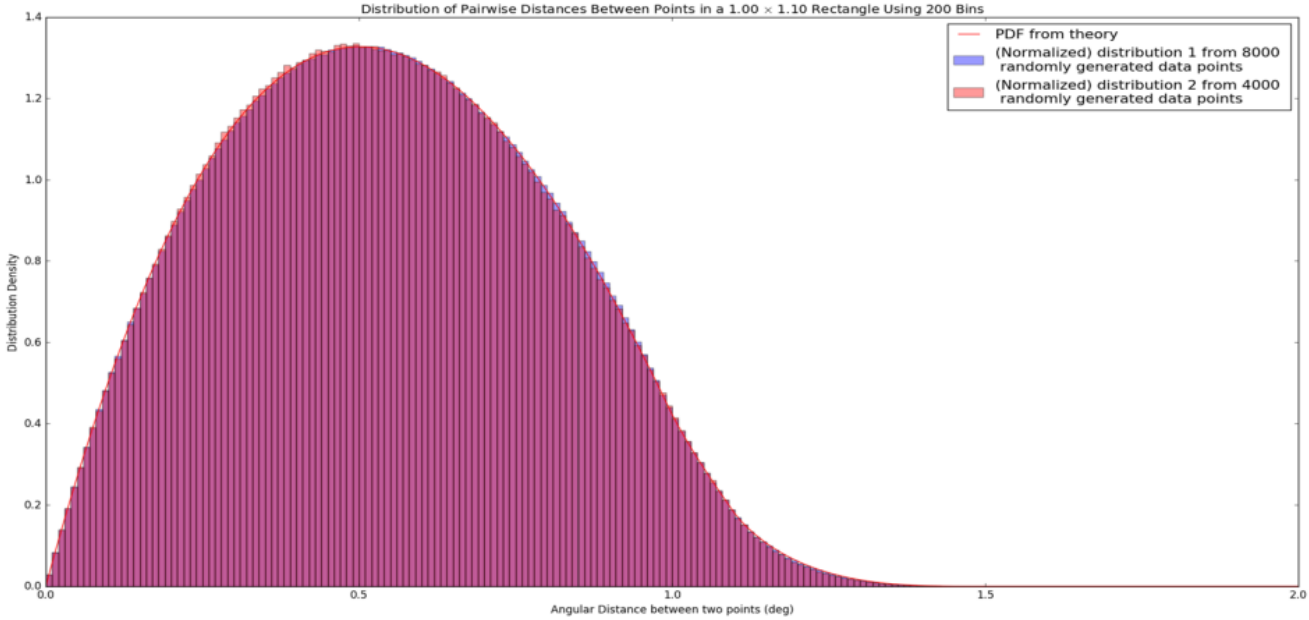

Figure 6: PwD PDF for random sample. It's hard to see, but the red line tracing the perimeter of the bell curve shows the theoretically derived PDF, while the blue and red histograms (which overlap to make purple) are PwD PDFs resulting from generating random data points.

Now, we can measure the over/under-density in a PwD PDFs as a function of (angular) scale by taking the bin-wise ratio of pdf heights. Specifically, let $H_{A,A}[\theta_i]$ be the Pwd PDF from the auto-pairs of some population A,

evaluated at the angular scale θ_i (the i[th] bin of the normalized histogram), and let $H_{RAND}[\theta_i]$ be the pdf corresponding to the random distribution (plotted directly above). The over/under-density of population A pairs separated by an angular distance θ_i, is then:

$$\xi_{A,A}[\theta_i] = \frac{H_{A,A}[\theta_i]}{H_{RAND}[\theta_i]} - 1 \qquad \textbf{(Eq. 2)}$$

$\xi[\theta]$ is known as the *two-point correlation function*, so we refer to $\xi_{A,A}[\theta]$ as an autocorrelation, and $\xi_{A,B}[\theta]$ as a cross-correlation. Note that the -1 guarantees that $\xi_{RAND}[\theta_i]$ = 0 identically for all θ_i; we define it such that there is no signal of structure in a purely random sample.

$\xi[\theta]$ is often defined in a less easily implementable (but nevertheless equivalent) way which is arguably more intuitive, so I'll include it here: "The spatial two-point correlation function is defined as the excess probability of finding a pair of galaxies at a given separation distance θ, compared with that expected of a random distribution."[viii]

$$\mathbf{d}P[\theta] = \bar{n}^2(1 + \xi[\theta]\,)\,\mathbf{d}V_1\,\mathbf{d}V_2 \qquad \textbf{(Eq. 3)}$$

Here, $\bar{n}$ denotes the mean galaxy density, so $\bar{n}^2\mathbf{d}V_1\,\mathbf{d}V_2$ would be the probability density of finding a pair at any given separation if the galaxies were distributed completely homogenously.

3.4. Power Spectra

Correlation functions aren't the end of the story, though; there's still more math to throw at the problem. Just as one can take a function in space/time and Fourier transform to represent the function in terms of its frequency basis components, we can apply an integral transform to map our 2-point correlation functions to the corresponding *power spectra*. If our data were sampled from a patch of Euclidean space with periodic boundary conditions, then a Fourier transform would be the appropriate procedure for translating correlation functions into power spectra. However, since our dataset only contains 2D projections of galaxy positions on the unit sphere, spherical harmonics[1] would really be the most appropriate basis to expand in/transform with[ix]. Fortunately the 2 square degrees observed by COSMOS comprise a very small solid angle[2], so to good approximation we can treat our sampled patch of sky as essentially flat. This flat sky approximation allows us to trade out our Spherical Harmonic Transforms in favor of the relatively simple *Hankel Transforms*[ix]:

$$\text{Hankel Transform:}\quad \Delta^2_\theta[k] = \int_0^\infty \xi[\theta]\; J_0[k\;\theta]\;\theta\;\mathbf{d}\theta \qquad \text{Hankel Inverse Transform:}\quad \xi[\theta] = \int_0^\infty \Delta^2_\theta[k]\; J_0[k\;\theta]\,\frac{\mathbf{d}k}{k} \qquad \textbf{(Eq. 4)}$$

Where $\xi[\theta]$ is the angular two-point correlation function as defined in section **3.3**, $\Delta^2_\theta[k]$ is the corresponding power spectrum (evaluated at some angular frequency k), and J_0 is the 0[th] Bessel function of the first kind: $J_0[z] = \frac{1}{\pi}\int_0^\pi e^{i\,z\,Cos[\theta]}\,\mathbf{d}\theta$. The utility of this mathematical machinery is the topic of section **4.3**.

4. Attempts, Results, And Conclusions

As discussed in section 2, the universe should exhibit different levels of structure at different stages in its evolution, going from a largely homogenous mass distribution during very early periods and gradually converging into a "cosmic web" of densely packed galaxies interwoven through vast swathes of empty space. Operating with this picture in mind, it's not unreasonable to assume that galaxies sampled at particular redshifts (particular stages in the universe's evolution) should exhibit pairwise distance distributions which are characteristic of that redshift, and more specifically that galaxies in lower redshift samples should be biased toward shorter separations, since the galaxies in these samples have had more time to collapse toward each other. This assumption, in conjunction with the more technical framework detailed in section **3**, formed the foundation for our attempts at developing a novel redshift estimation technique. The general strategy was as follows:

---

[1] The Spherical Harmonic expansion could then be reduced to a 1D Legendre Polynomial expansion by assuming isotropy in the power spectrum.

[2] The entire unit sphere is 4 π steradians, or about 41,253 square degrees.

(1) Obtain a collection of PwD PDFs characteristic of galaxies from particular redshift populations, and use these as basis in which to decompose pairwise distance distributions from more general samples.
(2) Find a linear combination of basis functions which most closely matches a sample's (non-normalized) PwD histogram. The coefficients yield an estimate for the numbers of galaxies belonging to particular redshift bins.
(3) Use this technique iteratively to identify samples with large concentrations of high-redshift galaxies, which can then be targeted for spectroscopic follow-up.

4.1. Mixed-Redshift Samples

In order to evaluate our methods, we of course needed to test them on samples for which the true redshift statistics are already known. To do this, we generated test samples by randomly selecting a randomly determined number of galaxies from a random subset of pre-specified redshift populations in our COSMOS sample. For example, say I wanted to generate a test sample of 8000 galaxies from a subset of the redshifts z = [0.25, 0.5, 1, 2], then I might end up with the collection of galaxies in the scatter plot below:

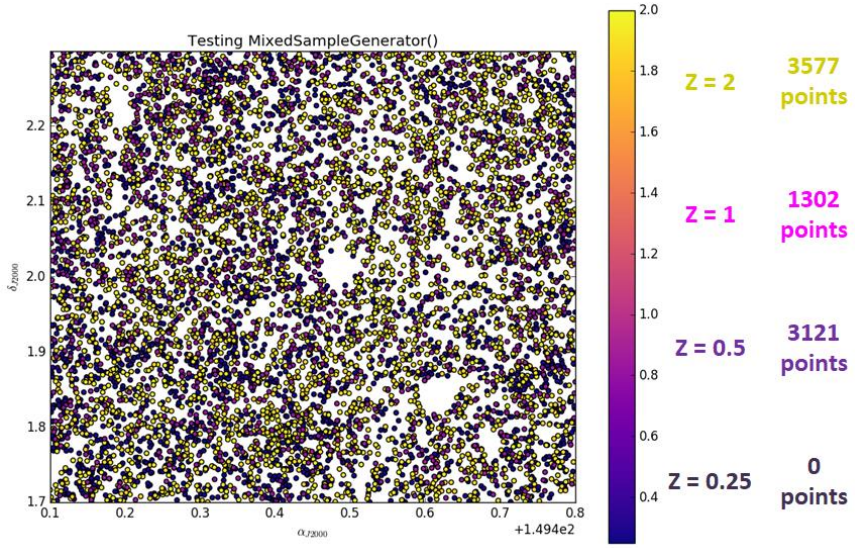

Figure 7: Scatter plot of mixed-redshift sample selected from COSMOS dataset

4.2. Regress, Progress, then Regress

As a first step, we tried decomposing mixtures from just two redshifts. This was accomplished with a simple single-variable regression. For a sample of N galaxies drawn from two populations there are N_A objects from a population at redshift Z_A, and N_B objects drawn from a population at redshift Z_B. As in section **3.2**, the PwD histogram from the mixed sample can be broken down into 3 components. Let $H[\theta_i]$ denote a normalized PwD PDF as in section 3.3, and let $\Gamma[\theta_i]$ denote an un-normalized PwD histogram.

$$\Gamma_{mix}[\theta_i] = \binom{N_A}{2} H_{A,A}[\theta_i] + \binom{N_B}{2} H_{B,B}[\theta_i] + N_A N_B H_{A,B}[\theta_i] \quad \textbf{(Eq. 5)}$$

$H_{A,A}$, $H_{B,B}$, and $H_{A,B}$ can all be estimated empirically by recording PwD PDFs from (different) random samples of A/B objects. We don't know N_A or N_B, but we can count the total number of objects, so we know N. Since we also know that $N = N_A + N_B$, or more suggestively that $N_B = N - N_A$, the coefficients in the equation above all depend on the single unknown parameter N_A. By fitting the right-hand side of the above equation to the bin height of $\Gamma_{mix}[\theta_i]$ over a set of values $[\theta_i]$, we can determine a best fit value for N_A, allowing us to estimate the number of redshift Z_A galaxies in mixed sample.
The two-bin decomposition was actually quite successful, generally predicting values for N_A to within 50, out of total mixed sample sizes N = 6000 (See Figure 8 for an example).

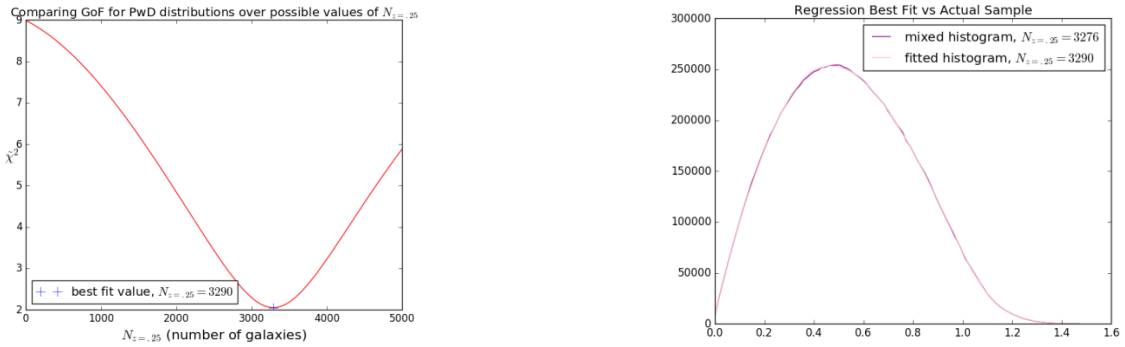

Figure 8: Goodness of Fit Results (Left) and Best Fit PwD Histogram (Right) for Mixed, Two-Redshift Decomposition

We had less success when we tried to generalize this to samples composed from more than 2 redshifts, however. While our multivariate regression functions were behaving properly, we believe that the problem arose from our empirically estimated (and therefore subject to statistical variance/error) basis functions. The large effects of noise are shown below in the correlation functions derived from the PwD PDFs for a large set of redshift (auto/cross) populations:

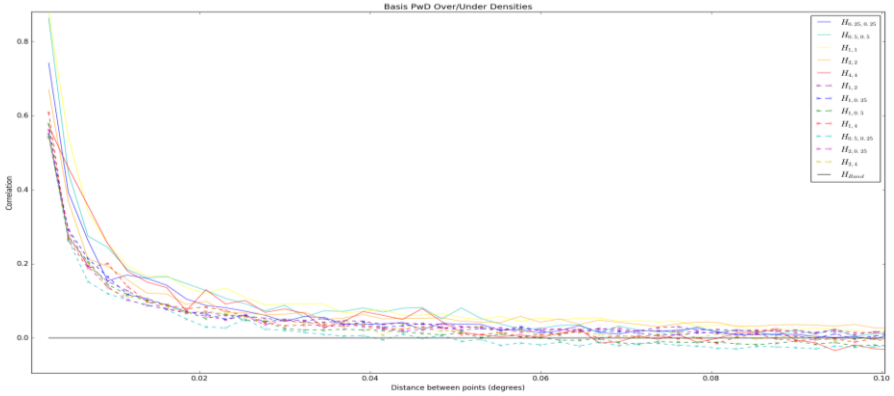

Figure 9: Two-Point Correlation Functions for Redshifts z = (.25, .5, 1, 2, 4): While the cross-correlations are uniformly weaker than the autocorrelations, it is curious that they aren't all zero. We suspect that this is due to masking/selection effects, but the mystery was never fully resolved.

While the overdensity of pairs at small scales can be clearly seen for each redshift, the correlation functions are all very similar, making it difficult to distinguish between many of the redshifts.

4.3. Fitting to Halofit

To get around the noise problem we decided that instead of estimating our basis PwD PDFs from noisy data, we would try to develop the basis distributions from theory. We accomplished this thanks to the assistance of Muhamed Rameez[x] and the Python library CAMB (Code for Anisotropies in the Microwave Background). CAMB doesn't directly calculate PwD distributions or correlation functions, however; it computes multipole coefficients C_ℓ of the matter power spectrum in spherical harmonics.

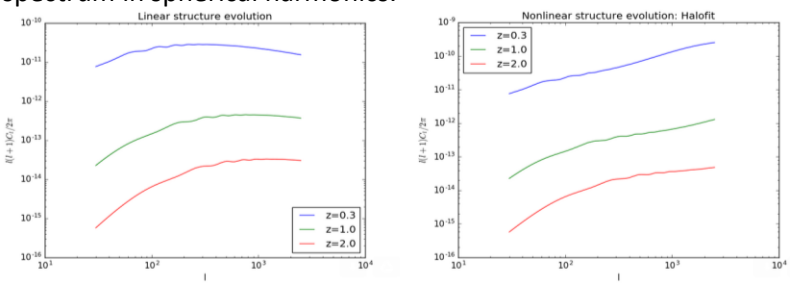

Figure 10: Multipole Power Spectra Predicted with CAMB

Linear (Left) and Nonlinear (Right) power spectra $\frac{l(l+1)}{2\pi}C_l$ for redshifts z=0.3, z=1, & z=2

As seen in Figure 10, the C_l were initially only computed between $\boldsymbol{l = 30}$ and $\boldsymbol{l = 3000}$, but because we'd had good success with our correlation functions/PwD PDFs using 400 bins over a maximum in-window separation of 1.4°, or in other words an angular precision of $1.4\ deg/400\ bins \approx .0036\ deg/bin$, we extended our spectra out to $l_{max} \approx 5 * 10^4$, since at this range we approached the Nyquist criteria $l_{max} \approx 180^o/.0036^o$ necessary to avoid aliasing issues.

Since these multipole coefficients are calculated at large ℓ, and we are restricting our analysis to small angular scales, we used the Limber approximation to map the C_ℓ to the corresponding coefficients $\Delta^2_\theta[k]$ for the flat-sky power spectrum [ix]:

$$k \approx l, \qquad \Delta^2_\theta[k] \approx \frac{(2\,l+1)^2}{4\,l\,(l+1)}\,C_l \qquad \textbf{(Eq. 6)}$$

Upon making this approximation, we constructed an interpolated function from the discrete $\Delta^2_\theta[k]$ values output from the simulations, then numerically integrated using the Hankel inverse transform defined in **(Eq. 4)**, finally finagling angular two-point autocorrelations out of the simulation.

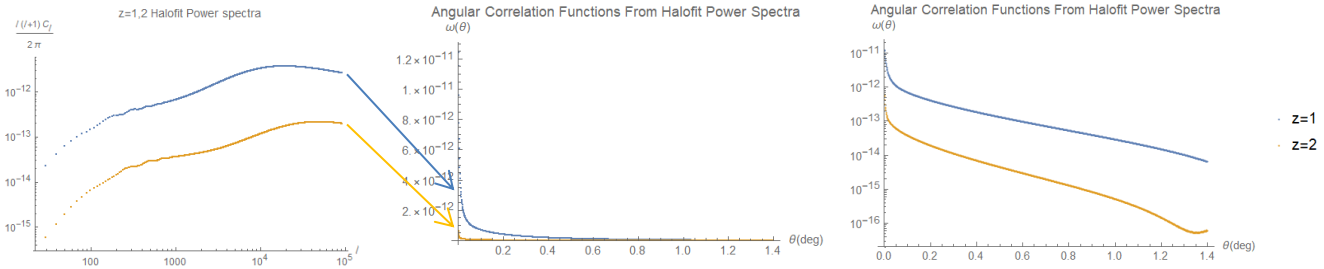

Figure 11: Nonlinear power spectra (Left) and transformed autocorrelations, with linear (Center) and log (Right) scale vertical axes.

We found that the nonlinear power spectra calculated from Halofit (with C_ℓ calculated from ℓ = 30 to ℓ = 100,000) resulted in correlation functions that most accurately resembled those which we had calculated empirically from COSMOS[3].

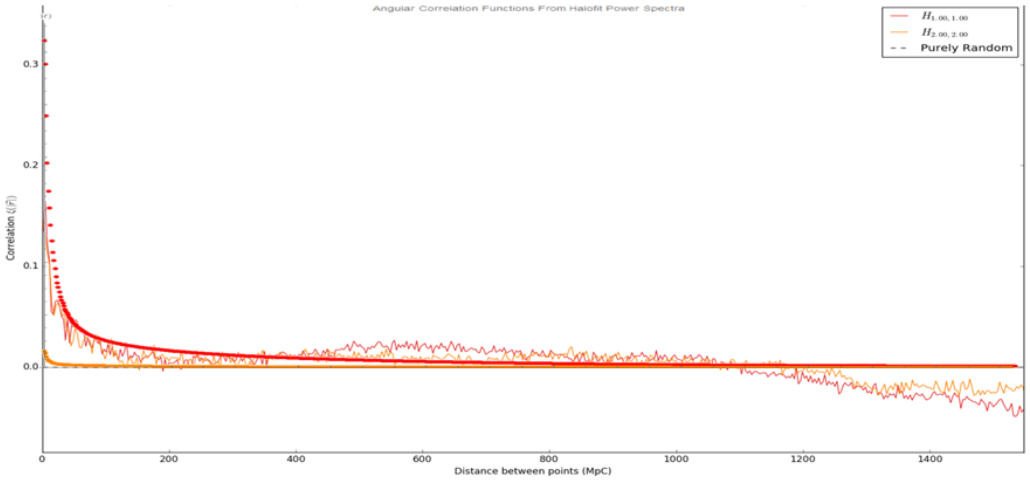

Figure 12: Overlay of z=1 and z=2 Autocorrelation Functions. The spiky pair of curves are from COSMOS data, while the cleaner curves result from inverse Hankel transforming Halofit power spectra.

This procedure for generating correlation functions from simulation was supposed to be an important development that increased our decomposition method's domain of applicability, because it could be used to estimate the correlation functions (and therefore PwD PDFs) of galaxies at *any* redshift, not just redshifts for which we already have photometric redshift estimates.

[3] Up to some (large) overall scaling factor, which probably arose from the numerical integration implementation.

4.4. Systematic Errors/Conclusions

Unfortunately, it wasn't meant to be. It seems that, in addition to the statistical variation, a large component of the correlations calculated from COSMOS should actually be attributed to systematic errors caused by telescope/sample-specific effects (e.g. survey selection function, masking, dust, cosmic variance, Finger of God). This would explain why the autocorrelations in figure 9 all look so similar, and why cross-correlations appear at all; any true measure of structure is just a small fraction of the total signal, if any.

Of course, many of these error sources can be overcome by a careful, thorough analysis of the survey's biases: By multiplying a simulated power spectrum by a survey's (approximated) selection function prior to taking the inverse Hankel transform, the selection bias can be baked right into the predicted correlation functions (convolution theorem). A library such as AstroStomp – if you can get it set up[4] – can infer a survey's masking function, and allows you to account for masked patches in your analysis. Surveys of intergalactic dust could conceivably enable you to correct for biases caused by redshift-dependent dust absorption.

But all of those corrections require a substantial amount of work. And worse still, the corrections would have to be applied on a per-telescope basis, which would defeat our goal of developing a method that can be generically applicable to any given beam survey. Perhaps if the selection bias correction can be systematized similar to how AstroStomp systematically corrects for masking, and if our intergalactic dust maps become more comprehensive, then the feasibility of our technique could be something to reconsider. But as things stand presently, we would be better served to continue looking for novel redshift estimators.

[4] We couldn't figure out how to set up AstroStomp, despite several days' worth of effort…

5. References

[i] Planck Collaboration, 2015, arXiv:1507.02704
[ii] M. Pettini: Introduction to Cosmology — Lecture 10
[iii] G. Djorgovski, lecture notes
[iv] A. Kravtsov (U. Chicago) (Here's a cool animated version: https://www.youtube.com/watch?v=8C_dnP2fvxk)
[v] J. Gott, A. Melott, M. Dickinson, 1986
[vi] Laigle et al. 2016
[vii] PHILIP, JOHAN. (2018). THE PROBABILITY DISTRIBUTION OF THE DISTANCE BETWEEN TWO RANDOM POINTS IN A BOX.
[viii] P. Murdin, 2001, Encyclopedia of Astronomy and Astrophysics
[ix] J. Peacock, 1998, Cosmological Physics
[x] Mohamed Rameez, July-August 2018, personal correspondence