PROBING DARK ENERGY WITH LARGE GALAXY SURVEYS: SYSTEMATICS QUANTIFICATION & MITIGATION

By

HUMNA AWAN

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Physics & Astronomy

Written under the direction of

Eric Gawiser

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2020

ABSTRACT OF THE DISSERTATION

Probing Dark Energy with Large Galaxy Surveys: Systematics Quantification & Mitigation

By HUMNA AWAN

Dissertation Director:

Eric Gawiser

Dark energy is a leading theory to explain cosmic acceleration, and forthcoming astronomical surveys have been specifically designed to probe this mysterious energy component of our universe. This thesis addresses aspects of using large galaxy surveys to study dark energy, which requires an unprecedented understanding and mitigation of systematics – a challenge that can be addressed on two fronts: quantification of the impacts of systematics, and new tools to mitigate them. Here, we specifically study the impacts of three key systematics: those induced by 1) the telescope observing strategy, 2) the Milky Way dust, and 3) uncertain photometric redshifts. Focusing on the Legacy Survey of Space and Time (LSST) carried out by the Vera C. Rubin Observatory, we quantify the impacts of LSST observing strategy on large-scale structure studies, which is a probe of dark energy. We demonstrate the effectiveness of large translational dithers – telescope-pointing offsets – in increasing LSST survey uniformity and reducing systematic uncertainties (Awan et al., 2016; LSST Science Collaboration et al., 2017) – a result that has now been adopted for the baseline LSST observing strategy. We also study the impacts of Milky Way dust on dark energy science and demonstrate that ∼25% of the default LSST survey area would not be useful for extragalactic static science given the Milky Way dust extinction, motivating the reconfiguration of the LSST survey footprint to avoid high-extinction regions of the sky (Lochner et al., 2018; Olsen et al., 2018). And finally, we present a new formalism that provides a novel way to correct for redshift contamination arising from

photometric redshift estimation (Awan & Gawiser, 2020). Specifically, we first introduce a general formalism to correct for sample contamination for photometric galaxy samples when measuring two-point angular correlation functions, and then a new weighted estimator that assigns each galaxy a weight in each redshift bin based on its probability of being in that bin, thereby fully utilizing the probabilistic distance information available for photometric galaxies. While these techniques are motivated by preparations for LSST, they are applicable to other large galaxy surveys like Dark Energy Survey (DES), Dark Energy Spectroscopic Instrument (DESI), Hobby-Eberly Telescope Dark Energy Experiment (HETDEX), Euclid, and Wide-Field Infrared Survey Telescope (WFIRST).

Acknowledgements

This thesis would not have been possible without so many things going just right – and while I consider myself very lucky to have been at the right place at the right time, there are specific people and entities who unarguably played a critical role in my success.

First, of course, my Ph.D. advisor, Eric Gawiser. There simply has been no encounter within my academic circle where I have not appreciated how incredibly spoiled I am to have had you as my advisor: from allowing me to follow the projects that piqued my interest to bearing with me as I asked you to repeat an explanation for the fifth time, you have been an indispensable support. There have been numerous times when I had no confidence in my abilities but having your support has always grounded me. Your commitment to student mentorship and exciting and impactful science inspires me, and I can only wish to follow in your footsteps.

I am also grateful to my thesis committee: Jack Hughes, Jolie Cizewski, Matthew Buckley and Phil Marshall. Jack, Jolie, Matt: thank you for always finding the time in your busy schedules for my annual meetings and keeping me on track. Phil: thank you for agreeing to be on my thesis committee and continuing your support outside of LSST DESC – I am very fortunate to have you as a mentor and to be able to bring power to junior researchers and motivate exciting science and hence great public talks.

I must also thank Andrew Baker who spearheaded the Rutgers REU program in which I was a participant in summer 2014 – the summer that I spent at Rutgers and realized how critical a good advisor can be and how a strong research community can shape one's interests. Not only that, Andrew has always been there for me, whether it was to address my fears of being an inadequate researcher or to find the right peer support.

Of course, here I must mention Ron Gilman who has been immensely supportive, not only in terms of enabling travel support to various conferences but also as a graduate program director; despite all his commitments, he always finds the time to hear out graduate students and their concerns, and works with them to find solutions. I truly thank him for this commitment and outlook, especially on the necessary but mundane details of a functional department. Also, I would like to thank Premi Chandra, who has been a beacon of inspiration when it comes to ensuring and creating inclusive research environments. I admire her tenacity in

constructing a supportive and amiable community in the department, especially for members of underrepresented groups.

I have also been very lucky to have met some of the most incredible graduate students: Willow Kion-Crosby, who made life as I knew it so much more palatable and gave me the treasure of a camaraderie, who would not only help me debug my code but also practice balance and go hike the world; Peter Doze who introduced me to painting minis and D&D, and hence a world of awesomeness; Charlotte Olsen, whose love for star-forming galaxies inspires me and whose commitment to the betterment of world motivates me to do the same; Kartheik Iyer, whose incredible research productivity sets a stretch goal for mine; Jack Hay, whose enthusiasm for microlensing, astronomy, and science in general is contagious and aspirational; and Yssa Camacho, Prasiddha Arunachalam, and Deepti Jain, who were my partners in crime as we struggled through classical mechanics together in our first year of graduate school, in our good old ARC offices.

Of course, I must thank Shirley Hinds who had to bear with me these five some years as I burdened her with tons of paperwork for my various intricate appointments and asked too many questions. I would also like to mention Katherine Lamb and Lisa Rivera, both of whom have been such a pleasant presence during my time at the department. You guys are awesome; thank you for taking care of everyone with all that goes unnoticed too often.

I would also like to thank Michael Scott and his office for uplifting my spirits so many times; to the staff at Gerlandas for amazing coffee and delish sandwiches to fuel my thoughts; to Git and GitHub for existing and making life so much easier; to LSST DESC for allowing me the space to thrive as a junior researcher and setting standards for healthy research collaborations; and to the Department of Physics and Astronomy here at Rutgers for teaching me invaluable skills and connecting me with invaluable people.

Finally, I would be nowhere without my family: my dad, whose commitment to hard work is instilled in me and motivates me; my mom, whose pursuit of knowledge has been a driving force behind mine; my sister, who I am ever so glad to have in my life as not only someone who I can share the agony of graduate school with but, of course, also the dire impacts of the broader patriarchal and colonialist world we live in; and my brother, whose strength and commitment to pursuing his aspirations despite hardships motivates me to do the same.

Here's to great friends and exciting science, without which this world will be a boring place .. and who'd want that?

Dedication

To doing what you love and hence never working a day in your life .. for everyone and not just the selected few.

Table of Contents

Chapter 1
Introduction

Cosmology is the study of the origin and evolution of our universe. In recent decades, this field has become quantitative, with more advanced technology allowing detailed observations and providing the computational power needed to analyze the vast amount of data. This thesis specifically focuses on the cosmic acceleration – one of the phenomena unexplained by our models of the universe. To provide a proper context and address the motivation of this work, we start with a brief overview of the discovery of cosmic acceleration in Section 1.1, followed by the current models explaining the phenomenon in Section 1.2. We then discuss one of the observational probes in Section 1.3 and provide the context for this thesis in Section 1.4.

1.1 An Accelerating Universe: Discovery

In 1929, Edwin Hubble measured distances to nearby galaxies and plotted them against their recessional velocities. He found that the galaxies farther away from us appear to be receding faster from us. The trend, known as the Hubble's law, indicates an expansion of the universe, and is expressed quantitatively as $v = H_0 r$, where v is the recessional velocity, H_0 is the Hubble constant, and r is the distance.

Since Hubble's sample consisted mainly of nearby galaxies, he was able to use the classical Doppler shift to infer velocities from redshift, $z = v/c$, where redshift measures the difference between the observed and emitted wavelengths:

$$z = \frac{\lambda_{\text{observed}} - \lambda_{\text{emitted}}}{\lambda_{\text{emitted}}} \tag{1.1}$$

If the galaxies are receding from each other today, they must have been closer together before. Therefore, the expansion of the universe applied reversed in time gave rise to the Big Bang model: an expanding universe that started from an infinitely dense region, marking the origin of the universe. The model, however, did not explain the nature of the expansion.

In 1998, two teams revealed another feature of our universe. Using Type Ia supernovae

(SNIa), Riess et al. (1998) and Perlmutter et al. (1999) extended the Hubble's law to high-z galaxies, as SNIa offer precise distance estimates to far away galaxies given that they are very bright objects with known intrinsic luminosities (Branch & Tammann, 1992) and their luminosities are standardizable using their light curve width (Phillips et al., 1999). Their results indicate an accelerating expansion of the universe, i.e., our universe is not only expanding, as Hubble had found, but the expansion rate has been increasing recently.

Figure 1.1 shows the results from Riess et al. (1998). The vertical axis shows the difference between the apparent magnitude m and absolute magnitude M, which is a direct measure of luminosity distance d_L:

$$m - M = 5 \log_{10} \left(\frac{d_L}{10\text{pc}} \right) \tag{1.2}$$

Luminosity distance is one of the measures of the cosmological distance to an object (Dodelson, 2003). Therefore, Figure 1.1 is an analog of the Hubble diagram. We note that in the high-z regime, the simple relation between the recessional velocity and redshift does not hold, as 'cosmological redshift' comes to play an important role (for a related discussion, see Bunn & Hogg, 2009). The lower panel in the figure makes clear that the galaxies farther away from us are at greater distances than their redshifts would imply for a non-accelerating universe (i.e., $\Omega_\Lambda = 0$).

We note that the Hubble constant is a function of time:

$$H = H(t) = \frac{da/dt}{a(t)}, \quad a(t) = \frac{1}{1+z} \tag{1.3}$$

where the scale factor $a(t)$ measures the expansion of the universe, and is taken to be unity today, i.e. $a_0 = a(t_0) = 1$; note that the scale factor is a dimensionless quantity. Therefore, the Hubble "constant" measures the rate of expansion at a given time.

Recent measurements of the Hubble constant today, H_0, range from $67.8 \pm 0.9\ \text{kms}^{-1}\text{Mpc}^{-1}$ (Planck Collaboration et al., 2016) to $73.2 \pm 1.7\ \text{kms}^{-1}\text{Mpc}^{-1}$ (Riess et al., 2016). Given the range of estimates, it has become customary to introduce a reduced Hubble constant h as

$$H_0 = 100h\ \text{kms}^{-1}\text{Mpc}^{-1} \tag{1.4}$$

and quote distance-dependent quantities in the units with h (Mo et al., 2010). We will return to the curves in Figure 1.1 later when we discuss how the high-z data favors a specific class of cosmological models.

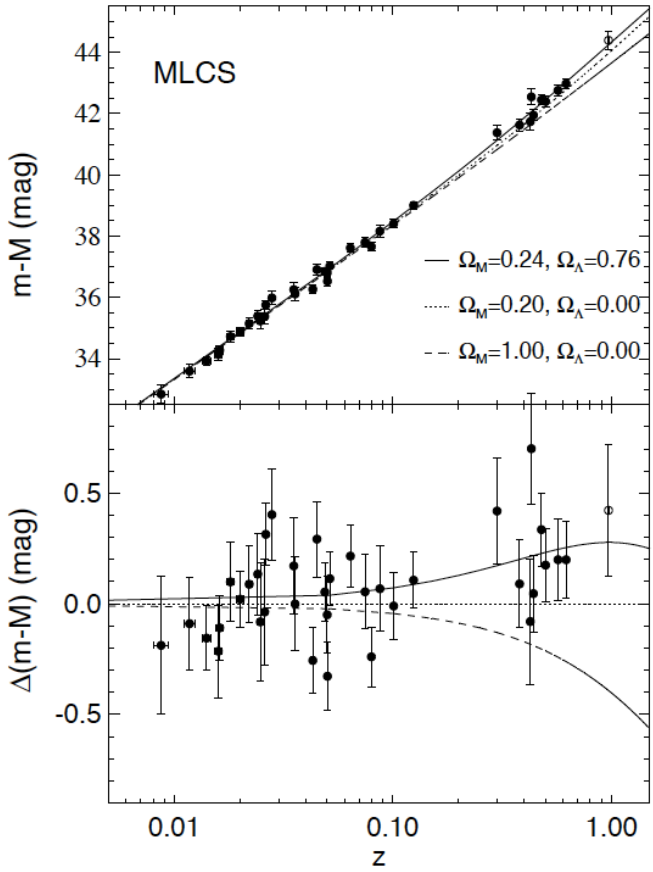

Figure 1.1: Figure 4 from Riess et al. (1998) showing the high-z Hubble diagram constructed using Type Ia supernovae measurements, showing an accelerating expansion of the universe; figure included here with permission. The vertical axis in the top panel shows the difference between the apparent magnitude m and absolute magnitude M, which is a measure of luminosity distance as in Equation 1.2, whereas the bottom panel plots the change in this quantity; the horizontal axis plots the redshift. The observed data is plotted as black dots, while the predictions for the relation based on three different cosmologies is shown as curves: the solid curve is for the cosmology with a non-zero energy density for Λ, the dotted one is for a universe with no Λ but not a matter-only universe, while the dashed line is for a matter-only universe. We see that the data prefers the cosmology with a non-zero energy density for Λ, hence indicating an accelerating universe.

1.2 An Accelerating Universe: Explanation

By the end of the 20th century, we knew that our universe is composed of at least two components: matter and radiation. Matter content not only includes the luminous, baryonic matter, but also the non-baryonic one, with the latter discovered by Vera Rubin et al. (Rubin et al., 1962; Rubin, 1965), and found to be dominating the matter content of the universe; observational data indicates that the non-baryonic matter interacts with baryonic matter only gravitationally and is non-relativistic (see more details in e.g., Mo et al., 2010), and hence is termed as *cold dark matter* (CDM), characterizing what is known of its nature of being non-luminous and non-relativistic. As for the radiation component, an isotropic presence, with small-scale inhomogeneities, is the

Cosmic Microwave Background (CMB) – discovered in 1964 (Penzias & Wilson, 1965), it is a relic from some 300,000 years after the Big Bang when photons decoupled from matter. The results indicating the expansion of the universe (Hubble, 1929; Riess et al., 1998; Perlmutter et al., 1999), however, suggest the existence of another component, dubbed *dark energy*, characterized by negative pressure that drives the expansion of the universe to accelerate.

In order to elaborate the requirements for an explanation of the expansion of the universe, we first consider the Einstein equations relating the geometry of the universe to its energy-momentum components:

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} \tag{1.5}$$

where the $G_{\mu\nu}$ is the Einstein tensor encoding the geometry of the universe, G is Newton's constant, and $T_{\mu\nu}$ is the energy-momentum tensor describing the different components of the universe; μ, ν are indices denoting the time coordinate and three spatial coordinates. Note that further details of Equation 1.5 and what follows can be found in e.g., Ryden (2003) and Carroll (2019).

For a flat universe (i.e., one without any intrinsic geometric curvature), Equation 1.5 leads to two Friedmann equations:

$$\left(\frac{\dot{a}}{a}\right)^2 = H^2 = \frac{8\pi G}{3}\rho \tag{1.6}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3\mathcal{P}) \tag{1.7}$$

where ρ is the energy density and $\mathcal{P}$ is the pressure of the fluid; a is the scale factor introduced in Equation 1.3, and $\dot{a}$ and $\ddot{a}$ are its time derivatives. Here, we note that Equation 1.6 defines the critical density ρ_{critical}, i.e., the energy density of a flat, non-expanding universe.

Modeling the fluid pressure as a function of density, with the equation of state parameter w, we have $\mathcal{P} = w\rho$. Applying the energy-momentum conservation, we arrive at the relation

$$\rho_i \propto a^{-3(1+w_i)} \tag{1.8}$$

For matter (both baryonic and non-baryonic), the density is inversely proportional to volume, so $\rho_m \propto a^{-3}$. On the other hand, the radiation energy density ρ_r is the photon number density times the photon energy; while the former is inversely proportional to volume, the latter is inversely proportional to the scale factor, accounting for the redshifting of the wavelength with the expansion of the universe. Therefore, $\rho_r \propto a^{-4}$.

For an accelerated expansion, $\ddot{a} > 0$. This combined with Equation 1.7 implies that any energy component driving the accelerating expansion of the universe must have

$$\rho + 3\mathcal{P} = \rho + 3w\rho < 0 \qquad \Rightarrow w < -\frac{1}{3} \tag{1.9}$$

The simplest model for dark energy considers a cosmological constant Λ with equation of state parameter $w = -1$ and hence a constant energy density (for details, see e.g., Peebles & Ratra, 2003).

In order to discuss and measure the density of different energy components, it is convenient to consider their density ratios versus the critical density:

$$\Omega_i \equiv \frac{\rho_i}{\rho_{\text{critical}}} \tag{1.10}$$

where i denotes different components, e.g., physical contributors like the dark energy (denoted as DE; with the specific case of a cosmological constant denoted as Λ), matter (m), baryons (b), and radiation (r). Also, since $\sum_i \Omega_i = 1$ to account for all the energy components of the universe, we have $\Omega_k \equiv 1 - \Omega_{\text{physical}}$, which is attributed to the geometry of the universe.

We can now vary the density parameters in our model and check its compatibility with observed data. In Figure 1.1, Riess et al. (1998) plot the Hubble relation for three different universes: matter-only universe $\Omega_m = 1$, $\Omega_m = 0.2$ universe without a cosmological constant, and the $\Omega_m = 0.24$, $\Omega_\Lambda = 0.76$ universe. Comparing the models with the data, we find that the high-z SNIa measurements strongly rule out the matter-only universes. Similar constraints are achieved when comparing the theoretical CMB prediction with the observed data.

1.2.1 Dark Energy

While the cosmological constant model for DE satisfies the parameter required to explain the accelerating expansion, as dictated by Equation 1.9, and leads to predictions that are in good agreement with the observed data, it leads to a formidable discrepancy when we consider the physical nature of the cosmological constant. By definition, the energy density of Λ remains constant. The only known kind of energy density that remains constant is the vacuum energy density – arising from quantum fluctuations in empty space, with virtual particle, antiparticle pairs zipping in and out of existence. While an exact theoretical estimate of the vacuum energy is not available, an order of magnitude estimate of this vacuum energy comes from the smallest length scale, the Planck length ℓ_{planck} and the related Plank energy E_{planck} (Ryden,

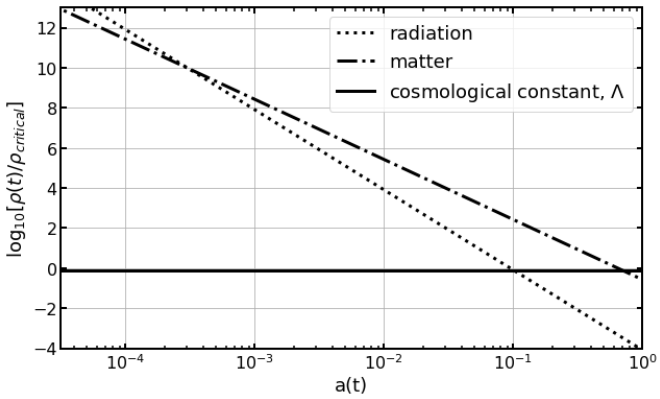

Figure 1.2: Figure modeled after Figure 1.3 from Dodelson (2003), showing the evolution of the different components in a flat universe. Specifically, the y-axis shows the density of each of the energy components as a function of the scale factor plotted on the x-axis.

2003; Carroll, 2001):

$$\rho_{\rm vac} \sim \frac{E_{\rm planck}}{\ell^3_{\rm planck}} \sim \mathcal{O}(10^{110}) \text{ ergs cm}^{-3} \tag{1.11}$$

On the other hand, observational data constrains the ratio of dark energy density to the critical density (e.g. Planck Collaboration et al., 2016):

$$\Omega_\Lambda = \frac{\rho_\Lambda}{\rho_{\rm critical}} \approx 0.7 \quad \Rightarrow \rho_\Lambda \approx \mathcal{O}(10^{-10}) \text{ ergs cm}^{-3} \tag{1.12}$$

The comparison of Equations 1.11-1.12 shows an $\mathcal{O}(10^{120})$ discrepancy between the observed and (potential) theoretical estimate of the energy density associated with the cosmological constant; this disagreement, known as *the cosmological constant problem*, marks a fundamental problem with the cosmological constant.

Furthermore, in the ΛCDM framework, we appear to be at a special point in time in the course of the evolution of the universe: Ω_Λ has only recently taken over Ω_m, while remaining subdominant to radiation and matter since inflation. Figure 1.2 shows the evolution of the different components of the universe, highlighting the peculiarity of the current state of the universe; this *cosmic coincidence* (Sahni, 2002) also generates criticism of the cosmological constant model for DE.

Since the cosmological constant leads to rather uneasy conclusions, other models are proposed for the dark energy equation of state, e.g. quintessence, which proposes an evolving equation of state parameter, e.g., $w = w_0 + w_a(1 + a)$ (see e.g., Barboza & Alcaniz, 2008). Yet other models propose modified gravity, i.e., departures from Einstein's theory of general relativity on large scales, and therefore do not introduce a different type of energy to explain the expansion of the universe (see e.g., Jain & Zhang, 2008).

In summary, the observational data reveal characteristic features of our universe, and we test our current cosmological model by measuring some of its key parameters. These include the dark energy density $\Omega_{\rm DE}$, baryon density Ω_b, matter density Ω_m, curvature density $\Omega_k = 1 - \Omega_m - \Omega_{\rm DE}$, and the DE equation of state parameters w_0, w_a.

1.3 Large-Scale Structure as a Probe

Various large galaxy surveys have revealed the large-scale structure (LSS) of the universe. Referred to as the *cosmic web*, LSS consists of galaxies, galaxy clusters and super-clusters forming 1D filaments that combine to form 2D sheets, separated by large nearly-empty spaces termed as voids. This peculiar arrangement hints at a hierarchal growth of structure: as matter gravitates into the potential wells set up by overdense regions, these regions get bigger and eventually merge, giving rise to the filamentary structure we see today (Cooray & Sheth, 2002).

Studying the evolution of LSS from initial conditions is a strong probe to study the expansion and acceleration of the universe, especially with the increased computational power which allows running detailed simulations of how the structures evolve and how the different components in the universe affect the growth of structure. This is shown in Figure 1.3, where the left panel shows the LSS mapped by the Sloan Digital Sky Survey (SDSS) (York et al., 2000), while the right panel shows the structure simulated by the Millennium Simulation (Springel et al., 2005). The details of both panels allow a comparison of the theoretical model predictions with observations, hence constraining the cosmological parameters in our models.

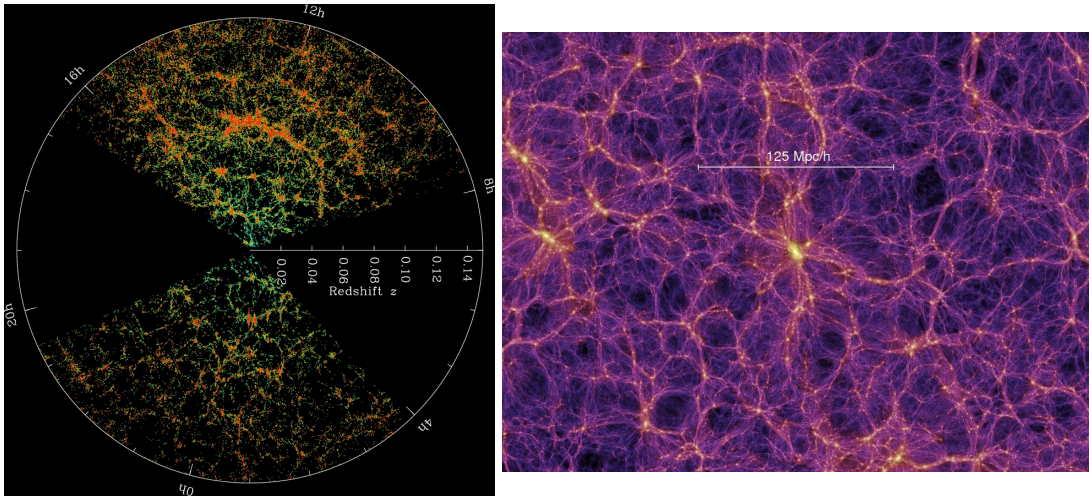

Figure 1.3: *Left*: Large-scale structure as mapped by the Sloan Digital Sky Survey (SDSS); image included here given the Creative Commons Attribution license. *Right*: Simulated large-scale structure from the Millennium Simulation (Springel et al., 2005); image included here with permission.

1.3.1 A Cosmological Ruler: Baryonic Acoustic Oscillations

A specific cosmological signature in large-scale clustering of galaxies is the Baryonic Acoustic Oscillations (BAO), arising due to an inherent scale introduced by the photon-baryon interaction in the early universe. Before recombination, when the photons and baryons were tightly coupled, their interaction gave rise to acoustic waves. Pressure drove the baryon-photon plasma outwards from the overdensities, before it returned to the overdensities due to gravitational pull. However, as the photons decoupled from neutral matter, they free-streamed instead of gravitating back with the baryons, leaving the baryonic shells 'frozen' as they were no longer driven out by pressure. These shells evolved under gravity, preserving the comoving scale of the original matter inhomogeneity as well as the baryonic shell (Eisenstein & Hu, 1998; Eisenstein et al., 2007).

The radius of the baryonic shells is the *sound horizon*, s, at recombination and marks a characteristic scale length; it is quantitatively defined as

$$s = \int_{z_{\rm rec}}^{\infty} dz \, \frac{c_s}{H(z)} \tag{1.13}$$

where $z_{\rm rec}$ is the redshift at recombination (~ 1000) and c_s is the speed of sound. Therefore, studying the BAO scale at various epochs can be instrumental in mapping out the expansion of the universe.

Specifically, measurement of the BAO signal provides direct constraints on the Hubble constant $H(z)$ and the angular diameter distance $d_A(z)$, which is a distance measurement based on the apparent angular size of an object (in this case, the BAO feature). More specifically, we can probe both the transverse and the line-of-sight (LOS) BAO scale s, and get two relations. The transverse measurement relates the angular scale of the BAO signal and $d_A(z)$ (Dodelson, 2003, Chapter 2):

$$\Delta\theta_{\rm BAO} = \frac{s}{d_{\rm A}(z)} \tag{1.14}$$

In a flat universe, $d_A(z)$ is defined as (Noh, 2013; Koehler, 2009)

$$d_{\rm A}(z) = \frac{c}{1+z}\int_0^z \frac{dz'}{E(z')}, \quad E(z) = \sqrt{\Omega_m(1+z)^3 + \Omega_{\rm DE}\exp\left[3\int_0^z dz\,\frac{1+w(z)}{1+z}\right]} \tag{1.15}$$

On the other hand, Hubble's law relates the LOS measurement of s to the change in redshift

and the Hubble constant (Hamilton, 1998; Dodelson, 2003):

$$c\Delta z_{\rm BAO} = sH(z) \tag{1.16}$$

Therefore, a measurement of the BAO signal can independently constrain the angular diameter distance through Equation 1.14, and the Hubble constant via Equation 1.16 (see e.g., Kazin et al., 2012). Since both $d_A(z)$ and $H(z)$ depend on the evolution of the universe, measuring the BAO signal at different epochs allows strong constraints on the evolution of the universe.

1.3.2 Two-Point Statistics and Systematics

Various statistics can be used to study the LSS, most common of which are two-point statistics like the two-point correlation function and its Fourier transform, the two-point power spectrum. Specifically considering photometric surveys, where redshift information is not the most precise and hence disallows precise measurements in 3D space, it is common to use 2D statistics – namely, the angular correlation function ($w(\theta)$) and the angular power spectrum (C_ℓ) – which are calculated in (thin) redshift bins to track time-evolution and hence constrain cosmology (e.g., see Camacho et al. (2019) for an analysis for Dark Energy Survey (DES) (Flaugher, 2005), and Nicola et al. (2019) for that for Hyper Suprime-Cam (HSC) (Aihara et al., 2018)).

Now, we expect to measure the BAO signal in two statistics as the sound horizon leads to excess clustering on its specific scale (Eisenstein et al., 1998a; Eisenstein & Hu, 1998; Eisenstein et al., 1998b); for instance, it is measured in the DES galaxy sample (Abbott et al., 2019) while its first detection came from SDSS (Eisenstein et al., 2005) using the 3D galaxy correlation function (which is measured in the 3D comoving space, instead of the projected one). Given the imperative need to not only precisely locate the BAO peak but also measure its amplitude, it becomes critical to avoid systematic uncertainties on the scales of interests: e.g., for a galaxy sample at $z \sim 1$, the BAO angular scale is expected to be $\mathcal{O}(1)$ degree[1], which corresponds to the multipole range of $100 \leq \ell \leq 300$ [2].

Another point to note is the impacts of uncertainties associated with photometric redshifts. The current methodology to handle these redshift uncertainties in cosmological analyses relies on Bayesian forward modeling each point in the cosmological parameters space to predict measurements, which then allows constraining the parameters (e.g., see Abbott et al. 2019).

[1]Calculated as $\Delta\theta_{\rm BAO}(z) = d_{\rm physical,BAO}(z)/d_A(z)$, where $d_{\rm physical}(z) = d_{\rm comoving}(1+z)^{-1}$ and $d_A(z)$ is the angular diameter distance from Equation 1.15; here $d_{\rm comoving,BAO} \approx 110$Mpc.

[2]Since $\theta \sim \ell/180$, where θ is the angular scale and ℓ is the multipole.

Specifically, this approach focuses on predicting the observed two-point statistics in redshift bins as long as the true number of galaxies in each redshift bins is correct (which can be done by calibrating these numbers using e.g., spectroscopic surveys). While this approach yields competitive constraints, it disallows handling the redshift uncertainties for each galaxy – a problem that may be a limiting factor as we get more precise redshift uncertainties with the upcoming surveys.

1.4 Motivation for this work

As discussed above, the statistics used to quantify LSS suffer from statistical and systematic uncertainties. The former is overcome by the forthcoming astronomical surveys given access to an unprecedented amount of data. These surveys include the Legacy Survey of Space and Time (LSST) carried out by the Vera C. Rubin Observatory (LSST Science Collaboration et al., 2009), Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al., 2016), HETDEX (Hill et al., 2008), Euclid (Laureijs et al., 2011), and WFIRST (Spergel et al., 2015). Since these surveys begin an era where, for the first time, science probes of cosmic acceleration like LSS will be systematics-limited as opposed to being statistics-limited, it is imperative to analyze the sources of systematic uncertainties in our measurements and find ways to mitigate them.

This thesis focuses on two specific aspects of using LSS studies with large galaxy surveys: understanding survey systematics and their impacts on our measurements, and developing statistical tools to fully realize the data-driven understanding of our universe. Specifically, we consider the impacts of three specific systematics that affect our data: the telescope observing strategy, Milky Way dust, and redshift uncertainties; these are discussed in detail in Chapters 2, 3, and 4 respectively. Also, since the work here is motivated by preparations for LSST, we provide a brief overview of the survey below.

1.4.1 Legacy Survey of Space and Time

The Legacy Survey of Space and Time (LSST) is a 10-year optical survey that will be carried out by the Vera C. Rubin Observatory during 2022-2032. During its 10-year term, the Observatory will scan the full southern sky with unprecedented detail, with the final survey covering 20,000 deg^2 in six different filters (spanning $\sim$300-1100 nm in wavelength space) and measuring angular positions and photometric redshifts of over 10 billion galaxies (LSST Science Collaboration et al., 2009). Specifically pertinent to dark energy science is the ability to use the same

dataset for different astrophysical probes of cosmic acceleration – large-scale structure, Type-Ia supernovae, weak gravitational lensing, strong lensing, and clusters of galaxies – thereby, enabling competitive combined probe analyses, alongside those from individual probes given the unprecedentedly large statistical sample. Figure 1.4 shows the forecast constraints on the dark energy equation of state from the ten-year LSST data; these are based on the Science Road Map (The LSST Dark Energy Science Collaboration et al., 2018) from LSST Dark Energy Science Collaboration (DESC), which is the collaboration focused on carrying out the research and development for dark energy constraints using LSST data.

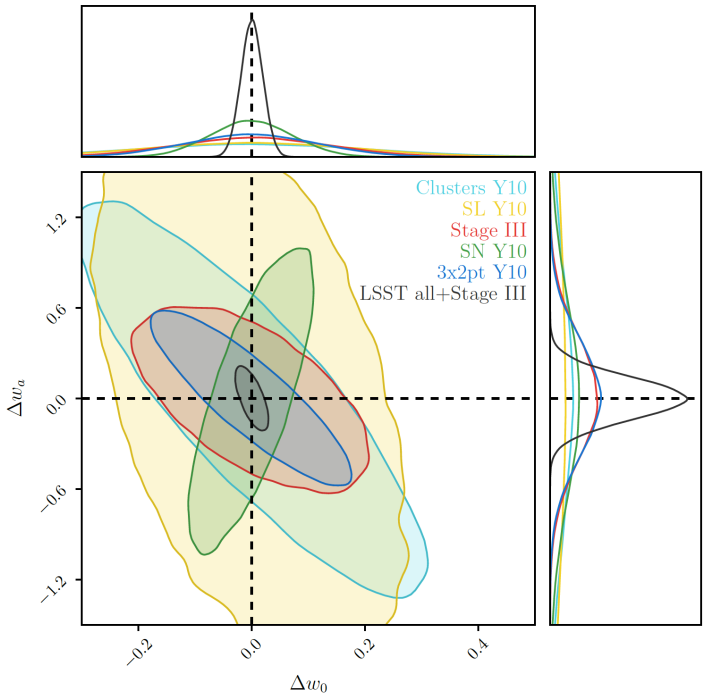

Figure 1.4: Figure adapted from Figure G2 in The LSST Dark Energy Science Collaboration et al. (2018) showing the forecast constraints on dark energy equation of state from the ten-year (Y10) LSST data; included here with permission. As discussed in Section 1.2.1, the DE equation of state can be characterized by $w = w_0 + w_a(1 + a)$, and this figure plots the constraints on the two equation of state parameters w_0 and w_a using individual DE probes (cyan, yellow, red, green contours), the joint static probes analysis (blue contour) and joint all-probes analysis (black contour).

Chapter 2

LSST Observing Strategy Systematics and Translational Dithering

This chapter is reproduced, aside from minor formatting changes and an additional appendix (2.B), from Awan et al. (2016): published in the Astrophysical Journal © AAS; titled *Testing LSST Dither Strategies for Survey Uniformity and Large-Scale Structure Systematics*; authored by Humna Awan, Eric Gawiser, Peter Kurczynski, R. Lynne Jones, Hu Zhan, Nelson D. Padilla, Alejandra M. Muñoz Arancibia, Alvaro Orsi, Sofía A. Cora, and Peter Yoachim. Reproduced here with permission.

Appendix 2.B, aside from its preface, is reproduced with permission from LSST Science Collaboration et al. (2017), a Community White Paper, titled *Science-Driven Optimization of the LSST Observing Strategy*, and authored by Phil Marshall, et al. (including Humna Awan) for the LSST DESC Collaboration. The specific source of the appendix is subsections 9.2.3-9.2.5, where Section 9.2 was authored by Humna Awan, Eric Gawiser, Peter Kurczynski, and Lynne Jones.

Note also that the content here is updated to use the new terminology: LSST now stands for the Legacy Survey of Space and Time, which is carried out by the Vera C. Rubin Observatory, which was previously known as the Large Synoptic Survey Telescope (abbreviated LSST).

2.1 Introduction

The Legacy Survey of Space and Time (LSST) is an upcoming wide-field deep survey, carried out by the Vera C. Rubin Observatory (Rubin Obs.) and designed to make detailed observations of the southern sky. A telescope with an effective aperture of 6.7m and a 3.2 Gigapixel camera, Rubin Obs. will survey about 20,000 deg^2 of the sky in *ugrizy* bands, over the course of ten years with $\sim$150 visits in each band to each part of the survey area (LSST Science Collaboration et al., 2009). While the survey has various goals, from studying near-Earth objects to transient phenomena, its imaging capabilities are particularly promising for studying dark energy. With its wide-deep observation mode, LSST will probe 1) the shear field from weak gravitational lensing, 2) Baryonic Acoustic Oscillations (BAO) in the galaxy power spectrum and correlation functions, 3) evolution of the galaxy cluster mass function, 4) Type Ia supernovae and their distance-redshift relationship, and 5) time delays from strong gravitational lenses, providing an opportunity to study dark energy from one dataset. The nature of these cosmic probes leads to requirements on the survey observing strategy, understood in terms of cadence, i.e. frequency of visits in a particular filter, and uniformity, i.e. survey depth across various regions

of the sky. For goals dependent on spatial correlations, such as BAO and additional large-scale structure (LSS) studies, survey uniformity is of critical importance, while time domain science often depends on high cadence.

The baseline LSST observing strategy tiles the sky with hexagons, each of which inscribes an LSST field-of-view (FOV) (LSST Science Collaboration et al., 2009). Given that the FOV is approximately circular, the hexagonal tiling leads to regions between the FOV and the inscribed hexagon that overlap when adjacent fields are observed. Therefore, observations at fixed telescope pointings lead to deeper data in these overlapping regions, decreasing survey uniformity and inducing artificial structure specifically at scales corresponding to the expected BAO signal at $z \sim 1$ (Carroll et al., 2014). While the double-coverage data could be discarded to make the survey uniform, the loss would comprise nearly 17% of LSST data (Carroll et al., 2014), equivalent to 1.5 years of survey time. On the other hand, correction methods have been developed for other surveys (e.g., Ross et al., 2012; Leistedt et al., 2016) to post-process and correct for the systematics in the observed data – such an approach could also work for LSST survey uniformity. Here, however, we address the approach of minimizing eventual survey systematics by designing an optimal observing strategy.

Dithers, i.e. telescope pointing offsets, are helpful in reducing systematics. While LSST plans to implement small dithers to compensate for the finite gaps between the CCDs (e.g., McLean, 2008), implementing large dithers on the scale of the FOV appears to offer a solution for LSST survey uniformity, reducing the artificial structure by a factor of 10 as compared to the undithered survey (Carroll et al., 2014). In this paper, we analyze various dither strategies, varying in both the geometric pattern and the timescale on which the pattern is implemented. We develop a methodology for a quantitative comparison of these strategies and explore their effects on survey depth and BAO systematic uncertainty. We introduce the LSST Operations Simulator and the Metrics Analysis Framework in Section 2.2. Then, in Section 2.3, we describe the variants of the dithers implemented, followed by a discussion of the impacts of the dither strategies on the coadded depth as well as artificial fluctuations in galaxy counts in Section 2.4. We conclude in Section 2.5, highlighting that our work illustrates the capability to assess the effectiveness of various dither strategies for LSST science goals.

2.2 The LSST Operations Simulator and Metrics Analysis Framework

The LSST Operations Simulator (OpSim) simulates 10-year surveys, accounting for realistic factors that affect the final data; these considerations include scheduling of observations, telescope pointing, slewing and downtime, site conditions, etc. (Delgado et al., 2014). More specifically, OpSim output contains realizations of LSST metadata, stamped with sky position, time, and filter (LSST Science Collaboration et al., 2009), allowing post-processing of the output to simulate different dither strategies.

As mentioned earlier, LSST OpSim tiles the sky with hexagonal tiles. In order to effectively account for the overlapping regions between the hexagons, we utilize the Hierarchical Equal Area isoLatitude Pixelization (HEALPix) package to uniformly tile the sky with equal area pixels (Górski et al., 2005). HEALPix uses nearly-square pixels to tile the sky with a resolution parameter N_{side}, leading to a total number of pixels $N_{pixels} = 12N_{side}^2$. In our analysis, we use $N_{side} = 256$, giving a total of 786,432 pixels, and effectively tiling each 3.5° FOV with about 190 HEALPix pixels. Here we note that our resolution is four-fold higher than that used in Carroll et al. (2014); this improvement ensures that we do not encounter signal aliasing in the angular scale range we study here.

We carry out our analysis within the Metrics Analysis Framework (MAF), designed for the analysis of OpSim output in a manner that facilitates hierarchical building of the analysis tools. MAF consists of various classes, of which most relevant here are $Metrics$ that contain the algorithm to analyze each HEALPix pixel and $Stackers$ that provide the functionality of adding columns to the OpSim database; for details, see Jones et al. (2014). Some of our code has already been incorporated into the MAF pipeline[1], and the rest can be found in the LSST GitHub repository[2].

2.3 Dither Strategies

We consider dither strategies with three different timescales: by season, by night, and by visit. A single visit is a set of two 15 second exposures (Ivezic et al., 2008). Since OpSim output does not have a season assignment for the simulated data, we define seasons separately for each field, starting from zero and incrementing the season number when the field's RA is overhead in the middle of the day. This leads to 11 seasons for the 10-year data, and we assign the 0th

[1] https://github.com/lsst/sims_maf

[2] https://github.com/LSST-nonproject/sims_maf_contrib/tree/master/mafContrib

and the 10th seasons the same dither position.

Since fields are scheduled to be visited at least two times in a given night, followed by a typical revisit time of three days (Ivezic et al., 2008), we implement two approaches for the by-night timescale: 1) FieldPerNight, where a new dither position is assigned to each field independently, and 2) PerNight, where a new position is assigned to all fields. The first approach tracks each field and assigns it a new dither position only if it is observed on a new night, while the second approach assigns a dither position to all the fields every night (regardless of whether a particular field is observed or not). For the by-visit timescale, we only consider FieldPerVisit, and for by-season strategies, we consider PerSeason only.

For the dithers, we implement a few geometrical patterns to probe the effects of dither positions themselves. Since the sky is tiled with hexagons inscribed within the 3.5° FOV, we restrict all dither positions to lie within these hexagons. For by-season strategies, given that there are only 10 seasons throughout the LSST run, we pick a geometry that allows choosing 10 dither positions uniformly across the FOV:

– Pentagons: points alongs two pentagons, one inside an inverted, bigger pentagon.

For by-night and by-visit timescales, we consider four different geometries:

– Hexagonal lattice dithers: 217 points arranged on a hexagonal lattice (Krughoff, 2016).
– Random dithers: random points chosen within the hexagon such that every dither position is a new random point.
– Repulsive Random dithers: after creating a grid of squares inside the hexagon, squares are randomly chosen without replacement. Every dither position is a random point within a chosen square.
– Fermat Spiral dithers: 60 points are chosen from the spiral defined by $r \propto \sqrt{\theta}$, where θ is a multiple of the golden angle 137.508° (geometry appears in nature; see Muñoz et al. (2014)).

Figure 2.1 shows these geometries and the possible dither positions. We also considered some other variants. For by-season timescale, we implemented a PentagonDiamond geometry where the first point is at the center of the FOV, followed by 9 points arranged along a diamond circumscribed by a pentagon. We find that PentagonDiamond leads to results similar to Pentagons, and discuss only the latter here. We also considered Spiral dithers, where equidistant points are chosen along a spiral centered on the FOV and the number of points and coils can be varied, as well as variants of Fermat spiral, in terms of the number of points and θ as a

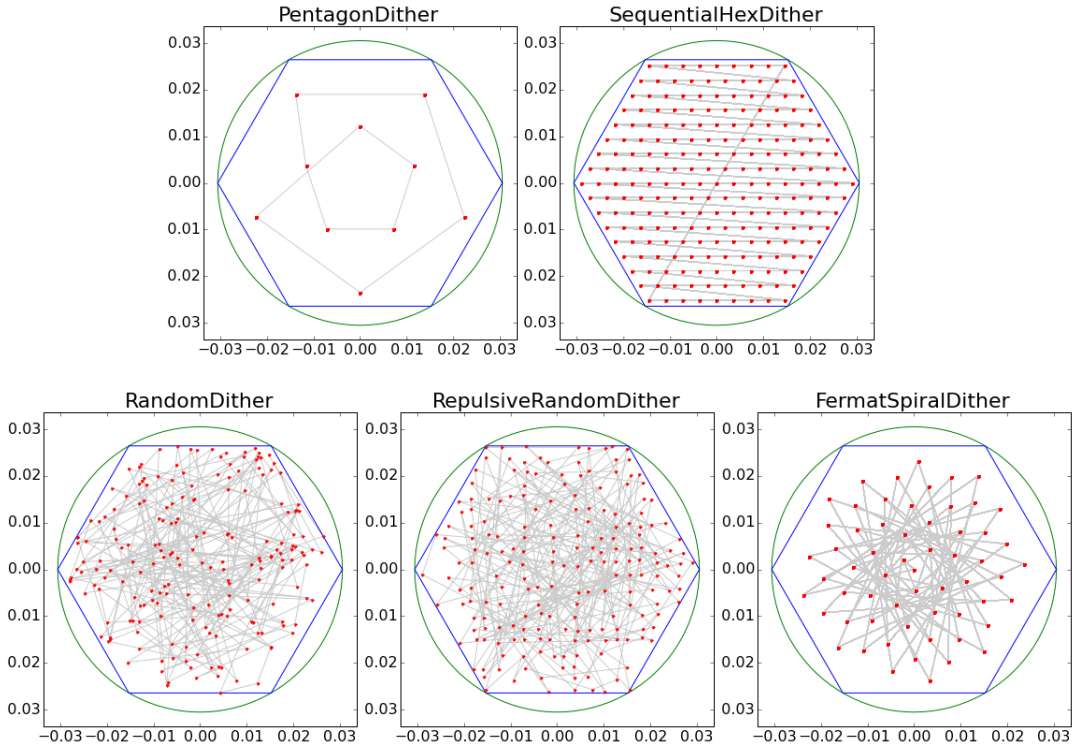

Figure 2.1: Dither geometries: PentagonDither is implemented only for per-season timescale, while the rest are implemented for per-visit, per-night and field-per-night timescales. The green curve represents the circular FOV with radius of 0.305 radians, the blue hexagon represents the hexagonal tiling of the sky originally adopted for the undithered observations, and the red points are the dither positions, connected with gray lines. The axes are labelled in radians. See Section 2.3 for details.

multiple of 77.508° or 177.508°. Our preliminary analysis shows that these spiral geometries behave similarly as the 60-point, golden-angle Fermat spiral described above.

To identify the various strategies, we follow a consistent naming scheme: [Geometry]Dither[Field]Per[Timescale], where the absence of 'Field' implies dither assignment to all fields, while its presence implies that each field is tracked and assigned a dither position independent of other fields. For instance, SequentialHexDitherPerNight assigns the new dither position to all fields every night, while SequentialHexDitherFieldPerNight assigns it to a field only when it is observed on a new night.

2.4 Analysis & Results

We use OpSim dataset `enigma_1189`[3], which includes the wide-fast-deep (WFD) survey region as well as five Deep Drilling fields; we focus only on WFD survey for our analysis. We implement various dithers within MAF by building *Stackers* corresponding to each dither strategy and post-processing the OpSim output to find the survey results using the dithered positions. First, we examine the r-band coadded depth (i.e. the final depth after the 10-year survey) as a function of sky location, followed by an analysis of the fluctuations in the galaxy counts, in order to probe the effects of dither strategies on large-scale structure studies.

2.4.1 Coadded 5σ Depth

In order to calculate the coadded depth, we use the modified 5σ limiting magnitude data from OpSim, where the limiting magnitude is 'modified' in order to represent a real point source detection depth (Ivezic et al., 2008). Assuming that the signal-to-noise ratio adds in quadrature, as it should for optimal weighting of individual images (see e.g., Gawiser et al., 2006), we calculate the coadded depth, $5\sigma_{\rm stack}$, in each HEALPix pixel from the modified 5σ limiting magnitude summed over individual observations, $5\sigma_{{\rm mod},i}$:

$$5\sigma_{\rm stack} = 1.25\log_{10}\left(\sum_i 10^{0.8\times 5\sigma_{{\rm mod},i}}\right) \tag{2.1}$$

We find that dithered surveys lead to shallower depth near the borders of the survey region, adding significant noise to the corresponding angular power spectra. In order to clean the spectra, we develop a border masking algorithm to discount pixels at the edges of the survey region, comprising nearly 15% of the survey area. See Appendix 2.A for details of the masking algorithm.

Figure 2.2 shows two projections for the r-band coadded 5σ depth for the various dither strategies after the shallow border has been masked. The first row shows the Mollweide projection of the coadded depth for NoDither and PentagonDitherPerSeason, while the second row shows the corresponding Cartesian projection, zoomed on the LSST WFD survey area ($-180°<$RA$<180°$, $-70°<$Dec$<10°$). To conserve space, we show only the latter projection for the rest of the dither strategies. We observe that the survey pointings without any dithering lead to deeper overlapping regions between the fields, and consequently a strong honeycomb pattern in the coadded depth. In contrast, the dithered skymaps have comparatively more

[3] https://confluence.lsstcorp.org/display/SIM/OpSim+Datasets+for+Cadence+Workshop+LSST2015

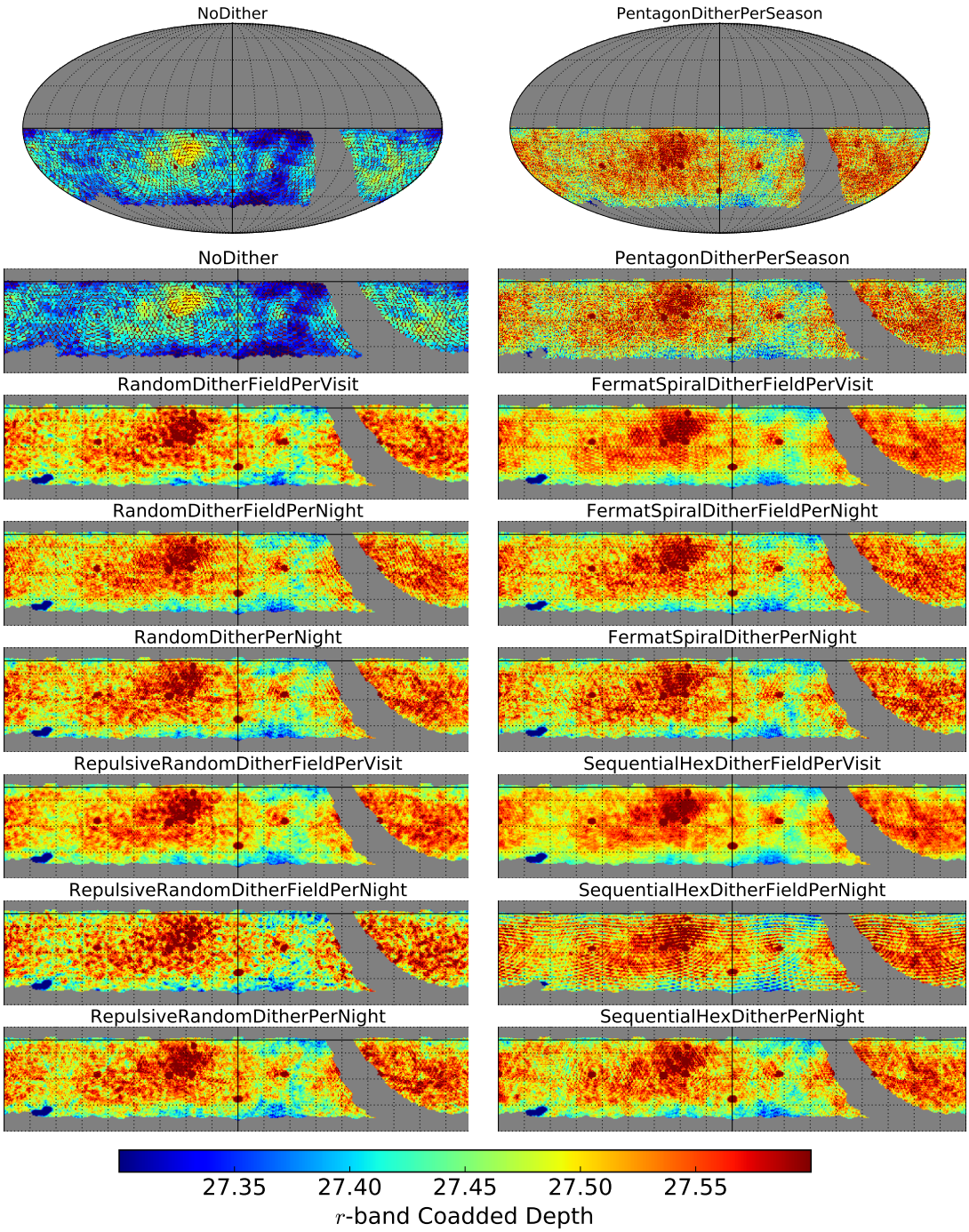

Figure 2.2: Plots for r-band coadded 5σ depth from various dither strategies, after masking the shallow-depth border. The top row shows the Mollweide projection for two observing strategies while the 2nd row shows the Cartesian projection restricted to 180°>RA>-180° (left-right), -70°<Dec<10° (bottom-top); we only show the latter for the rest of the strategies. We note that the strong honeycomb pattern present in the undithered survey is weaker in the dithered surveys, while for SequentialHexDitherFieldPerNight, we observe strong horizontal striping across the survey region. See Section 2.4.1 for further details.

uniform depth across the survey region, with smaller-scale variations amongst the dither strategies.

Here we note that although dithering in general weakens the honeycomb pattern seen in the undithered survey, we observe horizontal striping from SequentialHexDitherFieldPerNight; in contrast, SequentialHexDitherPerNight and SequentialHexDitherFieldPerVisit show no such behavior. This is an example where a specific dither strategy's behavior is highly dependent on the timescale on which it is implemented: for PerNight timescale, a new dither is assigned to all fields every night, implying that the 217-point lattice is traversed multiple times during the ∼3650-night survey. Similarly, for PerVisit timescale, although a new dither is assigned to each field every time it is visited, the lattice is traversed multiple times given that every field is visited ∼150 times in the r-band throughout the survey. In contrast, for FieldPerNight timescale, a new dither point is assigned to each field only when it is observed on a new night. Since a given field is only visited on ∼50 nights in a given filter, only the lower part of the lattice is traversed (as the lattice is traversed starting from bottom left), leading to horizontal striping. We verified this conclusion by rotating the hexagonal lattice by 90°, and observing vertical striping for FieldPerNight timescale.

In Figure 2.3, we show a histogram of the r-band coadded depth. We see that the undithered survey leads to a bimodal distribution, with the overlapped regions observed much deeper than the rest of the survey. On the other hand, all the dithered surveys lead to unimodal distributions, as dithering leads to observing the data more uniformly, in agreement with Carroll et al. (2014).

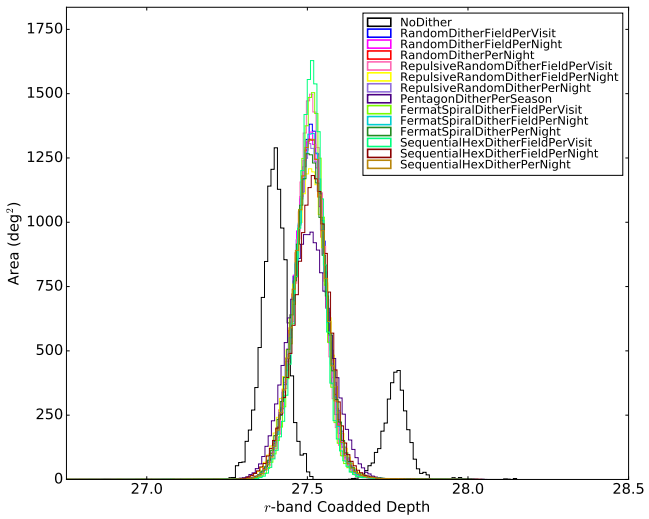

Figure 2.3: Histogram for the r-band coadded 5σ depth, indicating a bimodal distribution from the undithered survey, and unimodal distributions from the dithered ones.

In order to quantify the angular characteristics reflected in the skymaps, we measure the power spectrum associated with each of the skymaps. Figure 2.4 shows the power spectra for the coadded depth from each of the dither strategies considered here; we have removed the monopole and dipole using HEALPix routine remove_dipole. We note that the spectrum corresponding to the undithered survey has a very large peak around $\ell\sim$150, resulting from the strong honeycomb pattern. In comparison, we see over 10 times less power in the dithered surveys; the $\ell\sim$150 peak in these surveys is much more comparable to the rest of the spectrum. More specifically, we find that the FieldPerVisit timescale is the most effective in reducing the power for a given dither geometry, while Random and RepulsiveRandom dithers perform well on all three timescales. Also, we confirm the origins of the $\ell\sim$150 peak by creating a pure honeycomb, and observing a power spectrum similar to that from the undithered survey.

Furthermore, we see that the horizontal striping in the SequentialHexDitherFieldPerNight skymap generates a large peak around $\ell\sim$150, while the rest of the dithered spectra do not exhibit such a strong peak. Curiously, the PentagonDitherPerSeason strategy leads to two large peaks around $\ell\sim$270 – a characteristic different from the rest of the dither strategies' but similar to NoDither, with much less power.

To understand the origins of the characteristic patterns in the skymaps, we consider the $a_{\ell m}$ coefficients of their spherical harmonic transforms. This allows us to produce the skymaps corresponding to specific ranges of the angular scale ℓ. We show our results in Figure 2.5 for NoDither, PentagonDitherPerSeason and SequentialHexDitherFieldPerNight strategies. The top row includes the full power spectrum for each strategy, and the second row shows the corresponding Cartesian projection for $0^\circ<$RA$<50^\circ$, $-45^\circ<$Dec$<-5^\circ$. The third and fourth rows show the partial skymaps arising from each of the colored peaks shown in the power spectra in the top row. We observe that for the undithered survey, the $\ell\sim$150 peak arises from the strong honeycomb pattern, while the second peak arises from structure on the small angular scales. For PentagonDitherPerSeason, we see a milder honeycomb for the $\ell\sim$150 peak, while the $240<\ell<300$ peak arises from structure similar to the corresponding one in the undithered survey. Finally, for SequentialHexDitherFieldPerNight, we can see the source of the strong $\ell\sim$150 peak: the horizontal striping. For higher-ℓ peaks, we note the weaker structure as compared to the other two strategies. We also performed this $a_{\ell m}$ analysis individually for the two peaks in $240<\ell<300$ and found the underlying structure to be very similar.

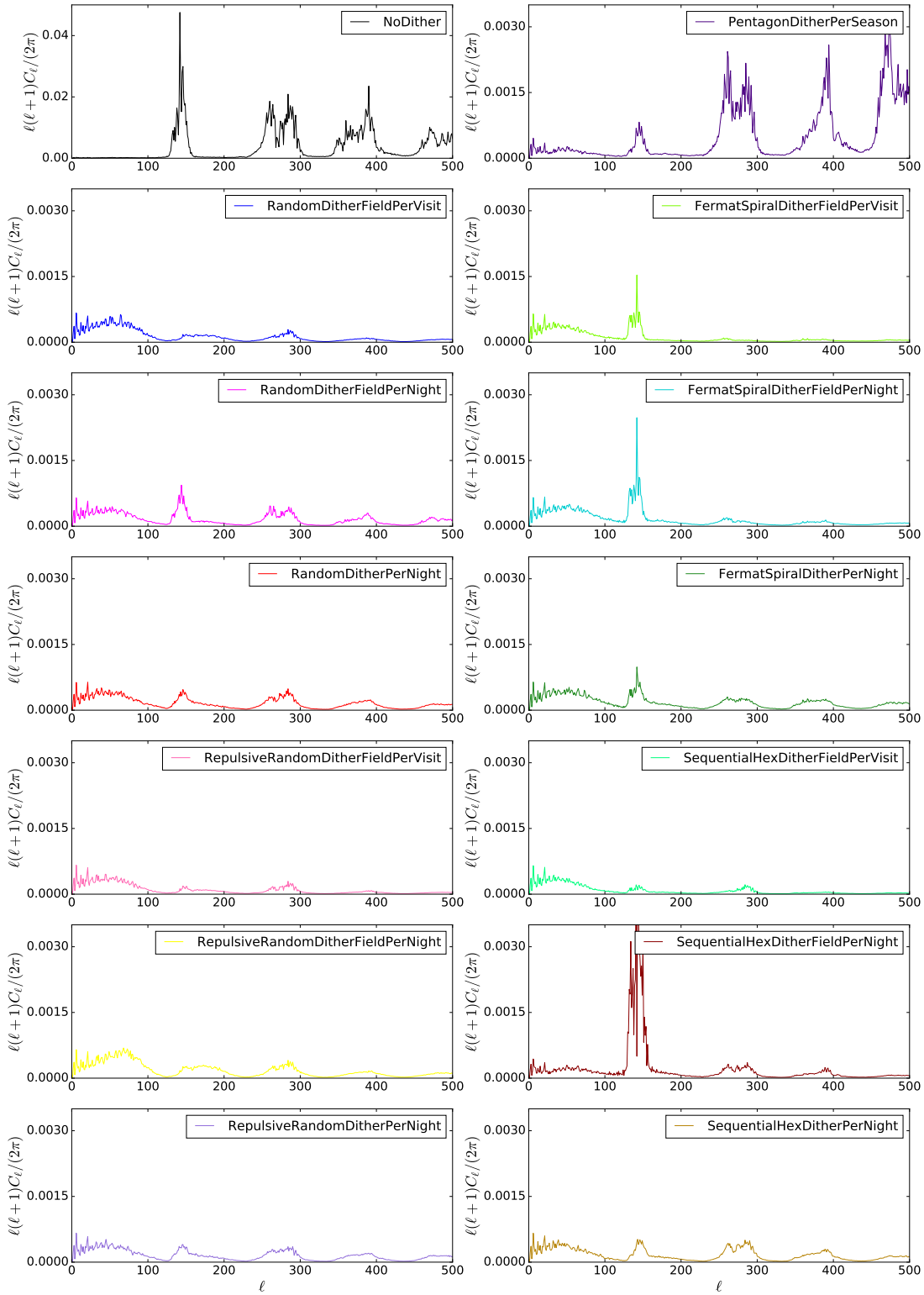

Figure 2.4: Angular power spectra for the r-band coadded depth for all the dither strategies. We note that dithering reduces the angular power by at least a factor of 10 as compared to NoDither. The honeycomb pattern in the undithered survey generates a large peak around ℓ∼150, while dithering of all kinds decreases the spurious power. The horizontal striping in SequentialHexDitherFieldPerNight also creates a moderate peak around ℓ∼150.

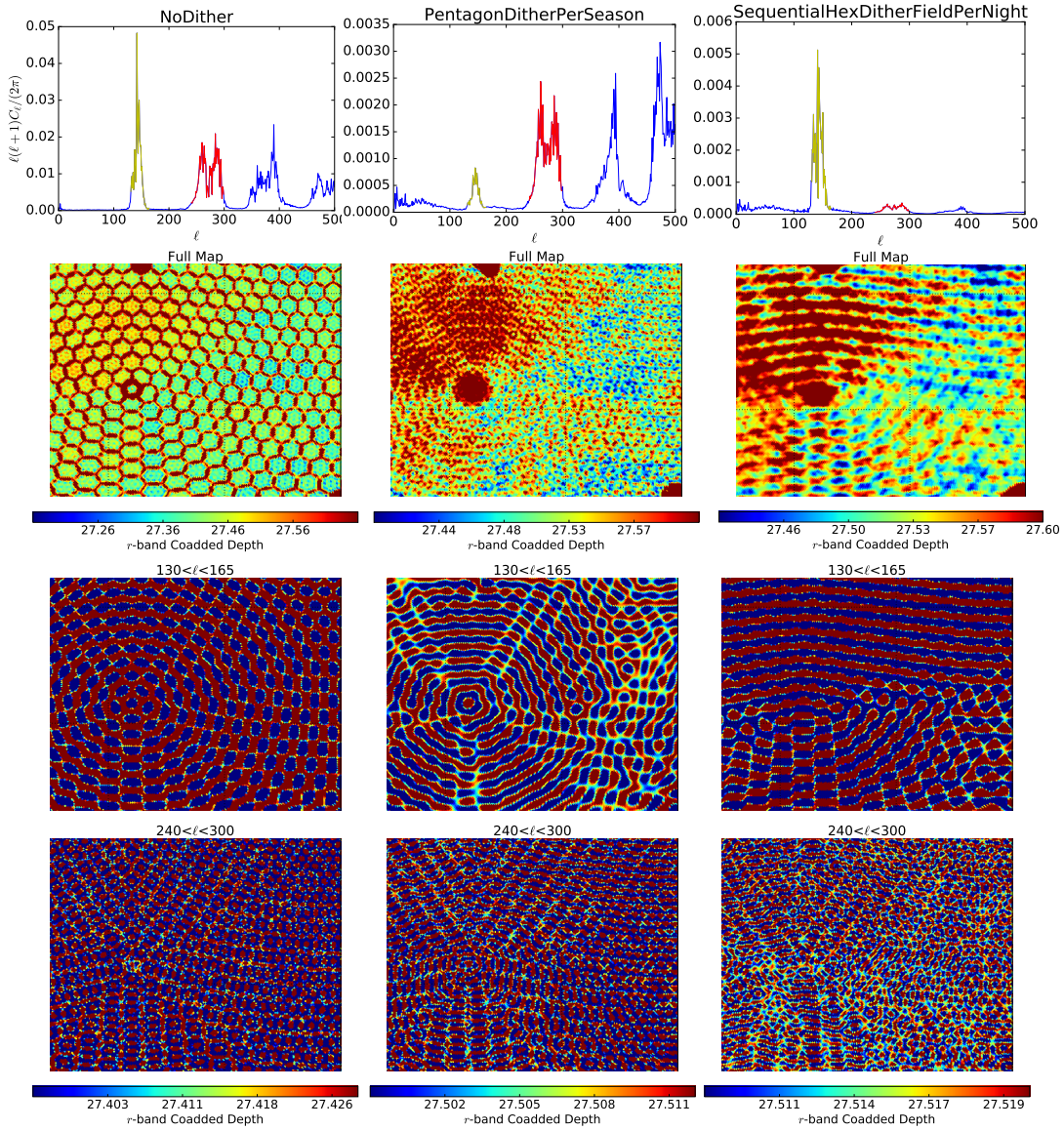

Figure 2.5: $a_{\ell m}$ analysis plots for two ℓ-ranges in the r-band coadded depth power spectra (colored peaks in the top row). The first row shows the full power spectrum for three observing strategies; the second row shows the corresponding skymaps for 50°>RA>0° (left-right), -45°<Dec<-5° (bottom-top). The third row is for 130<ℓ<165 (yellow in the power spectra in the first row),and the fourth is for 240<ℓ<300 (red in the top row), all in the same RA, Dec range as the second row. The leftmost column corresponds to NoDither, the middle one to PentagonDitherPerSeason, and the right one to SequentialHexDitherFieldPerNight. We see that the honeycomb pattern in the undithered survey and the horizontal striping in SequentialHex generates the $\ell\sim$150 peak. Also, we see one (partial) Deep Drilling Field at the top as well as a pentagonal tile at Dec= -30° resulting from the tiling of the sphere, both of which are smeared out by dithering.

2.4.2 Artificial Galaxy Fluctuations

Given our knowledge of the characteristics induced in the coadded depth due to the observing strategy, we now consider the effects of these artifacts on BAO studies. We model the artificial fluctuations in galaxy counts, accounting for photometric calibration errors, dust extinction, and galaxy catalog magnitude cuts. Since BAO studies are redshift dependent, we consider five redshift bins: $0.15<z<0.37$, $0.37<z<0.66$, $0.66<z<1.0$, $1.0<z<1.5$, and $1.5<z<2.0$.

We first estimate the number of galaxies in specific redshift bins detected in each pixel at a particular depth using a mock LSST catalog, which is constructed using the outputs of the SAG semi-analytic model for galaxy formation (Cora, 2006; Lagos et al., 2008; Tecce et al., 2010; Orsi et al., 2014; Gargiulo et al., 2015; Muñoz Arancibia et al., 2015). The model incorporates differential equations for gas cooling, quiescent star formation, energetic and chemical supernova feedback, the growth of a supermassive black hole, the associated AGN feedback, bursty star formation in mergers and disk instabilities, all coupled to the merger trees extracted from a dark matter simulation run with GADGET2 (Springel et al., 2005) assuming the standard ΛCDM model (Jarosik et al., 2011). The subhalo populations of merger trees are found using SUBFIND (Springel et al., 2001) after the DM haloes were identified using a friends-of-friends algorithm.

We normalize the total r-band galaxy counts to the empirical cumulative galaxy count estimates for LSST (see LSST Science Collaboration et al., 2009, Section 3.7.2 for details) at a magnitude cut $r<25.9$ (corresponding to the CFHTLS Deep survey completeness limit of $i<25.5$; see Hoekstra et al. 2006; Gwyn 2008 for details). In contrast with Carroll et al. (2014), where Fleming's function (Fleming et al., 1995) was used to account for the incompleteness near the 5σ limit, we use an erfc function. When multiplied by power-law number counts, Fleming's function causes completeness to drop to 20% of its peak at $r\sim30$ before rising again, while the erfc incompleteness function correctly damps down for higher magnitudes. We calculate the number of galaxies, $\mathrm{N_{gal}}$, in each HEALPix pixel in a given redshift bin as

$$\mathrm{N_{gal}} = 0.5 \int_{-\infty}^{\mathrm{m_{max}}} \mathrm{erfc}[a(\mathrm{m} - 5\sigma_{\mathrm{stack}})] 10^{\mathrm{c_1 m + c_2}} \mathrm{dm} \tag{2.2}$$

where a is the rollover speed and is chosen to be 1, $5\sigma_{\mathrm{stack}}$ is the coadded magnitude depth in the given HEALPix pixel, $\mathrm{m_{max}}$ is the magnitude cut, and $\mathrm{c_1}$ and $\mathrm{c_2}$ are the power-law constants determined from the mock catalogs for specific redshift bins. Here, we assume galaxies to have average colors, i.e. $u-g = g-r = r-i = 0.4$, and take this into account by modifying $\mathrm{c_2}$ and $\mathrm{m_{max}}$ in equation 2.2 for u, g, i vs. r. Given the sharp decline of the erfc function at

high magnitudes and the consequent decline in the differential galaxy counts, we consider a magnitude limit of r= 32.0 as no magnitude limit.

Using the number of galaxies in each pixel, we calculate the fluctuations in the galaxy counts $\Delta N/\overline{N}$ as $(N_{gal}/N_{avg}) - 1$, where N_{avg} is the average number of galaxies per pixel across the survey area. Within MAF, this procedure amounts to using a metric to calculate the number of galaxies and then post-processing the galaxy counts to find $\Delta N/\overline{N}$.

Here we note that artificial fluctuations in galaxy counts induced by the observing strategy (OS) scale the fluctuations arising due to actual LSS. In our calculations, we assume that LSS affects the local normalization of the galaxy luminosity function in a given redshift bin, not its shape. This assumption is valid as long as LSS does not alter the shape of the faint end of the luminosity function, which dominates the galaxy number counts. More precisely, in the ith pixel,

$$\left(\frac{N_{gal}}{N_{avg}}\right)_{observed,i} = \left(\frac{N_{gal}}{N_{avg}}\right)_{OS,i} \left(\frac{N_{gal}}{N_{avg}}\right)_{LSS,i} \tag{2.3}$$

Defining $\delta_i = \Delta N_i/\overline{N} = (N_{gal,i}/N_{avg}) - 1$, we have

$$(1 + \delta_{observed,i}) = (1 + \delta_{OS,i})(1 + \delta_{LSS,i}) \tag{2.4}$$

Since the ensemble average of LSS is zero, we have

$$\begin{aligned} \langle \delta_{observed,i} \rangle &= \langle \delta_{OS,i} \rangle + \langle \delta_{LSS,i} \rangle + \langle \delta_{OS,i} \delta_{LSS,i} \rangle \\ &= \delta_{OS,i} + \langle \delta_{OS,i} \delta_{LSS,i} \rangle \end{aligned} \tag{2.5}$$

where the angular brackets $\langle .. \rangle$ indicate an ensemble average defined as an average over many realizations of the Universe with one LSST survey. Hence, we have $\langle \delta_{OS,i} \rangle = \delta_{OS,i}$, as the OS-induced structure represents a fixed pattern on the sky for a given LSST observing strategy and OpSim run. Since there is generally no correlation between the OS–induced structure and LSS, the cross-term $\langle \delta_{OS,i} \delta_{LSS,i} \rangle$ should be negligible; we check and confirm this for a typical dither pattern. Also, we note that this assumption about the correlation between OS-induced structure and LSS breaks down if the survey strategy is correlated with LSS, e.g., Deep Drilling Fields focused on galaxy clusters, as then $\langle \delta_{OS,i} \delta_{LSS,i} \rangle \neq 0$.

Using equations 2.4-2.5, we calculate the power in $\delta_{observed,i}$:

$$\begin{aligned} \left\langle \delta^2_{observed,i} \right\rangle = \left\langle \delta^2_{OS,i} \right\rangle + \left\langle \delta^2_{LSS,i} \right\rangle + 2 \left\langle \delta_{OS,i} \delta_{LSS,i} \right\rangle \\ + 2 \left\langle \delta^2_{OS,i} \delta_{LSS,i} \right\rangle + 2 \left\langle \delta_{OS,i} \delta^2_{LSS,i} \right\rangle + \left\langle \delta^2_{OS,i} \delta^2_{LSS,i} \right\rangle \end{aligned} \tag{2.6}$$

As mentioned earlier, $\langle \delta_{\mathrm{OS},i} \delta_{\mathrm{LSS},i} \rangle$ is negligible since OS–induced structure and LSS are generally not correlated. To check how the higher order terms like $\langle \delta^2_{\mathrm{OS},i} \delta^2_{\mathrm{LSS},i} \rangle$ compare with $\langle \delta_{\mathrm{OS},i} \delta_{\mathrm{LSS},i} \rangle$, we calculate the cross-spectra for a typical dither pattern. We find that $\langle \delta_{\mathrm{OS},i} \delta_{\mathrm{LSS},i} \rangle$ is dominant over $\langle \delta^2_{\mathrm{OS},i} \delta^2_{\mathrm{LSS},i} \rangle$ and therefore these higher order terms are also negligible. Therefore,

$$\langle \delta^2_{\mathrm{observed},i} \rangle \approx \langle \delta^2_{\mathrm{OS},i} \rangle + \langle \delta^2_{\mathrm{LSS},i} \rangle = \delta^2_{\mathrm{OS},i} + \langle \delta^2_{\mathrm{LSS},i} \rangle \tag{2.7}$$

implying that the OS and LSS contribute independently to the observed power. $\delta^2_{\mathrm{OS},i}$ thus represents a bias in our measurement of LSS.

To consider realistic behavior of the observing strategies, we account for the uncertainties arising from photometric calibrations. Given that related systematic errors correlate with seeing (Leistedt et al., 2016) and are expected to decrease with the number of observations, we model the calibration uncertainty Δ_i in the ith HEALPix pixel as

$$\Delta_i = \frac{k \Delta \mathrm{s}_i}{\sqrt{\mathrm{N}_{\mathrm{obs},i}}} \tag{2.8}$$

where $\Delta \mathrm{s}_i$ is the difference between the average seeing in the ith HEALPix pixel and the average seeing across the map, $\mathrm{N}_{\mathrm{obs},i}$ is the number of observations in the ith pixel, and k is a constant such that the variance $\sigma^2_{\Delta_i} = 0.01^2$, ensuring the expected 1% errors in photometric calibration (LSST Science Collaboration et al., 2009). Figure 2.6 shows skymaps for these simulated uncertainties for example dither strategies. We note that while dithering does not alter the amplitudes of the photometric calibration uncertainties in our model, it helps mitigate the sharp hexagonal pattern seen in the uncertainties in the undithered survey.

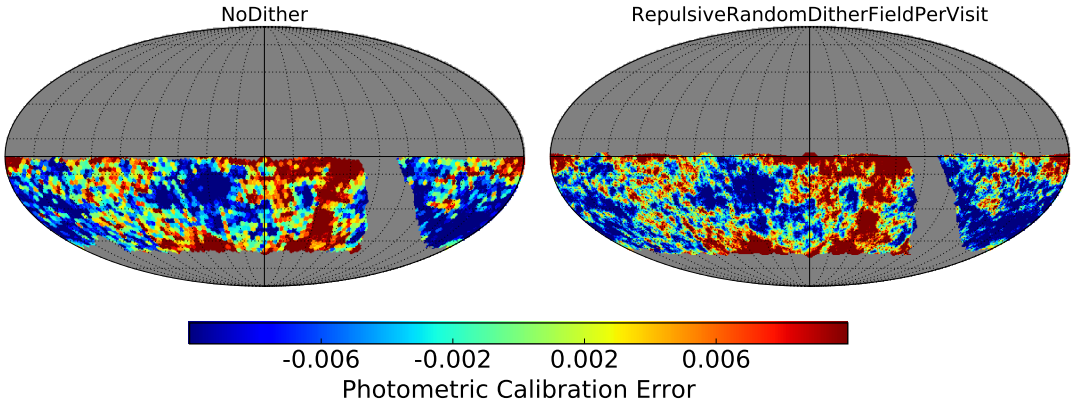

Figure 2.6: Skymaps of simulated photometric calibration uncertainties for example dither strategies.

In order to account for the fluctuations in the galaxy counts arising due to the photometric calibration uncertainties, we modify the upper limit on the magnitude in equation 2.2 to be $\mathrm{m}_{\mathrm{max}} + \Delta_i$ for the ith pixel. Since the calibration uncertainties are small, the skymaps for the fluctuations in the galaxy counts after accounting for the calibration uncertainties are indistinguishable from those without. These are shown in the top row in Figure 2.7.

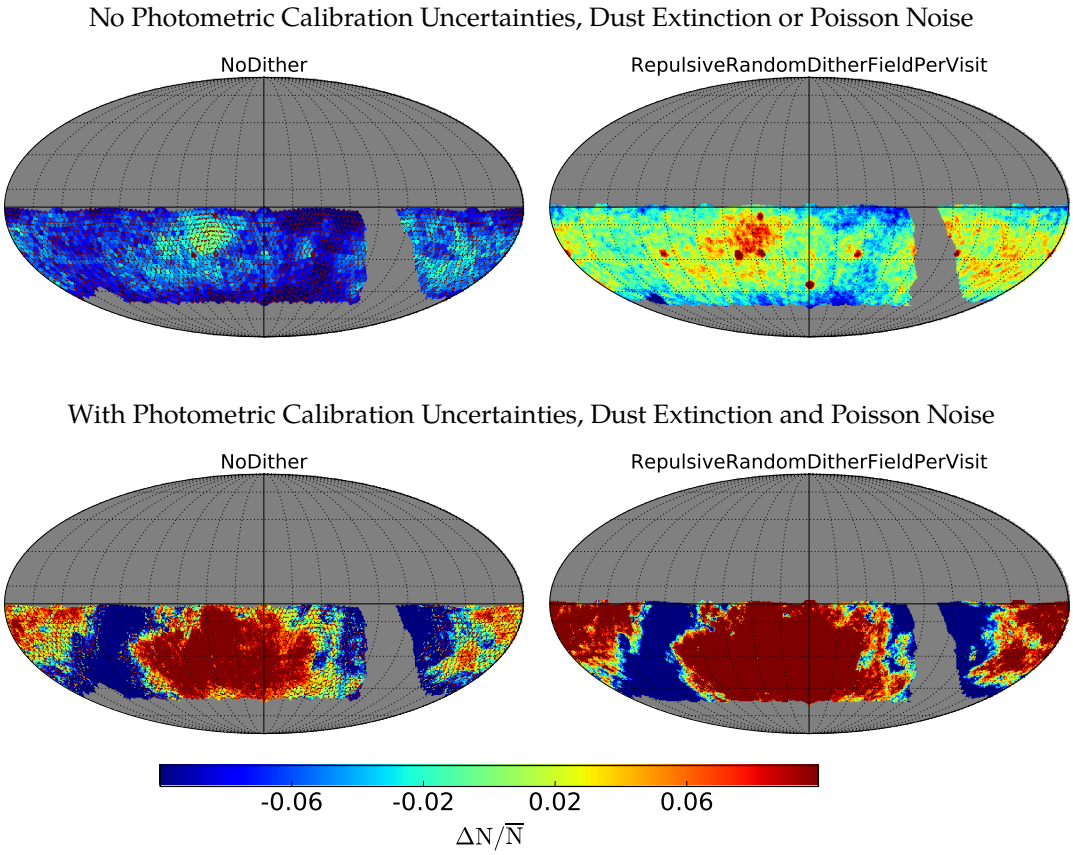

Figure 2.7: Skymaps for artificial galaxy fluctuations for example dither strategies for $0.66<z<1.0$. *Top row*: without calibration errors, dust extinction or poisson noise. *Bottom row*: After including calibration uncertainties, dust extinction and poisson noise. We do not see significant differences in the fluctuations after including the photometric calibration uncertainties or poisson noise; the skymaps match those in the top row. However, we see that dust extinction dominates the structure on large angular scales. These trends remain consistent across all five z-bins.

Furthermore, we include dust extinction by using the Schlegel-Finkbeiner-Davis dust map (Schlegel et al., 1998) when calculating the coadded depth as well as poisson noise in the galaxy counts after accounting for both dust extinction and photometric calibration. The bottom row in Figure 2.7 shows the skymaps for the artificial fluctuations for $0.66<z<1.0$ after accounting for photometric calibration uncertainties, dust extinction and the poisson noise. We find that dust extinction dominates both photometric calibration uncertainties and poisson noise; it induces power on large angular scales, but it does not wash out the honeycomb pattern in

the undithered survey or its low-level residual in the dithered surveys. These trends remain consistent across the five redshift bins.

Finally, in order to account for the spurious power introduced by the depth variations, we consider the relationship between the measured power spectrum and the true one, for a perfectly uniform survey:

$$\langle \mathrm{P}_{\mathrm{measured}}(\mathbf{k}) \rangle = \int \mathrm{d}\mathbf{k}' \, \mathrm{P}_{\mathrm{true}}(\mathrm{k}') \, |\mathrm{W}(\mathbf{k} - \mathbf{k}')|^2 \tag{2.9}$$

where $\mathrm{W}(\mathbf{k} - \mathbf{k}')$ is the survey window function, accounting for the effective survey geometry (Feldman et al., 1994; Sato et al., 2013). Projecting the 3D $\mathbf{k}$-space onto the 2D ℓ-space, we have

$$\mathrm{C}_{\ell,\mathrm{measured}} = \sum_{\ell'} |\mathrm{W}_{\ell-\ell'}|^2 \, \langle \mathrm{C}_{\ell'} \rangle + \delta \mathrm{C}_{\ell} \tag{2.10}$$

where $\langle \mathrm{C}_\ell \rangle$ is the expected power spectrum on the full sky, and $\delta \mathrm{C}_\ell$ is an error term whose minimum variance is given by (see Dodelson, 2003, Chapter 8 for details)

$$(\Delta \mathrm{C}_\ell)^2 = \frac{2}{\mathrm{f}_{\mathrm{sky}}(2\ell + 1)} \langle \mathrm{C}_\ell \rangle^2 \tag{2.11}$$

where $\mathrm{f}_{\mathrm{sky}}$ is the fraction of the sky observed, accounting for the reduction in observed power due to incomplete sky coverage. Since we consider only the WFD survey with masked shallow borders, $\mathrm{f}_{\mathrm{sky}} \approx 37\% - 39\%$ for the dithered surveys while $\mathrm{f}_{\mathrm{sky}} \approx 36\%$ for the undithered survey. The expected power spectrum can be defined as

$$\langle \mathrm{C}_\ell \rangle = \mathrm{C}_{\ell,\mathrm{LSS}} + \frac{1}{\bar{\eta}} \tag{2.12}$$

where $\bar{\eta}$ is the surface number density in steradians^{-1}; see Fall (1978), Huterer et al. (2001), Jing (2005) for details. The first term in equation 2.12 is the LSS contribution to the expected power spectrum while the second is the shot noise contribution arising from discrete signal sampling.

With no LSS and negligible shot noise, $\langle \mathrm{C}_\ell \rangle \to 0$. However, as shown in equation 2.7, the observing strategy induces a bias in the measured power spectrum, leading to non-zero power even when $\langle \mathrm{C}_\ell \rangle \to 0$. The uncertainty in this bias caused by imperfect knowledge of the survey performance limits our ability to correct for the OS-induced artificial structure. More quantitatively, we have

$$\left(\sigma_{\mathrm{C}_{\ell,\mathrm{measured}}}\right)^2 = (\Delta \mathrm{C}_\ell)^2 + \left(\sigma_{\mathrm{C}_{\ell,\mathrm{OS}}}\right)^2 \tag{2.13}$$

where the first term on the right is the minimum statistical uncertainty defined in equation 2.15, while the second term corresponds to the contribution from the uncertainty in the bias induced by the OS. Since the "statistical floor" ΔC_ℓ assumes no bias in C_ℓ measurements caused by the observing strategy, the OS-induced uncertainty $\sigma_{C_{\ell,\rm OS}}$ must be subdominant to the statistical floor for an optimal measurement of BAO at a given redshift, i.e.

$$\sigma_{C_{\ell,\rm OS}} << \Delta C_\ell = \sqrt{\frac{2}{f_{\rm sky}(2\ell+1)}}\left(C_{\ell,\rm LSS} + \frac{1}{\bar{\eta}}\right) \tag{2.14}$$

Here we note that the right-hand side in equation 2.14 is formally derived in Shafer & Huterer (2015a); also see Huterer et al. (2013). These papers offer a detailed theoretical treatment of artificial structure induced by calibration errors, and while our approach is similar to theirs, we incorporate the additional effects of dust extinction, variations in survey depth, and incompleteness in galaxy detection.

Considering the case where $C_{\ell,\rm LSS} = 0$, we find $C_{\ell,\rm measured}$, giving us $C_{\ell,\rm OS}$ for each band and magnitude cut. Since $ugri$ bands are the deepest and appear to have the greatest influence on photometric redshifts (Prakash, priv. comm.), we model the overall bias as the mean $C_{\ell,\rm OS}$ across the four bands. We calculate $\sigma_{C_{\ell,\rm OS}}$ as the standard deviation of $C_{\ell,\rm OS}$ across the $ugri$ bands, modeling uncertainties due to detecting galaxy catalogs in different bands. Therefore, $\sigma_{C_{\ell,\rm OS}}$ should provide a conservative upper limit on the true uncertainty in $C_{\ell,\rm OS}$.

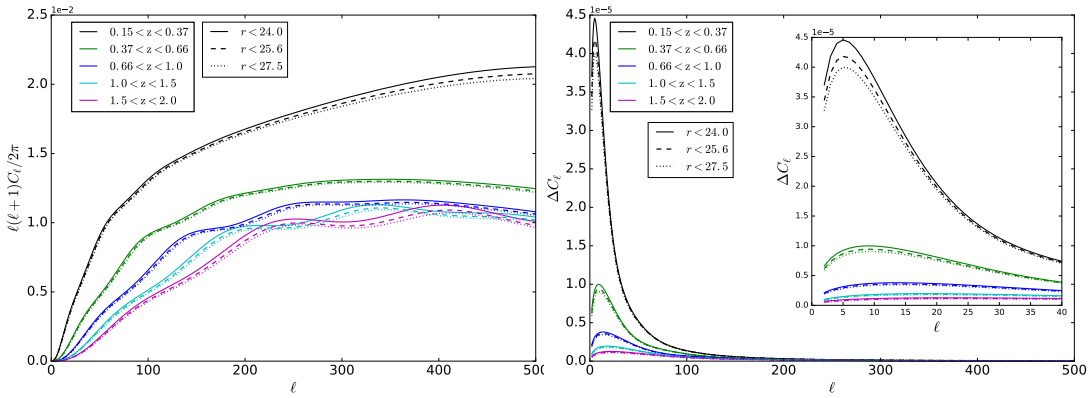

Figure 2.8: *Left*: Simulated full-sky, pixelized galaxy power spectra with BAO signal from five different redshift bins for three galaxy catalogs: r<24.0, 25.6, 27.5. *Right*: Minimum statistical error associated with measuring the signal in the left panel, with lower–ℓ range shown in the inset. We observe that neither curve changes significantly with magnitude cuts considered in Section 2.4.2.

The left panel in Figure 2.8 shows the full-sky galaxy power spectrum with the BAO signal for three galaxy catalogs, r<24.0, 25.6, 27.5, for the five redshift bins: 0.15<z<0.37, 0.37<z<0.66, 0.66<z<1.0, 1.0<z<1.5, and 1.5<z<2.0. These spectra are pixelized in order to account for the

finite angular resolution of our survey simulations, especially when comparing the uncertainties in $C_{\ell,\rm OS}$ with the minimum statistical error in measuring BAO. Assuming that all the HEALPix pixels are identical, the pixelized power spectra can be approximated by multiplying the galaxy power spectra with the pixel window function[4].

The galaxy power spectra are calculated using the code from Zhan (2006), with modifications to account for BAO signal damping due to non-linear evolution (Eisenstein et al., 2007). Using the galaxy redshift distribution from LSST Science Collaboration et al. (2009), galaxies are assigned to the five redshift bins according to their photometric redshifts, with a time-varying but scale-independent galaxy bias of $b(z) = 1 + 0.84z$ over scales of interest and a simple photometric redshift error model, $\sigma_z = 0.05(1 + z)$. Here we assume the cosmology with $w_0 = -1$, $w_a = 0$, $\Omega_m h^2 = 0.127$, $\Omega_b h^2 = 0.0223$, $\Omega_k = 0$, spectral index of the primordial scalar perturbation power spectrum $n_s = 0.951$ and primordial curvature power spectrum at k = 0.05/Mpc, $\Delta^2_R = 2 \times 10^{-9}$.

The right panel in Figure 2.8 shows the minimum statistical uncertainty for the five redshift bins for all three galaxy catalogs; the uncertainties are calculated using $f_{\rm sky}$ from the undithered survey. We observe that while shallower galaxy catalogs lead to larger C_ℓ and ΔC_ℓ, the difference is small and decreases with increasing redshift. For the lowest z-bin, 0.15<z<0.37, there is only about 8% increase in C_ℓ and ΔC_ℓ when comparing the r<25.6 catalog with r<24.0.

First we calculate $C_{\ell,\rm OS}$ and its uncertainties for 0.66<z<1.0 after only one year of survey in order to explore the quality of BAO study the first data release will allow. Figure 2.9 shows the $C_{\ell,\rm OS}$ uncertainties as well as the minimum statistical error for 0.66<z<1.0 for various observing strategies, for r<24.0 and r<25.7 (corresponding to the gold sample, i<25.3). We observe that the undithered survey leads to $C_{\ell,\rm OS}$ uncertainties 1-3× the minimum statistical uncertainty for the gold sample at ℓ >100, and only a few dither strategies are effective in reducing the difference. In particular, Random and RepulsiveRandom dithers are the most effective, reducing $\sigma_{C_{\ell,\rm OS}}$ to nearly 1-2× the statistical floor. We note that FermatSpiral and SequentialHex dithers perform nearly as poorly as NoDither when implemented on FieldPerVisit and FieldPerNight timescales, while the PerNight timescale is more effective. On the other hand, we see that FieldPerVisit and FieldPerNight lead to smaller uncertainties for Random and RepulsiveRandom geometries. As expected, we see that a shallower sample r<24.0 reduces the $C_{\ell,\rm OS}$ uncertainties; the undithered survey still leads to $\sigma_{C_{\ell,\rm OS}}$ about 3× the statistical floor, while Random and RepulsiveRandom dithers lead the uncertainties comparable to the

[4]See Appendix B in the HEALPix primer: http://healpix.sourceforge.net/pdf/intro.pdf

statistical floor on some timescales. Here we note that since we do not mask borders when considering the one-year data, $f_{sky} \approx 42\% - 45\%$ for the one year survey, depending on the dither strategy.

We then extend the calculation of the OS-induced power to the full 10-year survey. Figure 2.10 shows $\sigma_{C_{\ell,OS}}$ as well as ΔC_ℓ for $0.66<z<1.0$ for three different magnitude cuts: $r<24.0$, $r<25.7$ and $r<27.5$. We find that the undithered survey leads to $\sigma_{C_{\ell,OS}}$ 0.2-4 times the minimum statistical floor for $r<25.7$ and $r<27.5$; at $\ell > 100$, only a very strict cut of $r<24.0$ brings $\sigma_{C_{\ell,OS}}$ below ΔC_ℓ. However, most dither strategies reduce the uncertainties below the statistical floor for galaxy catalogs as deep as $r<27.5$, with exceptions of SequentialHex dithers on FieldPerVisit and FieldPerNight timescales. We note here that systematics correction methods such as template subtraction and mode projection can be applied to further reduce the contribution of $C_{\ell,OS}$ to the total C_ℓ uncertainties; e.g., see Elsner et al. (2016), Holmes et al. (2012). Such application appears necessary for the 1-year survey as optimizing the observing strategy alone does not reduce the uncertainties in $C_{\ell,OS}$ below ΔC_ℓ. However, the correction methods may not lead to significant improvements for a dithered 10-year survey, as optimizing the observing strategy is effective in reducing $C_{\ell,OS}$ well below the statistical floor.

To further our understanding, we repeat the 1-year and 10-year analysis for $1.5<z<2.0$. We find similar qualitative results as those from $0.66<z<1.0$ analysis: for the 1-year survey, Random and RepulsiveRandom perform well alongside FermatSpiral and SequentialHex on PerNight timescale, while most dither strategies are effective for the ten-year survey, with the exception of SequentialHex on FieldPerVisit and FieldPerNight timescales.

The effect of magnitude cuts is further illustrated in Table 2.1, which includes the estimated number of galaxies for $0.15<z<2.0$ from the r-band coadded depth for the 10-year survey after accounting for photometric calibration errors, dust extinction and poisson noise. We see that each magnitude cut eliminates a substantial number of galaxies. Also, as in Carroll et al. (2014), we see that dithering increases the estimated number of galaxies when compared to the undithered survey; the fractional difference in the number of galaxies from dithered to undithered surveys increases with shallower surveys.

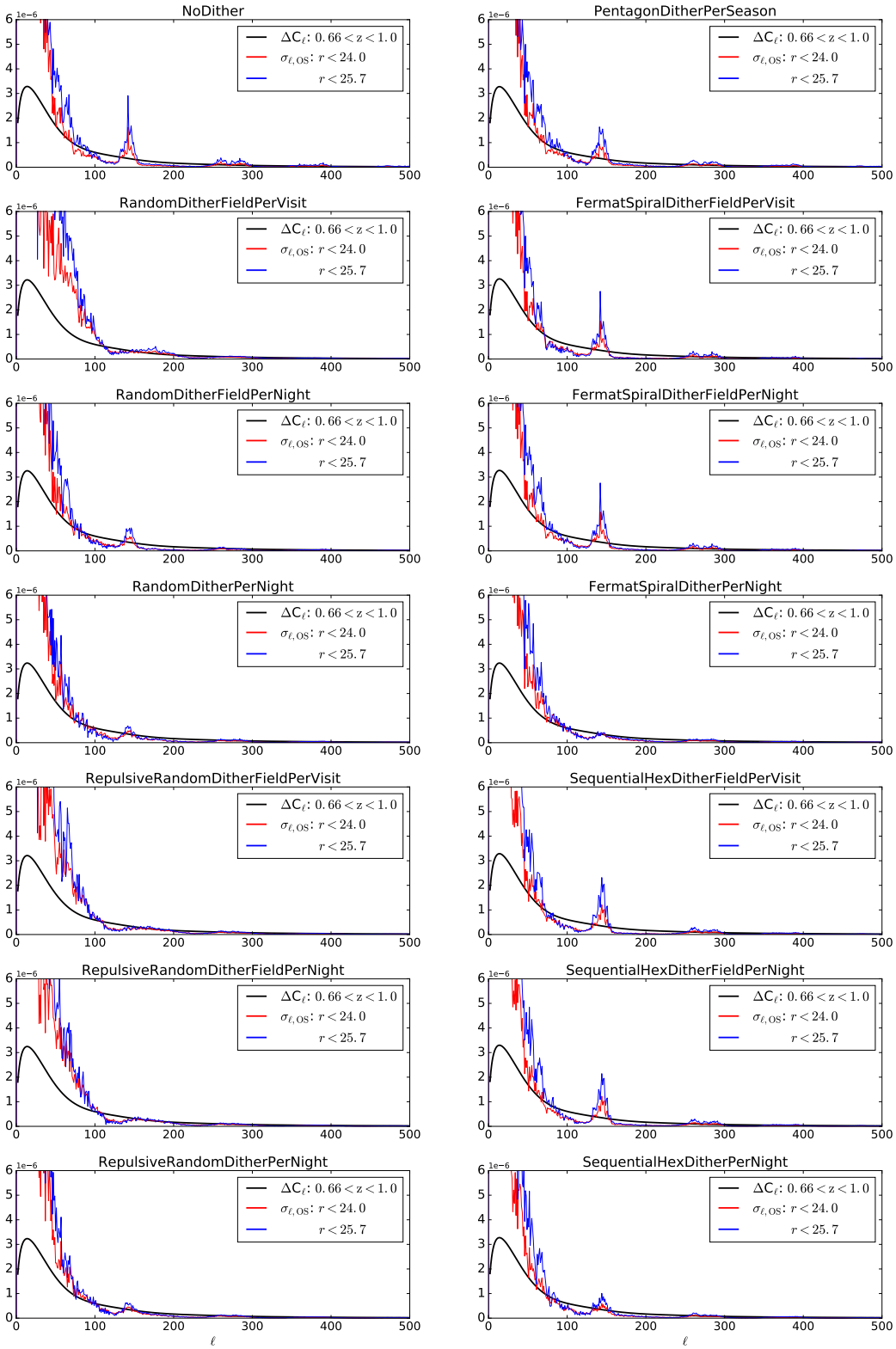

Figure 2.9: $\sigma_{\ell,\mathrm{OS}}$ comparison with the minimum statistical uncertainty ΔC_ℓ for $0.66<z<1.0$ for different magnitude cuts after only one year of survey.

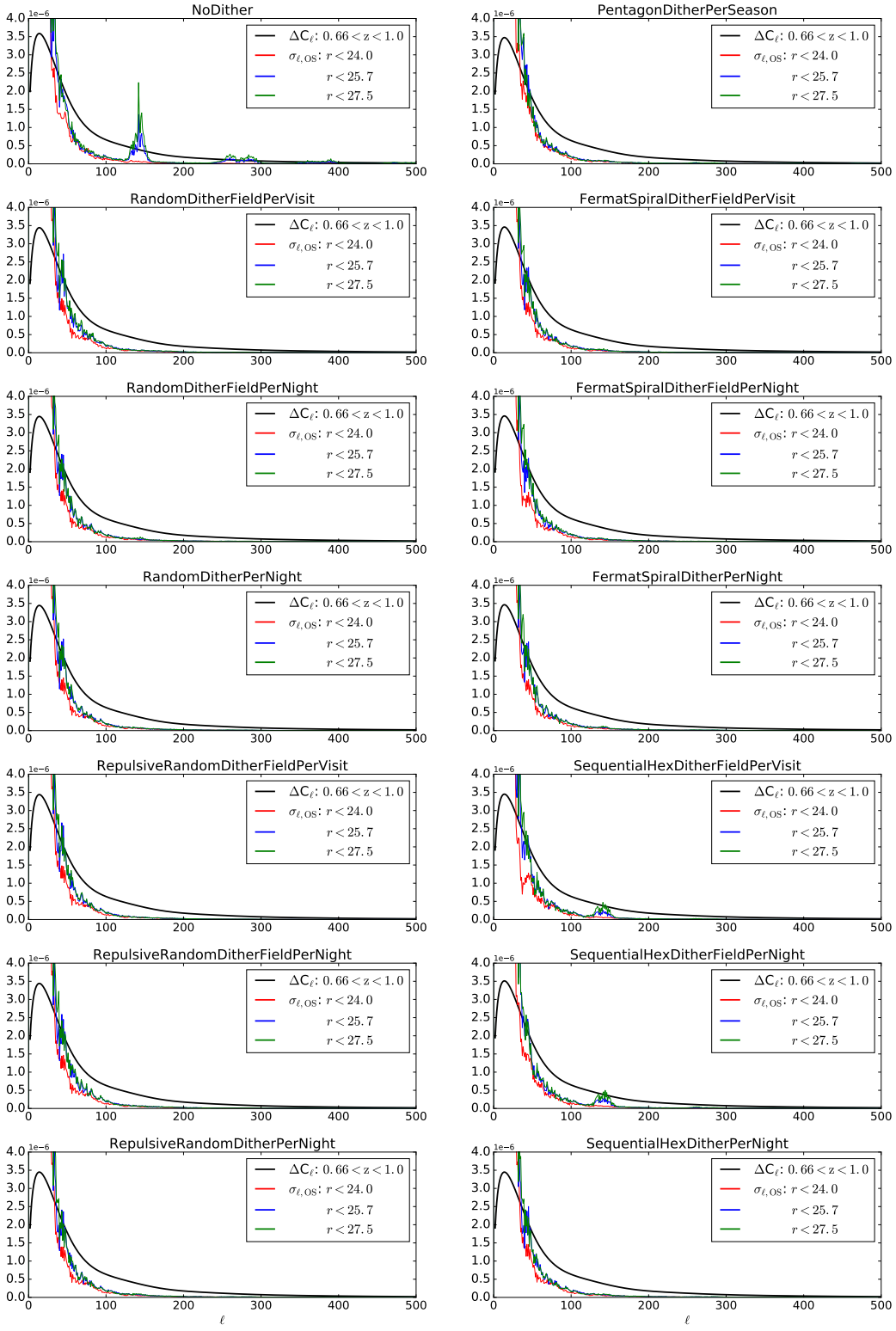

Figure 2.10: $\sigma_{\ell,\mathrm{OS}}$ comparison with the minimum statistical uncertainty ΔC_ℓ for $0.66<z<1.0$ for different magnitude cuts after the full 10-year survey.

Table 2.1. Estimated number of galaxies from r-band coadded depth after the 10-year survey for 0.15<z<2.0, after accounting for photometric calibration errors, dust extinction and poisson noise.

	r<27.5	r<25.7	r<24.0
Number of galaxies from NoDither	1.0×10^{10}	4.3×10^{9}	1.6×10^{9}
Percent improvements in comparison with NoDither			
PentagonDitherPerSeason	7.0	6.6	6.6
SequentialHexDitherFieldPerVisit	8.1	7.8	7.9
SequentialHexDitherFieldPerNight	4.9	4.3	4.4
SequentialHexDitherPerNight	8.3	8.0	8.1
FermatSpiralDitherFieldPerVisit	7.6	7.2	7.3
FermatSpiralDitherFieldPerNight	7.6	7.2	7.3
FermatSpiralDitherPerNight	7.4	7.0	7.1
RandomDitherFieldPerVisit	8.7	8.4	8.5
RandomDitherFieldPerNight	8.3	8.0	8.1
RandomDitherPerNight	8.5	8.2	8.3
RepulsiveRandomDitherFieldPerVisit	8.9	8.5	8.7
RepulsiveRandomDitherFieldPerNight	8.6	8.4	8.5
RepulsiveRandomDitherPerNight	8.3	7.9	8.0

Note. — We observe 6.5-9% improvement in the estimated number of galaxies from dithered surveys in comparison with undithered survey, across the three magnitude cuts. The exception is SequentialHexDitherFieldPerNight where the improvement is only 4-5%.

2.5 Conclusions

It is critical to develop an LSST observing strategy that will maximize the data quality for its science goals. In this work, we analyzed the effects of dither strategies on r-band coadded 5σ depth to study the feasibility of increasing the uniformity across the survey region. We investigated different dither geometries on different timescales, and illustrated how a specific geometrical pattern (e.g., hexagonal lattice) can perform quite differently when implemented on different timescales. We find that per-visit and per-night implementations outperform field-per-night and per-season timescales, while some dither geometries (like repulsive random dithers) consistently lead to less spurious power for all the timescales on which the dither positions are assigned. We also performed an $a_{\ell m}$ analysis to probe the origins of some of the characteristic patterns induced by the observing strategies. Our work illustrates the sensitivity of depth uniformity to the dither strategy.

We then considered how the artifacts in coadded depth produce fluctuations in galaxy counts; we calculate the uncertainties in the bias induced by the observing strategy, which limits our ability to correct for the spurious structure. We find that after accounting for photometric calibration uncertainties, dust extinction, poisson noise and reasonable magnitude cuts, dithers

of most kinds are effective in reducing the uncertainties in the observing-strategy-induced bias below the minimum statistical uncertainty in the measured galaxy power spectrum. Specifically, we find that RepulsiveRandom dithers implemented on per-visit and field-per-night timescales are the most effective for the $0.66<z<1.0$ sample after only one year of survey, although they do not bring down the uncertainties in the induced bias below the minimum statistical floor for $r<25.7$. As for the full 10-year survey, we find that all dither strategies (except per-visit and field-per-night SequentialHex dithers) bring down the uncertainties below the statistical floor for a galaxy catalog as deep as $r<27.5$. We find similar results for all redshift bins.

To precisely determine the limiting uncertainties in the bias induced by the observing strategy, more detailed LSST simulations are needed, including photometric redshifts, input large-scale structure and further systematics reduction methods, e.g., mode projection accounting for imperfect detectors and the consequent instrumental effects. Also, while our work illustrates the impact of dithers on large-scale structure studies, the differences between some dither geometries are small and therefore need more detailed investigation to determine a conclusively-best dither strategy, alongside an analysis of the impacts of various dither strategies on other science goals. Such analyses will facilitate a more definitive measure of the precision with which LSST data will allow high redshift studies of large-scale structure.

2.6 Acknowledgements

This research was supported by the Department of Energy (grant DE-SC0011636) and National Science Foundation (REU grant PHY-1263280). Hu Zhan was partially supported by the Chinese Academy of Sciences (grant XDB09000000), and Alejandra M. Muñoz Arancibia by FONDECYT (grant 3160776) and BASAL CATA PFB-06. Nelson D. Padilla also acknowledges support from BASAL CATA PFB-06 and FONDECYT (grant 1150300); the lightcones were run using the Geryon cluster hosted at the Centro de Astro-Ingenieria UC. Sofía A. Cora acknowledges support from Consejo Nacional de Investigaciones Científicas y Técnicas, Agencia Nacional de Promoción Científica y Tecnológica, and Universidad Nacional de La Plata, Argentina. We thank K. Simon Krughoff for suggesting the hexagonal dither pattern for LSST, and Abhishek Prakash, Jeff Newman, Andy Connolly, Rachel Mandelbaum, Dragan Huterer, Terry Matilsky and Seth Digel for helpful comments, conversations and insights. Finally, we thank the LSST Dark Energy Science Collaboration for feedback on design and conduct of this research.

Appendices

2.A Border Masking Algorithm

In Figure 2.A.1, we show skymaps (left column) and the corresponding power spectrum (right column) for the r-band coadded 5σ depth from the undithered survey and an example dithered survey. While the dithered survey does not have the strong honeycomb seen in the undithered case, we notice that the border of the dithered survey area is much shallower than the rest of the survey. This variation in depth carries over to the power spectrum as strong oscillations, especially at small ℓ. In order to minimize this effect, we develop a border masking algorithm to mask the pixels within a specific 'pixel radius' from the edge of the survey area. For this purpose, we utilize the distinction between out-of-survey and in-survey area in MAF: the former is masked, and the analysis only accounts for the data in the unmasked portion of the data array. Using this distinction and the HEALpix routine `get_all_neighbours`, we find the unmasked pixels with masked neighbors, effectively finding the edge of the survey. We parametrize the number of iterations for this neighbor finding algorithm, and choose the number of iterations (determined by what we call the pixel radius) that removes the shallow border. The masking algorithm can be found on GitHub[5].

Working at N_{side}= 256 resolution, we masked all the pixels within a 14-pixel radius from the edge of survey, effectively masking $\sim$15% of the survey area. The bottom row in Figure 2.A.1 shows the dithered skymap and the corresponding power spectrum after the shallow border has been removed. We notice a stark difference between the power spectrum before and after the border masking, as removing the shallow border allows the in-survey variations to be seen much more clearly.

[5] https://github.com/LSST-nonproject/sims_maf_contrib/blob/master/mafContrib/maskingAlgorithmGeneralized.py

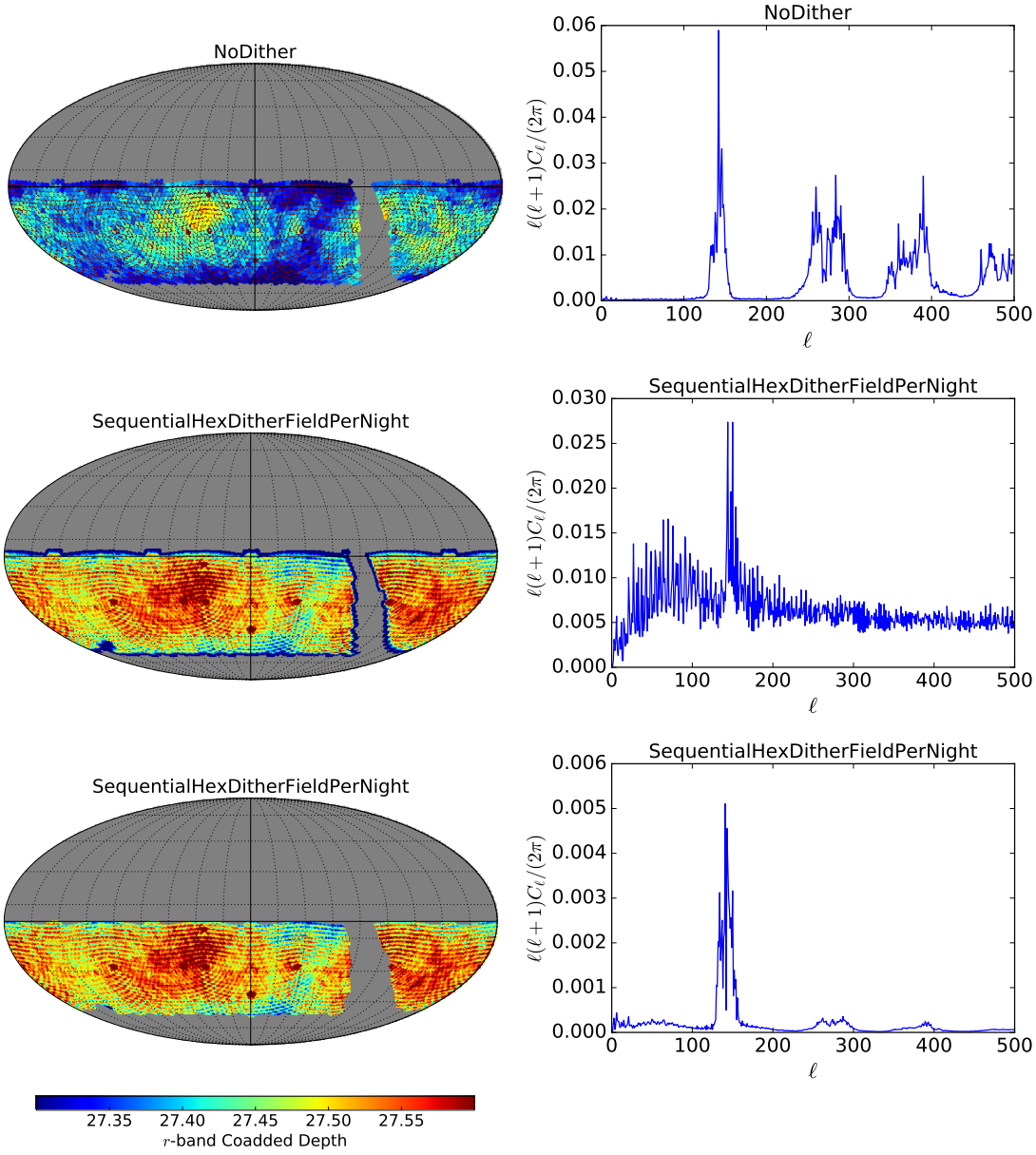

Figure 2.A.1: *Left column*: Skymaps for r-band coadded 5σ depth for example dither strategies. *Right column*: Angular power spectra corresponding to the skymaps in the first column. Top and middle rows show the data without any border masking. We note that the undithered survey does not lead to any shallow edges, while dithered survey does. The shallow-depth edge leads to a noisy power spectrum, shown in the middle right panel. After removing the shallow-border by implementing 14 pixel-radius masking, we see a reduction in the low-ℓ power, and therefore a cleaner spectrum.

2.B Dithering to Improve Survey Uniformity

Following our work in Awan et al. (2016), we extend our analysis to not only test the latest simulations but also develop a Figure of Metric for LSS studies to quantify the impacts of systematics induced by the observing strategy on the specific probe[6]. The latest simulation outputs include `minion_1016` which is the new baseline (as opposed to `enigma_1189` used in the analysis earlier), `minion_1020` which covers the maximum survey area of 27,400 deg^2, and `kraken_1043` which is the same as `minion_1016` but relaxes the constraint of visit pairs.

2.B.1 Figure of Merit

As derived and discussed in Awan et al. (2016), the spurious power from the artificial fluctuations in the galaxy counts induced by the observing strategy (OS) represents a bias in our measurement of the LSS. Hence, the uncertainty in this bias becomes the limiting factor in our ability to correct for the structure induced by the observing strategy. More quantitatively, for an optimized LSS study, the uncertainties induced by the observing strategy, $\sigma_{C_{\ell,\mathrm{OS}}}$, must be subdominant to the statistical uncertainty ΔC_ℓ inherent to the measured power spectrum due to "cosmic variance" (Dodelson, 2003):

$$\Delta C_\ell = C_{\ell,\mathrm{LSS}} \sqrt{\frac{2}{f_{\mathrm{sky}}(2\ell+1)}} \tag{2.15}$$

where f_{sky} is the fraction of the sky observed, accounting for the reduction in the observed information due to incomplete sky coverage.

Since we do not include any input LSS in our pipeline, the power spectrum we measure for any given band is due to the power induced by the observing strategy, $C_{\ell,\mathrm{OS}}$, for that band. Modeling the overall bias induced by the observing strategy as an average across $ugri$ bands, we calculate the uncertainties in the bias $\sigma_{C_{\ell,\mathrm{OS}}}$ as the standard deviation across $C_{\ell,\mathrm{OS}}$ for $ugri$ bands to account for the effects of detecting the galaxy catalog through various bands. We then compare these uncertainties with the statistical floor for various redshift bins, where the statistical floor is based on the galaxy power spectra calculated using the code from Zhan (2006), which we pixelize to match the HEALPix resolution to account for the finite angular resolution

[6]Note that we do not explore a direct relation between our Figure of Merit (FoM) and cosmological parameters but focus only on developing a metric to quantify the impacts of observing strategy to measured power spectra. Our FoM does correlate nontrivially with the Dark Energy Task Force FoM, which is defined as the reciprocal of the area of the contour enclosing the 68% confidence interval constraining the dark energy parameters, w_0 and w_a, after marginalizing over other parameters (Albrecht et al., 2006). Given the lack of a direct connection with cosmological parameters, however, our FoM may be thought of as a diagnostic metric as opposed to a true figure of merit.

of our simulations.

To quantify the effectiveness of each observing strategy in minimizing $\sigma_{C_{\ell,\mathrm{OS}}}$, we construct a Figure of Merit (FoM) as the ratio of the ideal-case uncertainty in the measured power spectrum and the uncertainty arising from shot noise and the structure induced by the observing strategy:

$$\mathrm{FoM} = \sqrt{\frac{\sum_{\ell}\left(\sqrt{\frac{2}{f_{\mathrm{sky,max}}(2\ell+1)}}C_{\ell,\mathrm{LSS}}\right)^2}{\sum_{\ell}\left[\left(\sqrt{\frac{2}{f_{\mathrm{sky}}(2\ell+1)}}\left\{C_{\ell,\mathrm{LSS}}+\frac{1}{\bar{\eta}}\right\}\right)^2+\sigma^2_{C_{\ell,\mathrm{OS}}}\right]}} \tag{2.16}$$

Here, $\bar{\eta}$ is the surface number density in steradians^{-1}, and the term containing it accounts for the contribution from the shot noise to the measured signal (Huterer et al., 2001; Jing, 2005). This FoM measures the percentage of ideal-case information that can be measured in the presence of systematics. We note that the shot noise is negligible even for the shallowest (10-year) surveys we consider.

We define the ideal-case as being based on the largest coverage of the sky with LSST, i.e., $f_{\mathrm{sky,max}}$ is the largest WFD coverage with the baseline cadence. For `minion_1016`, the observing strategy with RepulsiveRandomDitherFieldPerVisit dithers leads to the largest f_{sky} ($\sim 39.5\%$). Note that this fraction is calculated after masking the shallow borders of the main survey; for details, see Awan et al. (2016).

2.B.2 A Comment on Terminology

For clarity, we make a note on the terminology we have introduced. Strictly speaking, the bias caused by the observing strategy is a window function bias, as the survey window function (W_i) accounts for the effective survey geometry which scales the fluctuations in the galaxy counts in each pixel: $1+\delta_{\mathrm{obs},i} = W_i(1+\delta_{\mathrm{LSS},i})$. Comparing this with Equation 4 in Awan et al. (2016), $1+\delta_{\mathrm{obs},i} = (1+\delta_{\mathrm{OS},i})(1+\delta_{\mathrm{LSS},i})$, we see that the bias induced by the observing strategy is directly related to the window function: $1+\delta_{\mathrm{OS},i} = W_i$

Then, for the total power, we have

$$\left\langle\delta^2_{\mathrm{obs},i}\right\rangle = \left\langle\delta^2_{\mathrm{LSS},i}\right\rangle\left\langle(1+\delta_{\mathrm{OS},i})^2\right\rangle + \left\langle\delta^2_{\mathrm{OS},i}\right\rangle = \left\langle\delta^2_{\mathrm{LSS},i}\right\rangle\left\langle W_i^2\right\rangle + \left\langle(W_i-1)^2\right\rangle \tag{2.17}$$

where the first equality is based on Equation 6 in Awan et al. (2016) and the second one holds given the relation between $\delta_{\mathrm{OS},i}$ and W_i. Since the bias induced by the observing strategy $\delta^2_{\mathrm{OS},i} = (W_i-1)^2$, the uncertainties in the bias are the window function uncertainties.

Generally the window function is assumed to be known perfectly and its uncertainties

are not explicitly identified as such. To avoid confusion and focus on the window function uncertainties arising from the observing strategy, we continue using the terms bias induced by the observing strategy and its uncertainties in favor of window function and its uncertainties.

2.B.3 OpSim Analysis and Results

For the purposes of our analysis, we use HEALPix resolution of $N_{\rm side} = 256$, effectively tiling each 3.5° FOV with about 190 HEALPix pixels. Using the metrics developed in the text, we analyze $\sigma_{C_{\ell,\rm OS}}$ from various observing strategies. First we present the results for the baseline cadence, `minion_1016`.

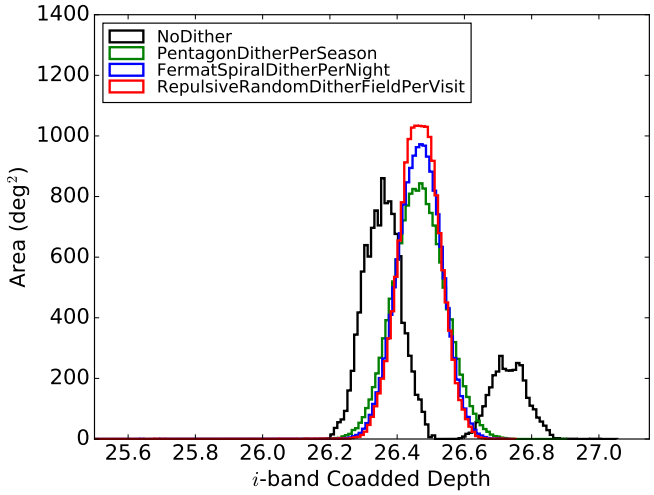

Figure 2.B.1: Histogram for the i-band coadded 5σ depth after the full, 10-year survey.

Figure 2.B.1 shows the histogram for the i-band coadded 5σ depth from `minion_1016` for the four observing strategies. We observe a bimodal distribution for the undithered survey – the deeper depth mode corresponds to the overlapping regions between the hexagons, while the rest of the survey contributes to the shallower mode. In contrast, all dithered surveys lead to unimodal distributions as the overlapping regions between the fields change frequently, leading to more uniformity. We also note that frequent dithering leads to deeper regions as we observe more peaked histograms for FieldPerVisit and PerNight strategies.

Figure 2.B.2 shows the plots for the i-band coadded 5σ depth for the observing strategies. As in Awan et al. (2016), we find that the undithered survey leads to a strong honeycomb pattern which is much weaker in all of the dithered surveys. We again observe that the dithered surveys are deeper than the undithered survey in terms of the median depth across the survey region.

In order to quantify the angular characteristics observed in the skymaps, we calculate

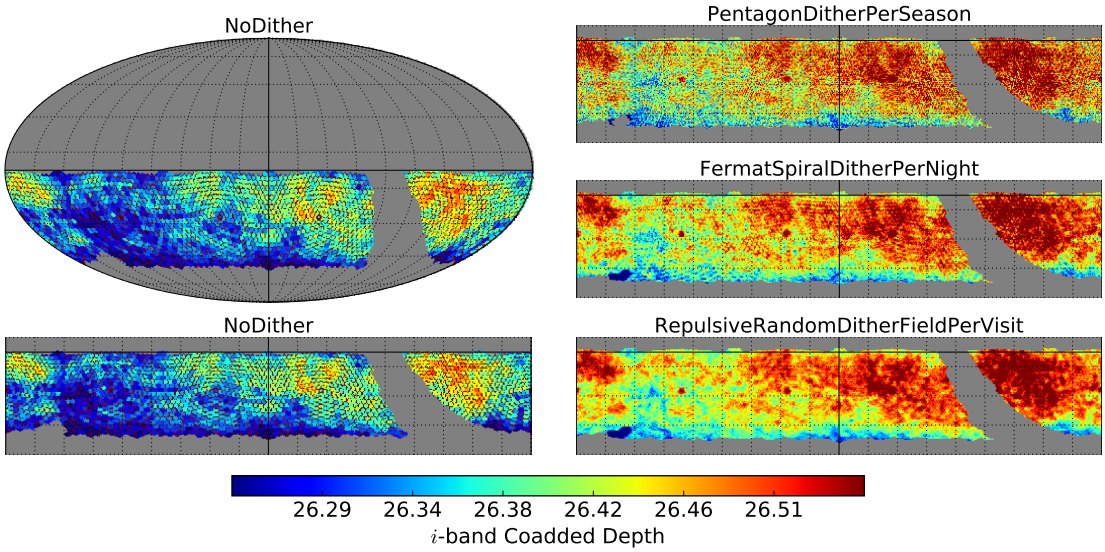

Figure 2.B.2: Plots for the i-band coadded 5σ depth based on `minion_1016` for various observing strategies. The top left plot shows the Mollweide projection for NoDither while the bottom left shows the corresponding Cartesian projection, restricted to 180° >RA> -180° (left-right), -70° <Dec< 10° (bottom-top). Only the latter is shown for the rest of the strategies.

the angular power spectra corresponding to the skymaps for the i-band coadded 5σ depth. Figure 2.B.3 shows these spectra for the four observing strategies. We observe a sharp reduction in the artificial power in the dithered surveys when compared to the undithered one: the strong honeycomb pattern in the undithered survey leads to a large peak around $\ell \sim 150$, while the peak is about 10 times weaker in the dithered surveys. We do, however, observe variations amongst the various dither strategies: while RepulsiveRandom dithers lead to small power for all timescales, PerSeason dithers lead to large power on larger angular scales, and both PerSeason and FermatSpiral lead to large power around $\ell \sim 150$ (which still is $< 10\times$ the corresponding peak from the undithered survey).

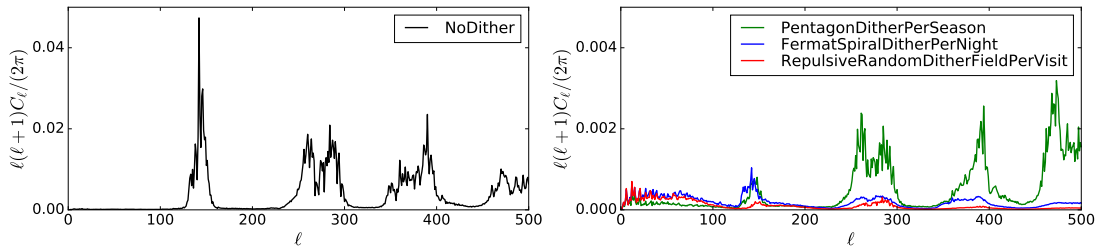

Figure 2.B.3: Angular power spectra for the i-band coadded 5σ depth from `minion_1016` for various observing strategies. We note that dithering reduces the spurious power by over $10\times$.

We then proceed to calculate the bias induced by the observing strategy and its uncertainty from the different observing strategies. First, we examine simulated results after only one year of survey. Figure 2.B.4 shows the comparison between $\sigma_{C_{\ell,\mathrm{OS}}}$ and ΔC_ℓ for $0.66 < z < 1.0$

after the 1-year survey for two magnitude cuts: $i < 24.0$ and $i < 25.3$. We observe that the undithered survey leads to $\sigma_{C_{\ell,\mathrm{OS}}}$ 1-5$\times$ the statistical floor around $\ell \sim 150$; PerSeason timescale does only slightly better. However, we see an improvement with frequent dithers: both FieldPerVisit and PerNight implementations lead to uncertainties 0.5-1$\times$ the statistical floor, although FermatSpiral dithers on PerNight timescale lead to a peak around $\ell \sim 150$ more pronounced than the one from RepulsiveRandom dithers on FieldPerVisit timescale.

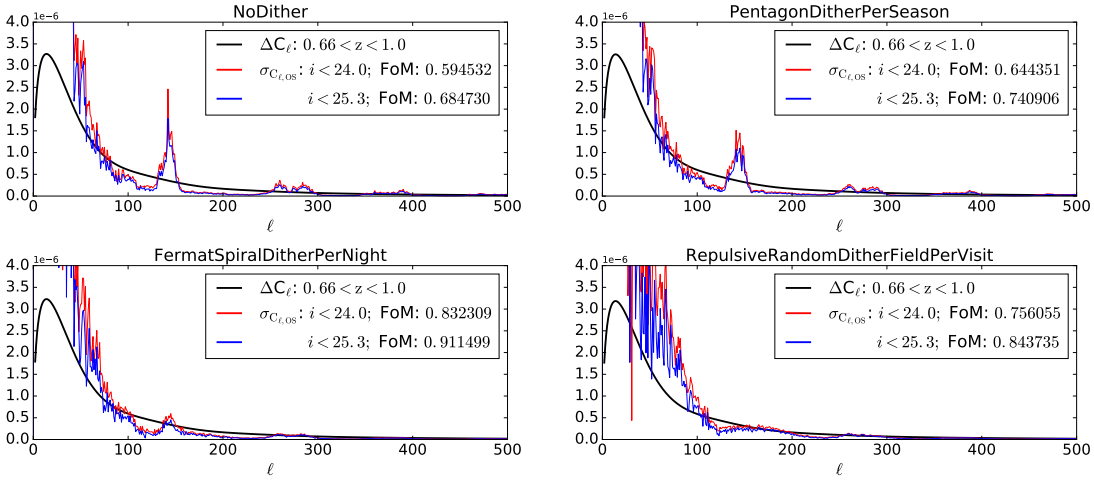

Figure 2.B.4: $\sigma_{C_{\ell,\mathrm{OS}}}$ comparison with the minimum statistical uncertainty ΔC_ℓ for $0.66 < z < 1.0$ for different magnitude cuts after only one year of survey based on `minion_1016`.

The trends are captured in the Figure of Merit, which we calculate using Equation 2.16 over the range $100 < \ell < 300$. We observe a smaller FoM for the shallower survey – realistic given that although there is less structure and therefore weaker artifacts induced by the observing strategy, the shot noise becomes significant and makes the FoM smaller. For the deeper survey, we find that FermatSpiralDitherPerNight outperforms all others with the highest FoM, while RepulsiveRandomDitherFieldPerVisit is more effective than PerSeason dithers. The undithered survey, as expected, performs the worst.

In Figure 2.B.5, we show simulated results after the full, 10-year survey for $0.66 < z < 1.0$ for three different magnitude cuts: $i < 24.0$, $i < 25.3$ and $i < 27.5$. We observe stark differences between the undithered and dithered surveys: the former leads to large uncertainties in the bias induced by the observing strategy while the latter is effective in bringing $\sigma_{C_{\ell,\mathrm{OS}}}$ well below the statistical floor. The effectiveness of all three dithered surveys in minimizing the uncertainties implies more flexibility in choosing the dither strategy for years 2-10.

Analyzing the FoM more closely, we observe that the gold sample leads to smaller FoM than both the shallower and deeper catalogs. The larger FoM for shallower catalog is realistic, given less structure with shallow depth leads to weaker artifacts and the shot noise is negligible over

the full ten-year survey, but the out-of-trend behavior of gold sample hints at a peculiarity of the variance across the $ugri$ bands at that depth for the baseline cadence. We investigate this behavior briefly and find that the u-band-induced artifacts add the most to the uncertainties in the bias induced by the observing strategy, as the gold sample u-band cadence in the `minion_1016` is different from gri cadences. This issue still needs to be further investigated, potentially incorporating the importance of each band to calculate an overall bias induced by the observing strategy. We note, however, that this peculiarity is particularly enhanced for the undithered survey.

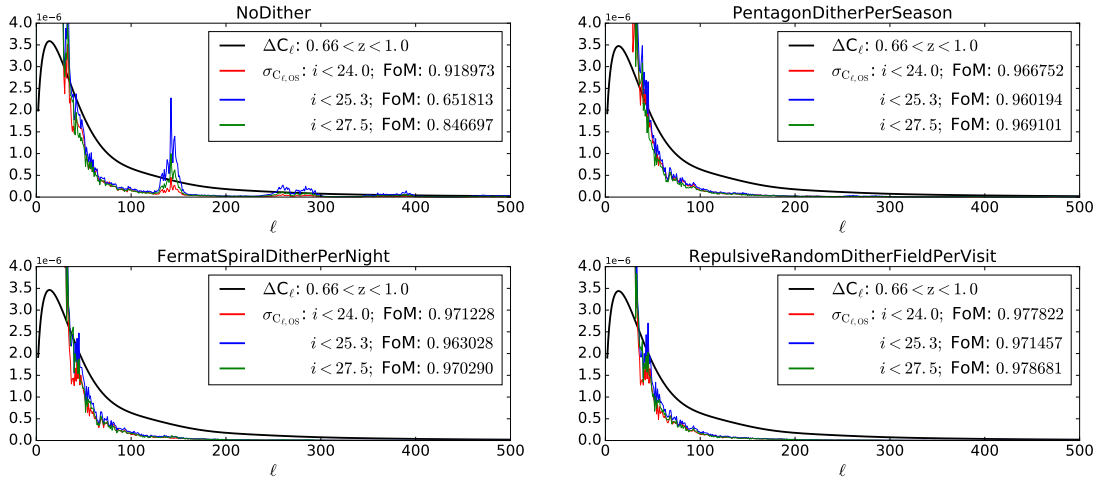

Figure 2.B.5: $\sigma_{C_{\ell,\mathrm{OS}}}$ comparison with the minimum statistical uncertainty ΔC_ℓ for $0.66 < z < 1.0$ for different magnitude cuts after the full, 10-year survey based on `minion_1016`.

The trends observed here remain consistent for all five redshift bins. We note that our choice of dithers is particularly important for the one-year survey as only one of the three dither strategies leads to a large FoM. Therefore, in the absence of effective dithers, systematics correction methods will become necessary after the one-year survey. However, these methods may not lead to significant improvements for a dithered 10-year survey as dithers of most kinds are effective in reducing the uncertainties well below the minimum statistical limit.

To further probe the effects of dithers, we run the 1-year and 10-year analyses for two cadences besides the baseline cadence: `kraken_1043` which does not require visit pairs, and `minion_1020` which implements a Pan-STARRS-like observing strategy offering a larger area coverage. In Figure 2.B.6, we compare the results from these two cadences with those from `minion_1016` for $0.66 < z < 1.0$ for the $i < 25.3$ galaxy sample after only one year of survey. We see that the undithered survey leads to large uncertainties in the bias induced by the observing strategy with all three cadences, with the peak uncertainty 5-15$\times$ the statistical floor. As expected, the undithered survey with the wider coverage `minion_1020` cadence leads to

stronger artifacts and a much smaller FoM (by $\sim 33\%$ in comparison with `minion_1016`), while not requiring visit-pairs is slightly more effective than the baseline (FoM increases by about 6%). We see very similar trends for the three cadences for PerSeason dithers although the peak $\sigma_{C_{\ell,\rm OS}}$ ranges between 3-9$\times$ the statistical floor; FoM based on `minion_1020` is worse than that from `minion_1016` by about 25% and `kraken_1043` improves on the baseline FoM by $\sim 5\%$.

As before, $\sigma_{C_{\ell,\rm OS}}$ improves with more frequent dithering. It is only about 1-3$\times$ the statistical floor for FermatSpiral dithers on PerNight timescale. In contrast to NoDither and PerSeason dithers, both `minion_1020` and `kraken_1043` perform better than baseline`minion_1016` with PerNight dithers: FoM from the wider coverage cadence is about 4.5% better than for the baseline cadence, while we see a 4% better FoM with `kraken_1043`.

For RepulsiveRandom dithers on FieldPerVisit timescale, we find that the uncertainties in the bias induced by the observing strategy are on the same scale as the statistical floor. The wider coverage cadence outperforms the baseline cadence significantly as the wider survey FoM is about 18% better than the baseline FoM while the improvement is about 3% when not requiring visit-pairs. We emphasize that the differences between results with different cadences are highly dependent on the observing strategy: the wider coverage with no or infrequent dithers performs quite poorly while it significantly improves the FoM when large, frequent dithers are implemented. On the other hand, not requiring visit-pairs leads to comparatively larger improvement for infrequent dithers than frequent ones (compared to the baseline).

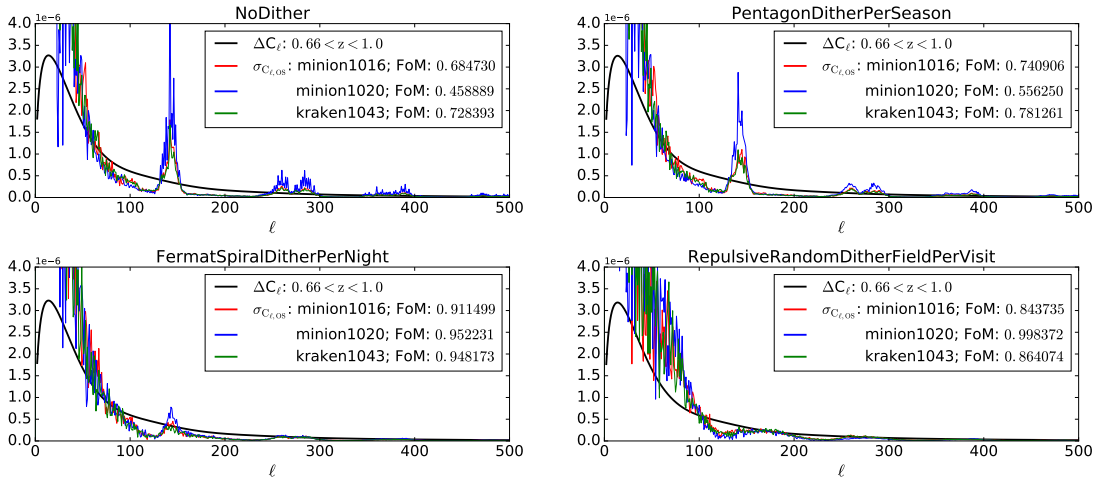

Figure 2.B.6: $\sigma_{C_{\ell,\rm OS}}$ comparison with the minimum statistical uncertainty ΔC_ℓ for $0.66 < z < 1.0$ for three different cadences for $i < 25.3$ after only one year of survey.

Finally, we show the simulated results for different cadences after the 10-year survey in Figure 2.B.7. As in Figure 2.B.5, we see that all the dithered surveys effectively minimize the uncertainties, regardless of the cadence. We do observe, however, that the wider coverage

`minion_1020` still underperforms significantly for the undithered survey (FoM about 30% less than baseline FoM) while all the dithered surveys see a stark improvement (FoM > 1 for all; $\sim 20\%$ improvement on the baseline FoM). The improvement from `kraken_1043` is comparable among the four observing strategies. Based on these results, we note that wider coverage offers significant improvements with large dithers on any implementation timescale.

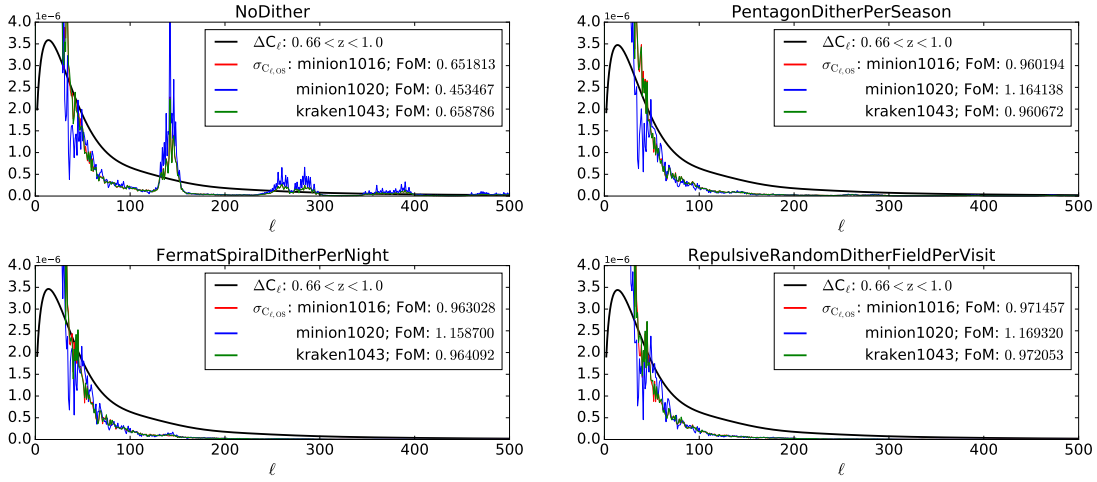

Figure 2.B.7: $\sigma_{C_{\ell,\mathrm{OS}}}$ comparison with the minimum statistical uncertainty ΔC_ℓ for $0.66 < z < 1.0$ for three different cadences for $i < 25.3$ after the full, 10-year survey.

Chapter 3

Milky Way Dust Systematics and LSST Survey Footprint

This chapter presents contributions from H. Awan's work for two White Papers: 1) Lochner et al. (2018), titled *Optimizing the LSST Observing Strategy for Dark Energy Science: DESC Recommendations for the Wide-Fast-Deep Survey* and authored by Lochner, M., Scolnic, D. M., Awan, H., and 33 authors; and 2) Olsen et al. (2018), titled *A Big Sky Approach to Cadence Diplomacy*, and authored by Olsen, K., and 9 authors, including Awan, H. These papers were written in response to the LSST Project White Paper Call in 2018, aimed at gathering feedback from various Science Collaborations on LSST observing strategy; content derived directly from the publications is reproduced here with permission.

3.1 Introduction

Milky Way dust affects our observation of all extragalactic sources, not only reducing the depth of our surveys but also inducing non-uniformities that correlate with the galactic dust, thus making dust uncertainties a limiting factor in our analyses. To address these issues, we investigate the impacts of Milky Way (MW) dust extinction on dark energy studies with LSST and demonstrate that nearly ∼25% of the current LSST Wide-Fast-Deep (WFD) survey area is not useful for dark energy science due to high dust extinction: the extinction renders this survey area too shallow and hence disallows access to constraining information for cosmological studies.

3.2 Methods and Results

Specifically, for the extragalactic static science using high S/N measurements, we must restrict our analysis to a footprint that will give us the deep, high S/N galaxy samples we need for our science. To achieve this, we implement an extinction cut and a depth cut, retaining only the survey area with E(B-V)<0.2 with limiting i-band coadded 5σ depth of 26.0 for Y10; the E(B-V) restriction ensures that we consider the area with small dust uncertainties (Schlafly & Finkbeiner, 2011) while the depth cut ensures that we have high S/N galaxies, with Y10 cut fixed by the LSST SRD goal of yielding a gold sample after Y10 (defined by $i < 25.3$ (LSST Science Collaboration et al., 2009), where i denotes the i-band coadded 5σ depth after

accounting for Milky Way dust extinction). Furthermore, in order to ensure good photometric redshift quality, we focus on the survey footprint that has coverage in all six LSST filters. For Y10, before any selection cuts, the LSST WFD survey comprises ∼19,091 deg^2 in the current baseline, `baseline2018a`. Implementing our selection cuts (on coadded depth and extinction, alongside requiring observations in all filters) reduces the usable footprint to ∼14,645 deg^2, hence discarding ∼23% of the WFD survey area[1]. Figure 3.1 shows the skymaps before and after our selection cuts for the i-band coadded 5σ depth after Y10.

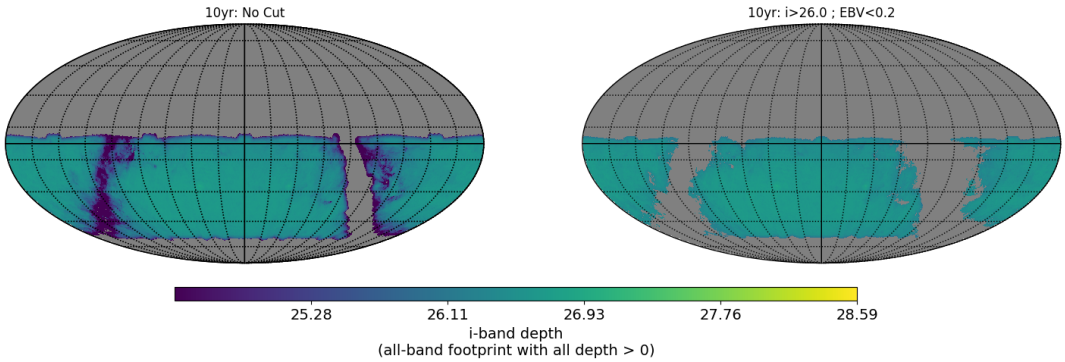

Figure 3.1: Skymaps for Y10 i-band coadded 5σ depth for `baseline2018a`, limited to the footprint with coverage in all six bands, with random, per night translational dithers. *Left*: Before any selection cuts; yielding ∼19,091 deg^2 in usable survey area. *Right*: After a depth cut of $i > 26.0$ and an extinction cut of E(B-V)< 0.2; yielding only ∼14,645 deg^2.

We find similar results when we implement selection cuts on Y1 data, now with a limiting 5σ magnitude of $i = 24.5$: the WFD footprint drops from ∼18,085 deg^2 to ∼13,613 deg^2, rendering ∼25% of the nominal survey area unusable for our purposes. Figure 3.2 shows the skymaps before and after our selection cuts for the i-band coadded 5σ depth.

In order to circumvent the issue of discarding a significant fraction of the WFD region, we propose to shift the nominal 18,000 square degree WFD footprint away from the Galactic Plane to lie entirely within the region where E(B-V)< 0.2, allowing for optimization of the WFD footprint that is usable for our science. To illustrate this, we consider the wider-coverage `OpSim` cadence, `pontus_2002` and implement similar selection cuts as we did for `baseline2018a`. For Y1, the final footprint consists of ∼15,544 deg^2 while Y10 footprint comprises ∼19,254 deg^2, illustrating that the WFD footprint can be optimized to yield a sufficiently large extragalactic footprint; Figures 3.3-3.4 show the analogs of Figures 3.1-3.2 for `pontus_2002`.

The optimization of the LSST WFD survey footprint effectively yields ∼ 25% more survey

[1]Note that the depth cut implemented here not only removes the shallow survey area resulting from high dust extinction but also the shallow borders around the main survey area that result from translational dithering on a fixed `HEALPix` grid, as is done in `OpSim`.

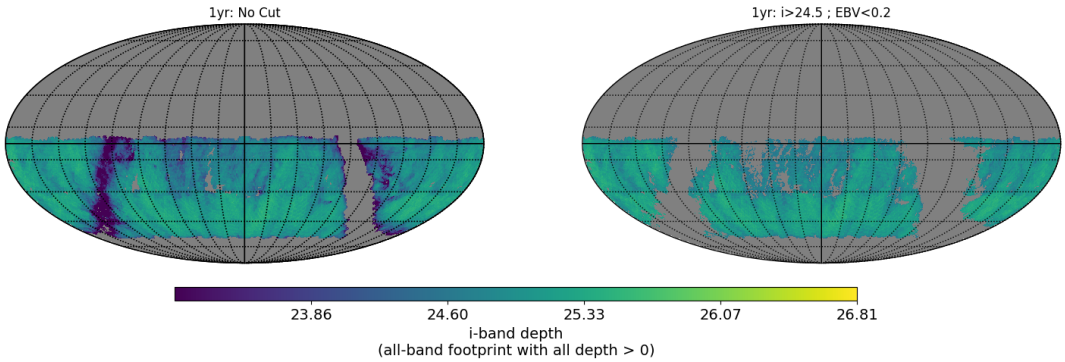

Figure 3.2: Skymaps for Y1 i-band coadded 5σ depth for `baseline2018a`, limited to the footprint with coverage in all six bands with random, per night translational dithers. *Left*: Before any selection cuts; yielding ∼19,091 deg^2 in usable survey area. *Right*: After a depth cut of $i > 24.5$ and an extinction cut of $E(B-V) < 0.2$; yielding only ∼14,645 deg^2.

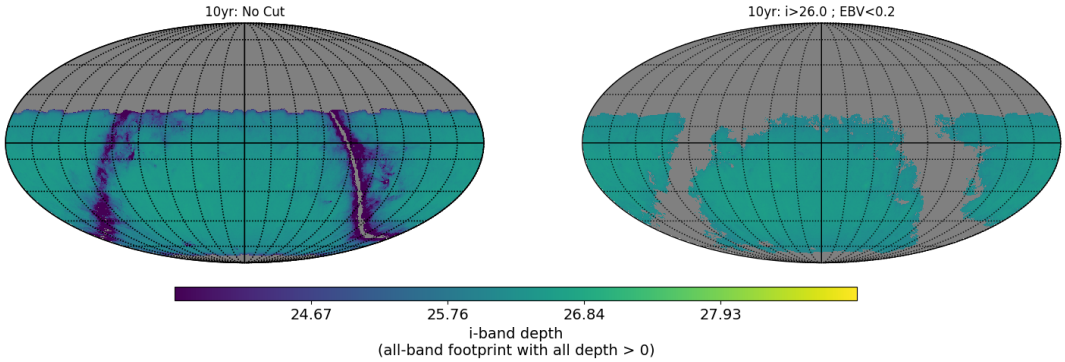

Figure 3.3: Skymaps for Y10 i-band coadded 5σ depth for `pontus_2002` (a wider-survey footprint), limited to the footprint with coverage in all six bands with random, per night translational dithers. *Left*: Before any selection cuts. *Right*: After a depth cut of $i > 26.0$ and an extinction cut of E(B-V)< 0.2; yielding ∼19,254 deg^2.

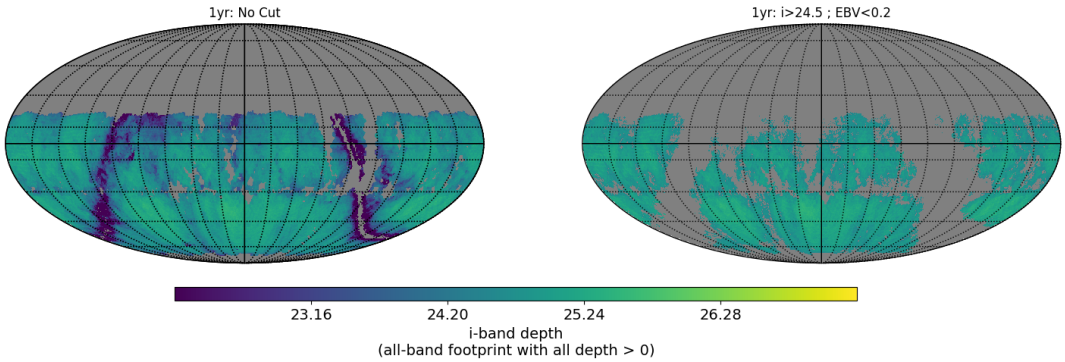

Figure 3.4: Skymaps for Y1 i-band coadded 5σ depth for `pontus_2002` (a wider-survey footprint), limited to the footprint with coverage in all six bands with random, per night translational dithers. *Left*: Before any selection cuts. *Right*: After a depth cut of $i > 24.5$ and an extinction cut of $E(B-V) < 0.2$; yielding ∼15,544 deg^2.

area, making the survey more constraining for dark energy science. This optimization was formally proposed to the LSST Project via Lochner et al. (2018). Figure 3.5 shows the skymaps,

comparing the baseline footprint with the optimized footprint; note that the proposal used the latest baseline cadence, `kraken_2002` – an upgrade from `baseline2018a` that was discussed above as it included dome crawls[2]. Not only does the optimized footprint yield more survey area for LSST, it also increases the overlap with spectroscopic surveys like 4MOST-TiDES (de Jong, 2011) and DESI (DESI Collaboration et al., 2016), which will be instrumental in calibrating the photometric samples from LSST.

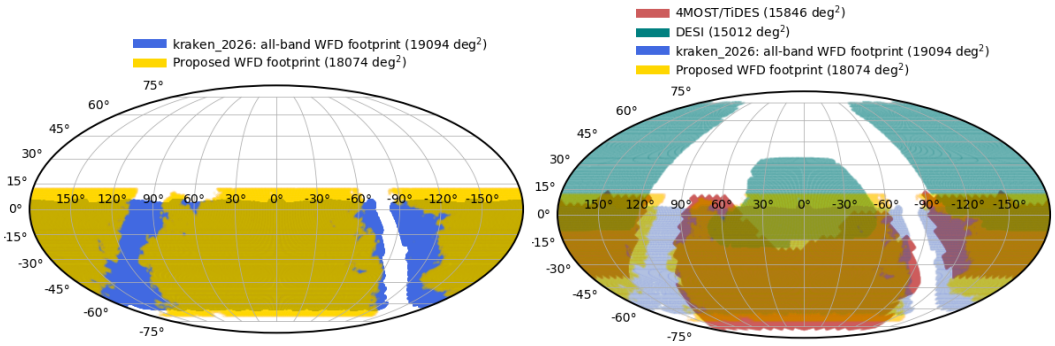

Figure 3.5: *Left*: Skymap showing the current baseline (in blue) and the optimized footprint (in yellow). *Right*: Figure 3 from Lochner et al. (2018), showing not only the baseline and the optimized footprint but also the expected survey footprints for upcoming spectroscopic surveys; included here with permission.

We also proposed a variation of the optimized footprint to the LSST Project via a cross-collaboration study Olsen et al. (2018). Here, the WFD survey was defined with selection cuts on the galactic latitude instead of the extinction (not optimal for dark energy science but better than no mechanism to avoid surveying the anti/galactic plane), with the minimum criteria for WFD visits met as outlined in LSST SRD. Figure 3.6 shows the skymaps for baseline (left) and proposed (right) LSST survey, with the optimized footprint significantly increasing the overlap with DESI (5912 deg^2 for WFD and 4538 deg^2 for non-WFD) vs. baseline (3739 deg^2 for WFD and 2233 deg^2 for non-WFD).

[2]Dome crawls refer to small dome movements during an exposure. More details regarding the crawling model can be found in the document detailing the description of the simulation.

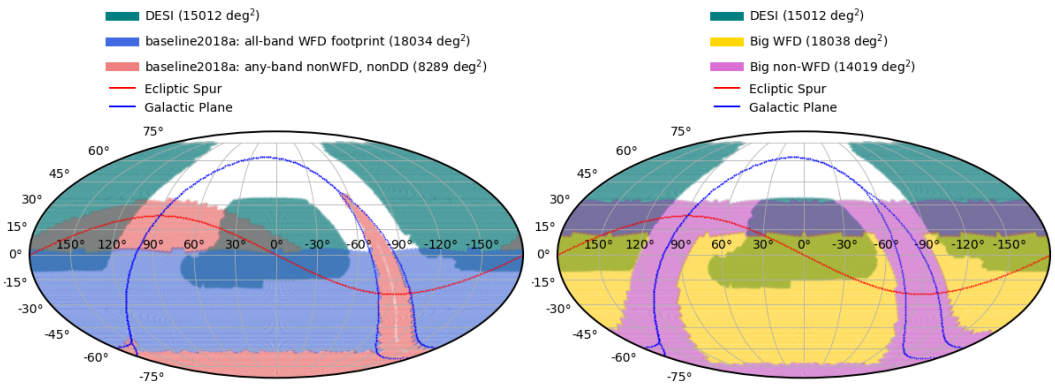

Figure 3.6: Figure adapted from Figure 1 in Olsen et al. (2018); included here with permission. *Left*: Skymap showing the current baseline footprint (WFD in blue; non-WFD in coral), alongside the survey footprint for DESI (in acqua green). *Right*: Skymap showing the proposed footprint (WFD in yellow; non-WFD in pink), alongside the survey footprint for DESI (in acqua green).

Further extending the analysis, we considered various cadences and calculated not only the useable survey footprint but also the overlap with spectroscopic samples. Figure 3.7 shows the area of the footprint resulting from various simulations that is useable for extragalactic science. We note that Y1 is especially sensitive to the specific cadence, and while the different kinds of cadences/footprints converge for Y3-Y10 area, very few simulations yield close to the 18,000 deg^2 WFD area for extragalactic science. Figure 3.8 shows the overlap area for the different cadences, where we see that only some cadences lead to a footprint that overlaps completely with at least one of the two spectroscopic surveys.

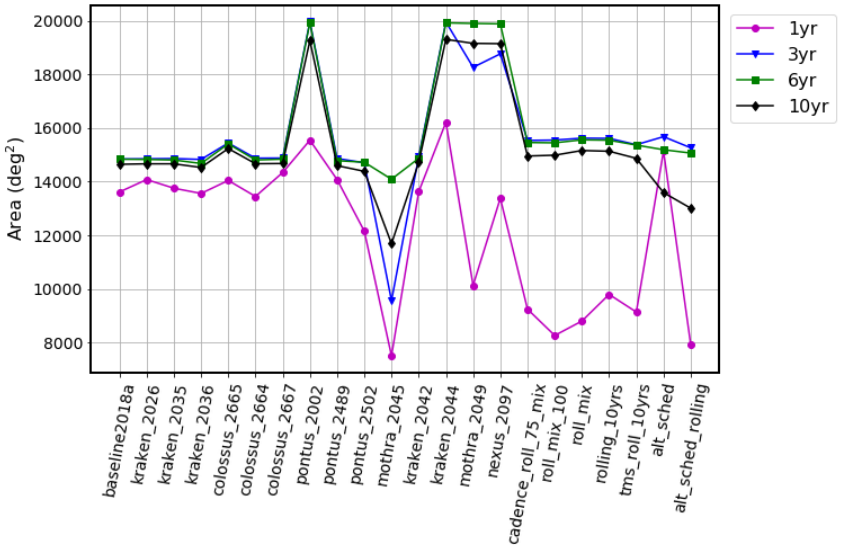

Figure 3.7: Footprint area for extragalactic science passing an extinction and depth cut alongside having coverage in all six of LSST filters, and hence yielding high S/N galaxies.

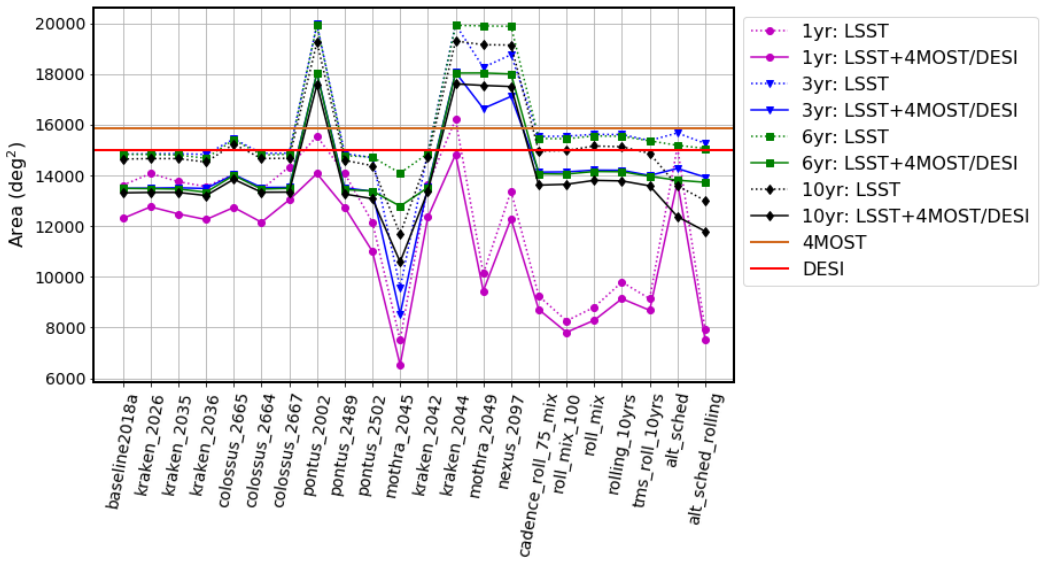

Figure 3.8: Overlapping area (in magenta, blue, green and black solid curves) between the extragalactic science footprint for LSST (i.e., that passing a depth and extinction cut, alongside having coverage in all six filters) and spectroscopic surveys like 4MOST and DESI; note that the full footprint area for 4MOST and DESI is shown in orange and red respectively while the overall extragalactic footprint area for LSST (from Figure 3.7) is shown here in dotted curves.

3.3 Conclusions

In this chapter, we considered the impacts of Milky Way dust extinction and analyzed the effective survey area and median survey depth at various times during the 10-year survey. While the efforts to optimize the observing strategy are underway, our work demonstrates the importance of considering the statistics at various intermediate points during the survey, e.g., since while it may make sense to expect depth cuts to scale as $\sqrt{t}$, the interim depth is sensitive to various factors and may require redefining depth cuts, with the final Y10 depth still as expected. We also note that this work continues to inform the ongoing survey optimization work, as the science collaborations work with the LSST Project; our analysis of the latest simulations is currently in development (Lochner et al. for the LSST-DESC, 2020, in prep).

Chapter 4

Redshift Contamination and Correlation Function Estimators

This chapter is reproduced, aside from minor formatting changes, from Awan & Gawiser (2020): published in the Astrophysical Journal © AAS; titled *Angular Correlation Function Estimators Accounting for Contamination from Probabilistic Distance Measurements*; authored by Humna Awan and Eric Gawiser. Reproduced here with permission.

Note also that, as for Chapter 2, the content here is updated to use the new terminology: LSST now stands for the Legacy Survey of Space and Time, which is carried out by the Vera C. Rubin Observatory, which was previously known as the Large Synoptic Survey Telescope (abbreviated LSST).

4.1 Introduction

Various probes exist to study the cause of cosmic acceleration, one of which is the evolution of large-scale structure (LSS) as traced by clustering in the spatial distribution of galaxies (Cooray & Sheth, 2002). The standard metric to quantify galaxy clustering is the two-point correlation function (CF) or its Fourier transform, the power spectrum. Galaxy clustering can be measured in 3D using spectroscopic surveys, where precise radial information is available, or by measuring the 2D correlations in tomographic redshift bins when only photometric data is available.

Several large astronomical surveys are coming online in the next decade, allowing access to an unprecedented amount of data and hence the ability to measure the evolution of LSS to high precision. These surveys include the Legacy Survey of Space and Time (LSST) (LSST Science Collaboration et al., 2009) carried out by the Vera C. Rubin Observatory (Rubin Obs.), Dark Energy Spectroscopic Instrument (DESI Collaboration et al., 2016), Euclid (Laureijs et al., 2011), and WFIRST (Spergel et al., 2015). The large datasets, however, present new challenges, among which are understanding, mitigating, and accounting for the impacts of systematic uncertainties that exceed the statistical uncertainties; these include uncertainties due to sample contamination, arising either due to photometric redshift uncertainties or spectroscopic line misidentification. Various studies have presented methods to mitigate these effects; e.g., Elsner et al. (2016) and Leistedt et al. (2016) present mode projection as a way to account for systematics, and Shafer & Huterer (2015b) present methodology to handle multiplicative errors like photometric calibration errors.

Various estimators exist to measure the CFs, with the most widely used one introduced in Landy & Szalay (1993) (referred to as LS93 hereafter); see e.g., Kerscher et al. (2000) for a comparison of the various analog estimators, while Vargas-Magaña et al. (2013) and Bernstein (1994) are example studies that consider involved optimizations of the estimators. These estimators can also be extended for various purposes using the overarching idea of 'marked' statistics, which employ weights, or 'marks', for different quantities: they can be used to account for additional dependencies in the correlation functions (see e.g., Sheth & Tormen, 2004; Harker et al., 2006; Skibba et al., 2006; White & Padmanabhan, 2009; Sheth et al., 2005; Robaina & Bell, 2012; Hernández-Aguayo et al., 2018; White, 2016), extract characteristic-dependent correlations (see e.g., Beisbart & Kerscher, 2000; Armijo et al., 2018), or be used to account for different systematics or to extract target features. For instance, Feldman et al. (1994) present a simple weighting that accounts for the signal-to-noise differences coming from each tomographic volume (which was applied e.g., when measuring the Baryonic Acoustic Oscillations (BAO) in Eisenstein et al. 2005); Ross et al. (2017) extend the weights in Feldman et al. (1994) to handle photometric redshift (photo-z) uncertainties for BAO measurements while Peacock et al. (2004) extend them to account for luminosity-dependent clustering, which then are extended by Pearson et al. (2016) for minimal variance in cosmological parameters; Zhu et al. (2015) and Blake et al. (2019) use weights to optimize the BAO measurements; Bianchi et al. (2018) employ weights to account for spectroscopic fibre assignment; Ross et al. (2012) use them to handle systematics, as do Morrison & Hildebrandt (2015); while Bianchi & Percival (2017) and Percival & Bianchi (2017) employ them for 3D correlations to not only correct for missing observations but to improve clustering measurements.

In this paper, we focus on the impacts of sample contamination on the angular correlation functions (ACF). As alluded to earlier, ACFs are especially relevant for photometric surveys, for which we can either measure the projected CFs (e.g., see Zehavi et al. 2002; Zehavi et al. 2011) or the ACFs in redshift bins (e.g., see Crocce et al. 2016; Balaguera-Antolínez et al. 2018; Abbott et al. 2018). Note that one can also measure the ACFs without the tomographic binning (e.g., as in Connolly et al. 2002; Scranton et al. 2002) but that disallows mapping the evolution of the galaxy clustering. Photo-z uncertainties make measuring ACFs in tomographic bins more challenging as the uncertainties introduce spurious cross-correlations across the redshift bins (e.g., see Bailoni et al. 2017 for a study on the impacts of bin cross-correlations on cosmological parameters) and smear out valuable cosmological information, including the BAO (e.g., as in Chaves-Montero et al. 2018). Since the traditional ACF estimators do not account for contamination arising from photo-z uncertainties, the standard tomographic clustering

analysis entails estimating $N(z)$, i.e., the number of galaxies as a function of redshift, in each nominal redshift bin and forward modeling the contaminated ACFs using the $N(z)$ estimates (e.g., as in Crocce et al. 2016; Balaguera-Antolínez et al. 2018; Abbott et al. 2018); also see e.g., Newman (2008) for a discussion on estimating $N(z)$. While this method allows cosmological parameter estimation, it suffers some key limitations as forward modeling is not commonly used outside of cosmology. Furthermore, the variance on the cosmological parameters could potentially be reduced if sample contamination were accounted for directly, instead of being forward modeled, to yield a higher S/N BAO signal from photometric samples.

We propose a method to measure the ACFs *while* accounting for contamination and without needing to forward model the $N(z)$. Specifically, we first introduce a formalism that uses the observed cross correlations to account for sample contamination. Using this formalism, we propose our first estimator, which still uses the photo-z point estimates and the standard CF estimator, but corrects for contamination. Then, we introduce a new estimator that incorporates not just the photo-z point estimates but each galaxy's entire photo-z probability distribution function (PDF; of which photo-z is only representative), by weighting each galaxy based on its photo-z PDF. We note that while the second estimator extends the idea of marked statistics, as discussed above, it differs from the applications in the literature on several fronts. In particular, it avoids the loss of information caused by placing galaxies in a single redshift bin based on their photo-zs, thereby allowing us to counter the impacts of sample contamination with the statistical power of a large dataset, as well as potentially allowing low-variance measurements of the full correlation functions. We return to some of these points for a more thorough discussion of the various differences between our work and that in the literature.

This paper is structured as follows: in Section 4.2, we formally introduce the ACF and its standard estimator. In Section 4.3, we introduce terminology to address sample contamination in the most general sense, followed by our first estimator to correct for sample contamination; we refer to this as the `Decontaminated` estimator. In Section 4.4, we introduce a weighted estimator in which the weights can be chosen to track the probability of each galaxy lying in each redshift bin; we refer to this as the `Weighted` estimator; it is followed by a `Decontaminated Weighted` estimator that estimates the true CFs. We present our validation method in Section 4.5, where we start with a toy example to illustrate the impacts of photo-z uncertainties, followed by a realistic example of measuring the ACFs in three redshift bins, demonstrating the effectiveness of the estimators in recovering the true correlation functions and their covariance matrices in the presence of sample contamination. We discuss our results in Section 4.6 and conclude in Section 4.7.

4.2 2D Two-Point Correlation Function

The most common statistic to study galaxy clustering is the two-point correlation function. The 2D angular correlation function $w_{\alpha\beta}(\theta)$ measures the excess probability of finding a galaxy of Type-α at an angular distance θ from a galaxy of Type-β, in comparison with a random distribution (Peebles, 1993):

$$dP_{\alpha\beta}(\theta) = \eta_\alpha \eta_\beta \left[1 + w_{\alpha\beta}(\theta)\right] d\Omega_\alpha d\Omega_\beta \tag{4.1}$$

where $dP_{\alpha\beta}(\theta)$ is the probability of finding a pair of galaxies of Type-$\alpha\beta$ at an angular distance θ, η_α is the observed sky density of Type-α galaxies in the projected catalog, and $d\Omega$ is the solid angle element at separation θ. An estimator for the correlation function can be constructed as the ratio of number of data-data pairs compared to the number of random-random pairs at a given angular separation:

$$w_{\alpha\beta}(\theta_k) = \frac{(DD)_{\alpha\beta}(\theta_k)}{(RR)_{\alpha\beta}(\theta_k)} - 1 \tag{4.2}$$

where $(DD)_{\alpha\beta}(\theta_k)$ is the normalized number of data-data pairs at angular separation θ_k, and $(RR)_{\alpha\beta}(\theta_k)$ is that for the random-random pairs; the index k emphasizes the binned nature of the estimator. We note that Equation 4.2 leads to an auto-correlation function when $\alpha = \beta$ and cross-correlation otherwise; for the cross-correlation, we explicitly consider independent random catalogs for the two populations, accounting for the case when the two samples do not completely overlap in their angular range. We also note that each histogram can be written using the Heaviside step function, defined as

$$\Theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{4.3}$$

For instance, for the auto-correlation, we have

$$\begin{aligned} (DD)_{11}(\theta_k) &= \frac{\sum_i^{N_1} \sum_{j>i}^{N_1} \Theta(\theta_{ij} - \theta_{\min,k})[1 - \Theta(\theta_{ij} - \theta_{\max,k})]}{\sum_i^{N_1} \sum_{j>i}^{N_1}} \equiv \frac{\sum_i^{N_1} \sum_{j>i}^{N_1} \bar{\Theta}_{ij,k}}{\sum_i^{N_1} \sum_{j>i}^{N_1}} \\ &= \frac{\sum_i^{N_1} \sum_{j\neq i}^{N_1} \bar{\Theta}_{ij,k}}{\sum_i^{N_1} \sum_{j\neq i}^{N_1}} \end{aligned} \tag{4.4}$$

where

$$\bar{\Theta}_{ij,k} \equiv \Theta(\theta_{ij} - \theta_{\min,k})[1 - \Theta(\theta_{ij} - \theta_{\max,k})] \tag{4.5}$$

Here, θ_{ij} is the angular separation between the ith and jth galaxy in the data sample of N_1 galaxies, and we have explicitly written out the histogram: the kth bin counts the number of galaxy pairs at separations $\theta_{\min,k} \leq \theta_{ij} < \theta_{\max,k}$. Note that the normalized histograms can be calculated either by considering all unique pairs or with double counting, as long as the normalization accounts for the total pairs; the denominator in the case where we count only the unique pairs yields the familiar count of $N_1(N_1-1)/2$ pairs.

Similar to Equation 4.4, we can write the histogram for the cross-correlation function as

$$(DD)_{12}(\theta_k) = \frac{\sum_i^{N_1}\sum_j^{N_2}\bar{\Theta}_{ij,k}}{\sum_i^{N_1}\sum_j^{N_2}} \tag{4.6}$$

where sample α contains N_α galaxies.

We note here that the estimator in Equation 4.2 differs only slightly from the estimator introduced in LS93 (referred to hereafter as the LS estimator). In the absence of sample contamination, the LS estimator is unbiased and has Poissonian variance but we choose to work with the simpler estimator since the LS estimator accounts for edge-effects that become subdominant to sample contamination when using large galaxy surveys. Specifically, we note that the DD/RR estimator presented above is as (un)biased as the LS estimator (see Equation 48 in LS93) and its variance reduces to Poissonian variance in the limit of large N (see Equations 42, 48 in LS93). We refer to the DD/RR estimator as the `Standard` estimator, when comparing with the new estimators.

4.3 Standard Estimator and Contaminants

We start with the case of two galaxy types in the observed sample, Type-A and Type-B; either one acts as a contaminant in relation to the other. We assume that we have some method that gives us the probability of each observed galaxy i of being Type-A, $\mathfrak{q}_i^A$ or Type-B, $\mathfrak{q}_i^B$; example methods include, e.g., integration of a galaxy's photo-z PDF in the target redshift bin or a Bayesian classifier as presented in Leung et al. (2017). Assuming that our observed galaxy sample comprises only the two types of galaxies, we have $\mathfrak{q}_i^A + \mathfrak{q}_i^B = 1$, where i runs over all the galaxies in the observed sample.

Now, assuming that the classifier is unbiased, we can use the classification probabilities to estimate the fraction of objects that are contaminants for a given target sample. For this purpose, however, we must divide the full observed sample into target subsamples, i.e., in

the 2-sample case, the observed Type-A and Type-B galaxies.[1] Then, our classifier provides the probability of each observed Type-A galaxy i to be truly of Type-A, q_i^{AA}, as well as the probability of each observed Type-A galaxy to be truly of Type-B, q_i^{AB}. Hence, we have

$$q_i^{AA} + q_i^{AB} = q_j^{BA} + q_j^{BB} = 1 \tag{4.7}$$

where i runs over the observed Type-A sample and j runs over the observed Type-B sample. We can then use the classification probabilities on the observed subsamples to estimate the contamination. That is, we have the fraction of observed Type-A galaxies that are true Type-A or Type-B galaxies given by

$$f_{AA} = \left\langle q_i^{AA} \right\rangle \; ; \; f_{AB} = \left\langle q_i^{AB} \right\rangle \tag{4.8}$$

where the average is over the observed Type-A sample. Equation 4.7 translates into the expected identities on the fractions:

$$f_{AA} + f_{AB} = f_{BA} + f_{BB} = 1 \tag{4.9}$$

These ideas can be generalized to M galaxy samples of Types $A_1, A_2, ..., A_M$, with the classification probabilities on the entire observed sample given by $\mathtt{q}_{A_1}, \mathtt{q}_{A_2}, ..., \mathtt{q}_{A_M}$. Once the full observed catalog is divided into M target subsamples, we have the probability of ith observed galaxy of Type-A_j being of Type-A_m given by $q_i^{A_j A_m}$ and the fraction of observed Type-A_j galaxies that are Type-A_m galaxies given by $f_{A_j A_m}$.

4.3.1 Decontamination

Using the standard ACF estimator, correlations from known contaminated samples can be corrected for by using the fractions $f_{\alpha\beta}$ as defined in Equation 4.8; see e.g., Grasshorn Gebhardt et al. (2019), Addison et al. (2019) for a similar approach. Formally, this is done by writing the observed correlation functions in terms of the true correlation functions by considering the type of galaxy that contributes to each data pair. Here we work with two target galaxy samples, Type-A and Type-B; the generalized case is discussed in Appendix 4.D.1.

Since we have two types of galaxies, we aim to calculate two auto-correlations and one cross-correlation from the contaminated sample: $w_{AA}^{\rm true}(\theta_k), w_{AB}^{\rm true}(\theta_k), w_{BB}^{\rm true}(\theta_k)$. However, if we calculate the correlations on the subsamples directly, we get $w_{AA}^{\rm obs}(\theta_k)$, $w_{AB}^{\rm obs}(\theta_k)$, $w_{BB}^{\rm obs}(\theta_k)$,

[1] A simple way to do this would be to assign all galaxies with $\mathtt{q}_i^A > 0.5$ to target sample A and the rest to target sample B.

which differ from the true correlations due to sample contamination. To construct the relation between the two, lets consider $w_{AB}^{\rm obs}(\theta_k)$ which gets its contributions from four types of pairs: 1) Observed Type-A galaxies that are true Type-A, paired with observed Type-B that are true Type-A, contributing $f_{AA}f_{BA}w_{AA}^{\rm true}(\theta_k)$ to the observed correlation, 2) Observed Type-A that are true Type-A, paired with observed Type-B that are true Type-B, contributing $f_{AA}f_{BB}w_{AB}^{\rm true}(\theta_k)$, 3) Observed Type-$B$ that are true Type-A, paired with observed Type-A that are true Type-B, contributing $f_{AB}f_{BA}w_{AB}^{\rm true}(\theta_k)$, and 4) Observed Type-$A$ that are true Type-B, paired with observed Type-B that are true Type-B, contributing $f_{AB}f_{BB}w_{BB}^{\rm true}(\theta_k)$. Therefore, we have

$$w_{AB}^{\rm obs}(\theta_k) = f_{AA}f_{BA}w_{AA}^{\rm true}(\theta_k) + \{f_{AA}f_{BB} + f_{BA}f_{AB}\}\, w_{AB}^{\rm true}(\theta_k) + f_{AB}f_{BB}w_{BB}^{\rm true}(\theta_k) \tag{4.10}$$

The auto correlations follow similarly, leading us to

$$\begin{bmatrix} w_{AA}^{\rm obs}(\theta_k) \\ w_{AB}^{\rm obs}(\theta_k) \\ w_{BB}^{\rm obs}(\theta_k) \end{bmatrix} = \begin{bmatrix} f_{AA}^2 & 2f_{AA}f_{AB} & f_{AB}^2 \\ f_{AA}f_{BA} & f_{AA}f_{BB} + f_{AB}f_{BA} & f_{AB}f_{BB} \\ f_{BA}^2 & 2f_{BB}f_{BA} & f_{BB}^2 \end{bmatrix} \begin{bmatrix} w_{AA}^{\rm true}(\theta_k) \\ w_{AB}^{\rm true}(\theta_k) \\ w_{BB}^{\rm true}(\theta_k) \end{bmatrix} \tag{4.11}$$

where we note that the contribution from the true cross correlation to the observed auto correlations simplifies (as opposed for that to the observed cross correlation). We also present a formal derivation of the result above using Equation 4.1 in Appendix 4.A.1. Now, using these equations, we can construct the `Decontaminated` estimators $\widehat{w}_{AA}(\theta_k), \widehat{w}_{BB}(\theta_k), \widehat{w}_{AB}(\theta_k)$ for the true correlation functions $w_{AA}^{\rm true}(\theta_k)$, $w_{BB}^{\rm true}(\theta_k)$, $w_{AB}^{\rm true}(\theta_k)$ given by

$$\begin{bmatrix} \widehat{w}_{AA}(\theta_k) & \widehat{w}_{AB}(\theta_k) & \widehat{w}_{BB}(\theta_k) \end{bmatrix}^T = [D_{\rm S}]^{-1} \begin{bmatrix} w_{AA}^{\rm obs}(\theta_k) & w_{AB}^{\rm obs}(\theta_k) & w_{BB}^{\rm obs}(\theta_k) \end{bmatrix}^T \tag{4.12}$$

where $[D_{\rm S}]$ is the square matrix in Equation 4.11, which must be invertible[2]. Appendix 4.D.1 presents the `Decontaminated` estimators for the generalized case of working with M target subsamples. We also note that this decontamination formalism could be easily applied to the LS estimator; the decontamination matrix $[D_{\rm S}]$ does not inherently depend on the usage of the DD/RR estimator.

Given their construction, the `Decontaminated` estimators are unbiased (under the assumption that the contamination fractions are represented by the average classification probabilities);

[2]For the matrix to be non-invertible, its determinant must be zero, which, after many algebraic manipulations, simplifies to the constraint $(f_{AA}f_{BB} - f_{AB}f_{BA})^3 = 0$. Given Equation 4.9, this leads to $f_{AA} = f_{BA}$ and $f_{BB} = f_{AB}$, implying that $w_{AA}^{\rm obs}(\theta_k) = w_{AB}^{\rm obs}(\theta_k) = w_{BB}^{\rm obs}(\theta_k)$, i.e., all the observed correlation functions are equal and hence disallow distinguishing the contributions from the true correlation functions. We do not expect the contamination rate to be high enough to enable this special case.

see Appendix 4.A.2 for more details. As for the variance, the decontamination leads to a quadrature sum of the variance of the standard estimators for each of the auto- and cross-correlations in the absence of covariance between the observed correlations; the closed form expression for the variance as well as the general covariance of the estimators is presented in Appendix 4.A.3. Note that this overarching idea of using contamination fractions is similar to that presented in Benjamin et al. (2010) but their focus is on estimating the contamination fractions *from* the contaminated correlations, for which they resort to approximating the decontamination matrix as diagonal. Since we expect sufficiently strong correlations across the different target samples (e.g., between the neighboring photo-z bins for a tomographic clustering analysis), the simplification of ignoring some contamination fractions becomes undesirable.

4.4 A New, Weighted Estimator

Here, we present an estimator for the observed correlation function that accounts for pair weights, i.e., each pair of galaxies is weighted to account for its contribution to the target correlation function, e.g., by the classification probability of each contributing galaxy (alongside other parameters). This way, we consider the *entire* observed catalog, containing $N_{\rm tot}$ galaxies of both Type-A and Type-B, each with their respective classification probabilities. That is, we propose a `Weighted` estimator for the observed correlation function:

$$\widetilde{w}^{\rm obs}_{\alpha\beta}(\theta_k) = \frac{(\widetilde{DD})_{\alpha\beta}(\theta_k)}{RR(\theta_k)} - 1 \tag{4.13}$$

where α, β are the types, e.g., $\widetilde{w}^{\rm obs}_{AA}$ denotes the estimator for the observed Type-A auto-correlation while $\widetilde{w}^{\rm obs}_{AB}$ denotes the cross-correlation. Here, we define weighted data-data pair counts as

$$(\widetilde{DD})_{\alpha\beta}(\theta_k) = \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \tag{4.14}$$

where $\mathtt{w}_{ij}^{\alpha\beta}$ is the pair weight, with the pair comprised of the ith and jth galaxies, while the weighting is over all $N_{\rm tot}$ galaxies in the observed catalog. We note that the normalization is needed to match the normalization of unweighted correlation functions (Equations 4.4, 4.6). Equation 4.14 therefore allows us to calculate the different weighted data-data pair counts, e.g., $(\widetilde{DD})_{AA}, (\widetilde{DD})_{AB}, (\widetilde{DD})_{BB}$. We also note that $RR(\theta_k)$ is formally $(RR)_{\alpha\beta}(\theta_k)$ since different galaxy samples can have different selection functions. However, since we consider all the galaxies in the observed sample, not just the target subsamples, we take $RR(\theta_k)$ to trace the full survey geometry. We also note that using the DD/RR estimator allows us to introduce

pair weights more naturally here; the LS estimator would make it difficult given the DR term to account for. We include some notes on the implementation of the `Weighted` estimator in Appendix 4.C.2.

In the simplest scenario, the pair weight could be linearly dependent on the probabilities of ith and jth objects being of Type α, β respectively, i.e., $\mathtt{w}_{ij}^{\alpha\beta} = \mathtt{w}_i^{\alpha}\mathtt{w}_j^{\beta} = \mathtt{q}_i^{\alpha}\mathtt{q}_j^{\beta}$. Note that this approach does not require us to break the observed sample into target subsamples as long as intelligent weights are assigned to each galaxy pair. Explicitly, if we have two observed galaxy types in our observed catalog, as was discussed at the beginning of Section 4.3, $\mathtt{w}_i^A = q_i^{AA}$ for observed Type-A while $\mathtt{w}_i^A = q_i^{BA}$ for observed Type-B galaxies. Similarly, $\mathtt{w}_i^B = q_i^{AB}$ for observed Type-A while $\mathtt{w}_i^B = q_i^{BB}$ for observed Type-B. Also note that $N_{\rm tot} = N_{\rm obs}^A + N_{\rm obs}^B = N_{\rm true}^A + N_{\rm true}^B$. Finally, we highlight that our `Weighted` estimator reduces to the `Standard` estimator if $\mathtt{w}_i^{\alpha}$ is set to 1 for observed Type-A galaxies and to 0 for observed Type-B galaxies, and $\mathtt{w}_i^{\beta}$ is set to 0 for observed Type-A galaxies and to 1 for observed Type-B.

4.4.1 Estimator Bias and Variance

The estimator in Equation 4.13 is biased, as it considers the entire sample, including contaminants with different correlation functions. In order to estimate the true correlations using unbiased estimators, $\widehat{w}$, we require that their expectation value approach the true correlations. That is, we have

$$\left\langle \begin{bmatrix} \widehat{w}_{AA}(\theta_k) \\ \widehat{w}_{AB}(\theta_k) \\ \widehat{w}_{BB}(\theta_k) \end{bmatrix} \right\rangle = \left\langle [D_W] \begin{bmatrix} \widetilde{w}_{AA}^{\rm obs}(\theta_k) \\ \widetilde{w}_{AB}^{\rm obs}(\theta_k) \\ \widetilde{w}_{BB}^{\rm obs}(\theta_k) \end{bmatrix} \right\rangle = \begin{bmatrix} w_{AA}^{\rm true}(\theta_k) \\ w_{AB}^{\rm true}(\theta_k) \\ w_{BB}^{\rm true}(\theta_k) \end{bmatrix} \tag{4.15}$$

where $[D_{\rm W}]$ is a decontamination matrix, designed to make the estimators unbiased. It is analogous to the decontamination matrix $[D_{\rm S}]$ in Equation 4.12. Here we explicitly work with the two-sample case, with only Type-A and Type-B galaxies present in our sample.

As done to decontaminate the `Standard` estimators in Section 4.3.1, we calculate the contributions that are coming from each of the true correlation functions to any given weighted

correlation function. That is, we have

$$\left\langle \widetilde{w}^{\rm obs}_{\alpha\beta}(\theta_k) \right\rangle = \frac{\left(\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{q}_i^A \mathtt{q}_j^A \right) w^{\rm true}_{AA}(\theta_k) + \left(\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \left\{ \mathtt{q}_i^A \mathtt{q}_j^B + \mathtt{q}_i^B \mathtt{q}_j^A \right\} \right) w^{\rm true}_{AB}(\theta_k) + \left(\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{q}_i^B \mathtt{q}_j^B \right) w^{\rm true}_{BB}(\theta_k)}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \tag{4.16}$$

We present the full derivation of Equation 4.16 in Appendix 4.B. Consolidating the terms as done in Equation 4.11, we have

$$\begin{bmatrix} \left\langle \widetilde{w}^{\rm obs}_{AA}(\theta_k) \right\rangle \\ \left\langle \widetilde{w}^{\rm obs}_{AB}(\theta_k) \right\rangle \\ \left\langle \widetilde{w}^{\rm obs}_{BB}(\theta_k) \right\rangle \end{bmatrix} = \begin{bmatrix} \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AA} \mathtt{q}_i^A \mathtt{q}_j^A}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AA} \left\{ \mathtt{q}_i^A \mathtt{q}_j^B + \mathtt{q}_i^B \mathtt{q}_j^A \right\}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AA} \mathtt{q}_i^B \mathtt{q}_j^B}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AA}} \\ \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB} \mathtt{q}_i^A \mathtt{q}_j^A}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB} \left\{ \mathtt{q}_i^A \mathtt{q}_j^B + \mathtt{q}_i^B \mathtt{q}_j^A \right\}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB} \mathtt{q}_i^B \mathtt{q}_j^B}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB}} \\ \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BB} \mathtt{q}_i^A \mathtt{q}_j^A}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BB} \left\{ \mathtt{q}_i^A \mathtt{q}_j^B + \mathtt{q}_i^B \mathtt{q}_j^A \right\}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BB} \mathtt{q}_i^B \mathtt{q}_j^B}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BB}} \end{bmatrix} \begin{bmatrix} w^{\rm true}_{AA}(\theta_k) \\ w^{\rm true}_{AB}(\theta_k) \\ w^{\rm true}_{BB}(\theta_k) \end{bmatrix} \tag{4.17}$$

Therefore, the `Decontaminated Weighted` estimators are given by

$$\left[\widehat{w}_{AA}(\theta_k) \;\; \widehat{w}_{AB}(\theta_k) \;\; \widehat{w}_{BB}(\theta_k) \right]^T = [D_{\rm W}]^{-1} \left[\widetilde{w}^{\rm obs}_{AA}(\theta_k) \;\; \widetilde{w}^{\rm obs}_{AB}(\theta_k) \;\; \widetilde{w}^{\rm obs}_{BB}(\theta_k) \right]^T \tag{4.18}$$

where $[D_{\rm W}]$ is the square matrix in Equation 4.17. We note that each row in Equation 4.18 corresponds to final, unbiased weights on each pair, comprised of a sum of three weights – a fact that can be utilized when optimizing weights for minimum variance. We present an example optimization that decontaminates while estimating the correlation functions in Appendix 4.C.3.

We have checked Equation 4.18 in various limiting cases to confirm the validity of its form. Specifically, we first divided the total observed sample into subsamples, and then applied the simplifications that reduce the `Decontaminated Weighted` estimators to `Decontaminated` estimators (i.e., setting the pair weights for the target subsample to unity and the rest to zero, and approximating the classification probabilities with their averages); we confirm that Equation 4.18 does indeed reduce to Equation 4.12, demonstrating that `Decontaminated Weighted`

is the generalized estimator. We then tested the two limiting cases of no contamination and 100% contamination, working with just the observed subsamples and using pair weights that are a linear product of the respective classification probabilities; we confirm that the reduced estimator recovers the truth when there is no contamination while it is indeterminate when there is 100% contamination. Finally, we considered the entire observed sample and tested the limiting cases of no contamination and 100% contamination, with pair weights that are a linear product of the respective classification probabilities, and arrive at true correlations both when there is no contamination and when there is 100% contamination – an advantage of using the full sample. We also present the analytical form of the variance of the `Weighted` estimator in Appendix 4.C.1; since the variance is a function of a four-point sum and depends non-trivially on the pair weights, we choose to estimate the variance numerically using bootstrap as described in Section 4.5.1. Finally, we present the generalized estimator, i.e., applicable to M target samples, in Appendix 4.D.2.

4.5 Validation and Results

In order to test our estimators, we consider the simplest relevant application: tomographic clustering analysis, i.e., the measurement of the ACF for galaxies in different redshift bins. Then, in the context of our terminology in Sections 4.3-4.4, the different 'types' of galaxies are essentially the galaxies in the different redshift bins. For this purpose, we use the publicly available v0.4_r1.4 of MICE-Grand Challenge Galaxy and Halo Light-cone Catalog. The catalog is generated by populating the dark matter halos in MICE, which is an N-body simulation covering an octant of the sky at $0 \leq z \leq 1.4$. Most importantly for our purposes, the catalog follows local observational constraints, e.g., galaxy clustering as a function of luminosity and color, and incorporates galaxy evolution for realistic high-z clustering – allowing for a robust test of the estimators. More details about the catalog can be found in MICE publications: Fosalba et al. (2015a); Crocce et al. (2015); Fosalba et al. (2015b); Carretero et al. (2015); Hoffmann et al. (2015). We query the catalog using CosmoHub (Carretero et al., 2017).

In order to test our method, we must have photo-zs that are realistic for upcoming surveys like the LSST. Since MICE catalog photo-zs are biased and exhibit a large scatter, we simulate adhoc photo-zs using the true redshifts and assuming $\sigma_z = 0.03(1+z)$, the upper limit on the scatter mentioned in the LSST Science Requirements Document[3]. Specifically, we model

[3] https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17; see also LSST Science Collaboration et al. (2009).

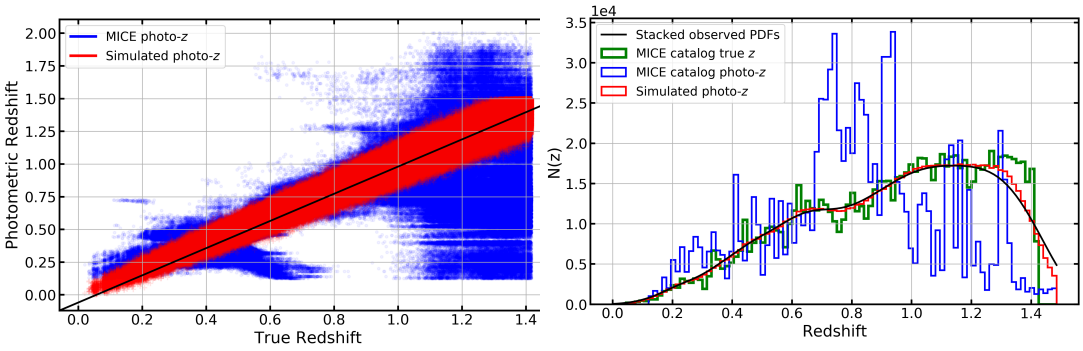

Figure 4.1: Illustration of the simulated photo-zs. *Left*: Comparison between true redshift and MICE catalog photo-zs (blue) vs. those simulated here (red). *Right*: Comparison between the different $N(z)$ distributions: true $N(z)$; those based on MICE catalog photo-zs vs. those simulated assuming Gaussian PDFs with $\sigma_z = 0.03(1+z)$. The red, blue, green are $N(z)$ estimates from binning the respective redshifts, while the black curve is based on stacking the observed photo-z PDFs. We see that our simulated photo-zs are well-behaved and are able to recover the true $N(z)$ effectively. These plots are created using only the galaxies with $0 \leq$ RA ≤ 5 deg, $0 \leq$ Dec ≤ 5 deg, yielding 994,863 galaxies at $0 \leq z \leq 1.4$.

the photo-z probability distribution function (PDF) for each galaxy as a Gaussian with its true redshift as the mean and σ_z as the standard deviation. Then, we randomly draw from the PDF and assign the draw as the photo-z of the galaxy; the "observed PDF" is then a Gaussian with the random draw as the mean and σ_z as the standard deviation. This method generates unbiased photo-zs in a simple way.

Figure 4.1 illustrates our simulated photo-zs: the left panel compares the MICE catalog photo-zs and the simulated photo-zs with the true redshifts, while the right panel shows $N(z)$, the number of galaxies as a function of redshift, as estimated by binning the redshifts as well as by stacking the photo-z PDFs. We see that our simulated photo-z PDFs and the consequent photo-zs effectively recover the overall true galaxy number distribution. Also note that the $N(z)$ from simulated photo-z (solid red) and observed (solid black) PDFs are very similar, indicating that our simulated observed photo-z PDFs are nearly unbiased.

Now, the true catalog essentially consists of the location of the galaxies on the sky (RA, Dec) and the true redshift, while the observed catalog consists of the RA, Dec and photo-zs. In order to test the effects of contamination, we must work with observed subsamples, i.e., galaxies with photo-zs in the target redshift bin; these differ from the true subsamples, which are galaxies with their true redshifts in the target redshift bins. Note that this subsampling is not necessary for the `Weighted` estimator, introduced in Section 4.4, which only needs the photo-z PDFs for all the observed galaxies. We use `TreeCorr` (Jarvis et al., 2004) to calculate the correlation functions.

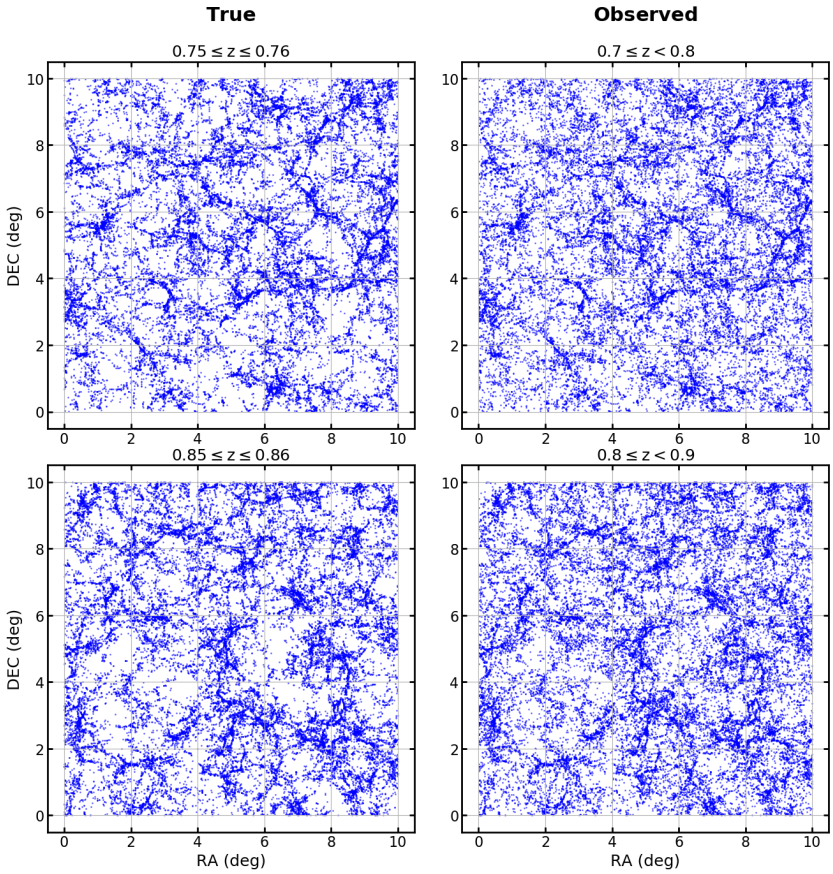

Figure 4.2: True and observed positions of galaxies for the idealized galaxy sample of Section 4.5.1, where all the true galaxies lie at $0.75 \leq z \leq 0.76$, $0.85 \leq z \leq 0.86$. We see that redshift binning of galaxies based on photo-z point estimates modifies the LSS due to the redshift contamination.

4.5.1 Toy Example

In order to illustrate the impacts of photo-zs, we consider a toy example: a clustering analysis using only two tomographic bins ($0.7 \leq z < 0.8$, $0.8 \leq z < 0.9$) with the true galaxy sample having galaxies only at $0.75 \leq z \leq 0.76$, $0.85 \leq z \leq 0.86$, but with the photo-z scatter as mentioned before, i.e., $\sigma_z = 0.03(1+z)$. We query the true galaxies in nine 10x10 deg^2 patches along Dec = 0; all patches have a similar number of galaxies (66K-78K) and face similar photo-z contamination rates (22-25%, 18-21% in the two tomographic bins, respectively). To make explicit the impacts of redshift binning based on photo-z point estimates, we show the true and observed positions of the galaxies in the two redshift bins in Figure 4.2, where we can see that the two distributions are different, with photo-z uncertainties mixing the LSS between the two bins. Figure 4.3 shows the distributions of the true and photometric redshifts using one of the patches (with 66,927 galaxies, and 23% and 20% contamination in the two tomographic bins, respectively).

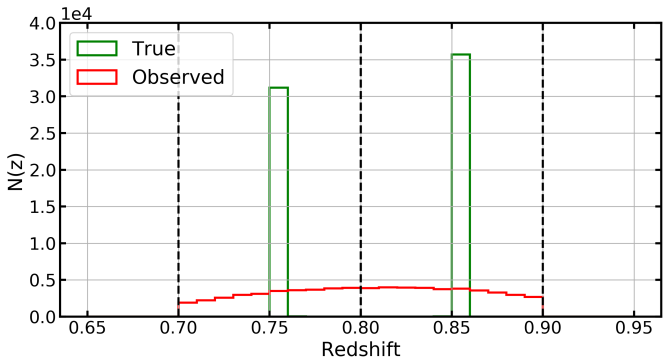

Figure 4.3: True and observed redshift histograms for the idealized galaxy sample of Section 4.5.1, with redshift bin edges shown using the vertical dashed lines. We see that photo-z uncertainties lead to a smearing of the redshift information.

Then, using the observed photo-z PDFs, we calculate the classification probabilities as the integral of the PDFs within the target redshift bin. Note that since we are simulating only two bins, we use Gaussian PDFs truncated at $z = 0.7$ and $z = 0.9$ to ensure that we conserve the number of true and observed galaxies; this yields a slight bias in the PDF integrations, which we correct to make the overall classification probabilities unbiased, i.e., $\left\langle q_i^{AB} \right\rangle = f_{AB}$, where the average is checked over redshift intervals with $\Delta z = 0.02$, while ensuring the de-biased probabilities remain in the range 0-1. For real data, this debiasing should be possible utilizing a limited set of spectroscopic redshifts. Figure 4.4 shows the distribution of the final classification probabilities for all the galaxies in our observed sample. In order to estimate the various correlation functions (two auto, one cross) and their variance, we consider the 9 patches: the mean across the nine samples gives us the mean estimate of the respective correlation function while we calculate the estimator variance as $\left\langle \{\widehat{w}_i(\theta_k) - w_i^{\rm true}(\theta_k)\}^2 \right\rangle$ where i runs over all the correlations (both auto and cross) and the expectation value is over all the realizations; note that this variance is not sensitive to the sample variance but only a measure of the estimator variance, which we can calculate explicitly given that we have access to the true CFs in each of the nine patches. Note that for each of the patches, we calculate five types of the three correlation functions: those in the true subsamples; those using the `Standard` estimator on the contaminated observed subsamples, followed by those from the `Decontaminated` estimators; and those using the `Weighted` estimator, followed by the `Decontaminated Weighted` ones. Also, we use a random catalog that is 5x the size of the data catalog, and restrict CF calculation to 0.01-3deg scales. Figure 4.5 shows our results, with both the correlation functions and their variance. As expected, the cross correlations with contamination are non-negligible, taking signal away from the two auto-correlations. Decontamination lowers the amplitude of the cross-correlations, and we find that both estimators correct for the contamination and reduce the bias,

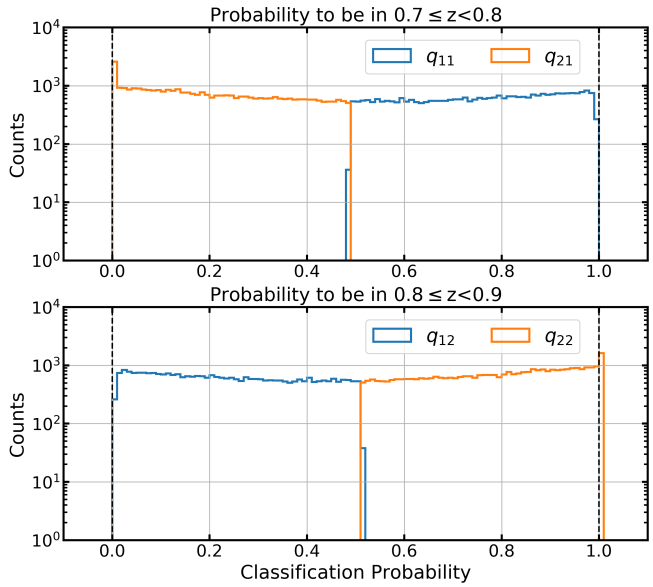

Figure 4.4: Distribution of the classification probabilities to be in bin 1 (upper panel) or bin 2 (lower panel) for the toy galaxy sample of Section 4.5.1. As introduced in Section 4.3, $q_{\alpha\beta}$ is the probability of the observed Type-α galaxy to be a true Type-β galaxy. We see that given the photo-z uncertainties, the probability to be in a given target tomographic bin has a broad range. Note that the two panels are mirror images of one another, as dictated by the identity in Equation 4.7.

leading to estimates closer to the truth. This is more apparent in Figure 4.6, where we show the bias in the correlation functions (i.e., difference from the truth calculated as $\langle \widehat{w}_i(\theta_k) - w_i^{\rm true}(\theta_k) \rangle$ where i runs over all the correlations (both auto and cross) and the expectation value is over all the realizations). We note that the `Decontaminated Weighted` estimator is unbiased after decontamination – a reassuring result. We also note that our decontaminated estimators reduce the variance on the CF estimates, as indicated by the error bars in Figure 4.5.

4.5.2 Realistic Example: Optimistic Case

Now we consider a more realistic scenario: a true galaxy sample with $0.7 \leq z \leq 1.0$, with three redshift bins ($0.7 \leq z < 0.8$, $0.8 \leq z < 0.9$, $0.9 \leq z < 1.0$) for the tomographic clustering analysis. As before, we query the galaxies in nine 10x10 deg^2 patches along Dec = 0, and model their photo-zs assuming Gaussian PDFs for all the galaxies with $\sigma_z = 0.03(1+z)$ as discussed at the beginning of Section 4.5; all patches have a similar number of galaxies (1080K-1147K) and face similar contamination (23-26%, 44-46%, 19-23% in the three tomographic bins, respectively). Note that our chosen bins are realistic, as a tomographic analysis for 10 redshift bins with $\Delta z = 0.1$ is currently planned for dark energy science studies with LSST (The LSST Dark Energy Science Collaboration et al., 2018); our treatment of photo-zs, however, is optimistic in the assumption of Gaussian photo-z PDFs.

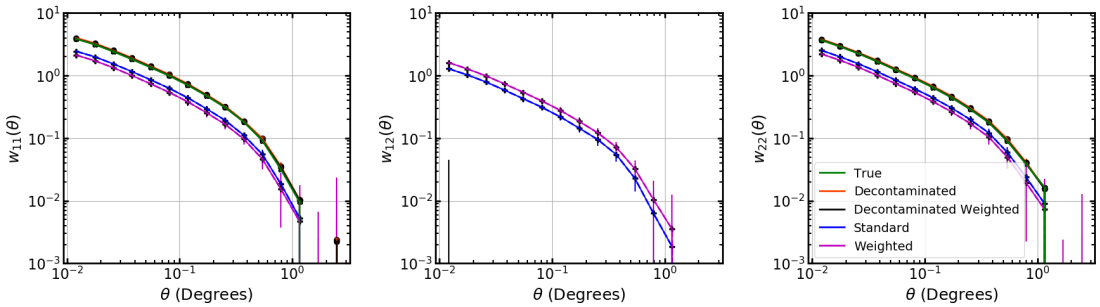

Figure 4.5: Correlation functions estimates and the estimator variance in the toy galaxy sample with only two redshift bins (presented in Section 4.5.1). We see that just as `Decontamination` (red) recovers the truth (green) using the correlations on the contaminated subsamples (blue), the `Decontaminated Weighted` estimator (black) recovers the truth from the `Weighted` correlations on the entire observed sample (magenta), without needing to divide the observed sample into subsamples. We also note that the decontaminated estimators reduce the variance on the CF estimates, as indicated by the error bars here.

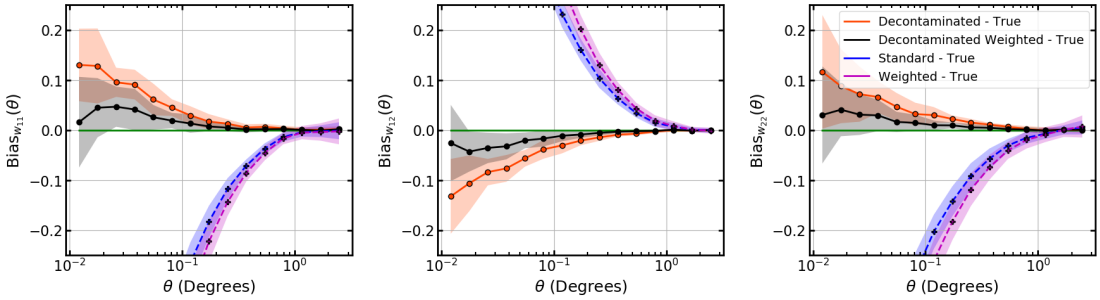

Figure 4.6: Bias in correlation functions for the toy galaxy sample of Section 4.5.1, with 1σ uncertainties in each estimator indicated with the shaded regions. We see that the `Decontaminated Weighted` estimator (black) leads to a bias smaller than that from the `Decontaminated` estimator (red); the green line indicates zero bias.

Figure 4.7 shows the distributions of the true redshifts and the photo-zs using one of the patches (with 1,095,404 galaxies, and 24%, 45% and 22% contamination in the three redshift bins, respectively). We note that the middle bin sees the largest and most realistic contamination – the case that will be true for most of the LSST bins, hence making this example a relevant one. Note that the bin edges see the impacts of artificially having contamination from only one side.

Figure 4.8 shows the distribution of the classification probabilities for all the galaxies. Again we note that given the large contamination rates for the middle bin, the classification probabilities are far from unity, indicating that no observed galaxy has a very high probability to be in any target bin. As before, we calculate the various correlations for each of the nine patches and estimate the mean and the variance across the calculations. Figure 4.9 illustrates our results, showing only the estimator bias for brevity, where we see that the `Decontaminated Weighted` estimator leads to a bias that is comparable to that using the `Decontaminated` estimator, both of

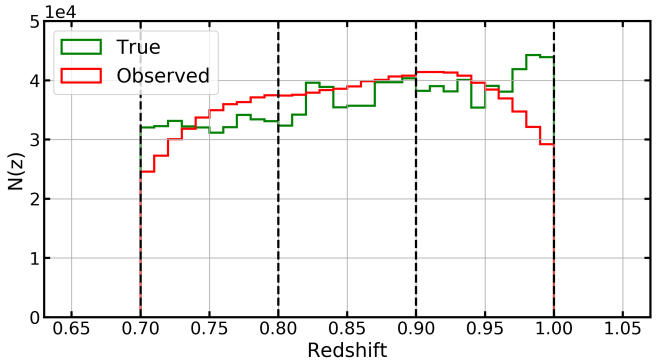

Figure 4.7: True and observed redshift histograms for the mock galaxy sample of Section 4.5.2, with bin edges shown using the vertical dashed lines. We see that the photo-z uncertainties lead to a smearing of the redshift information, while the truncation of the edge-bins makes the $N(z)$ biased near the outermost edges.

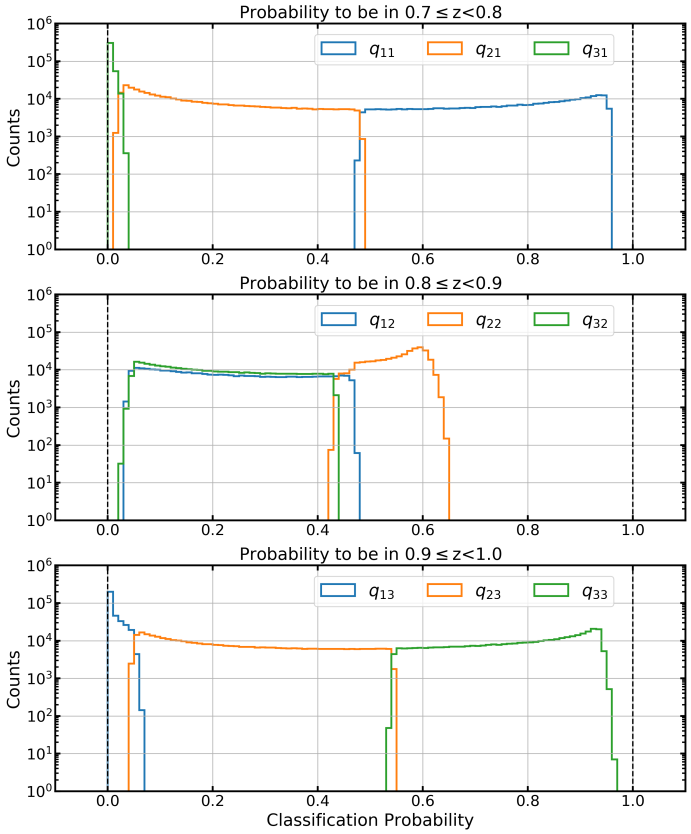

Figure 4.8: Distribution of the classification probabilities to be in the three target redshift bins for the mock galaxy sample of Section 4.5.2. The middle bin sees the largest contamination and therefore has no objects that have a very high probability to be in any target bin.

which are smaller than from those without decontamination. We note that the `Decontaminated` estimator performs similar to `Decontaminated Weighted`, potentially due to the correlation functions in the three redshift bins being similar. We also note that there is a weak residual

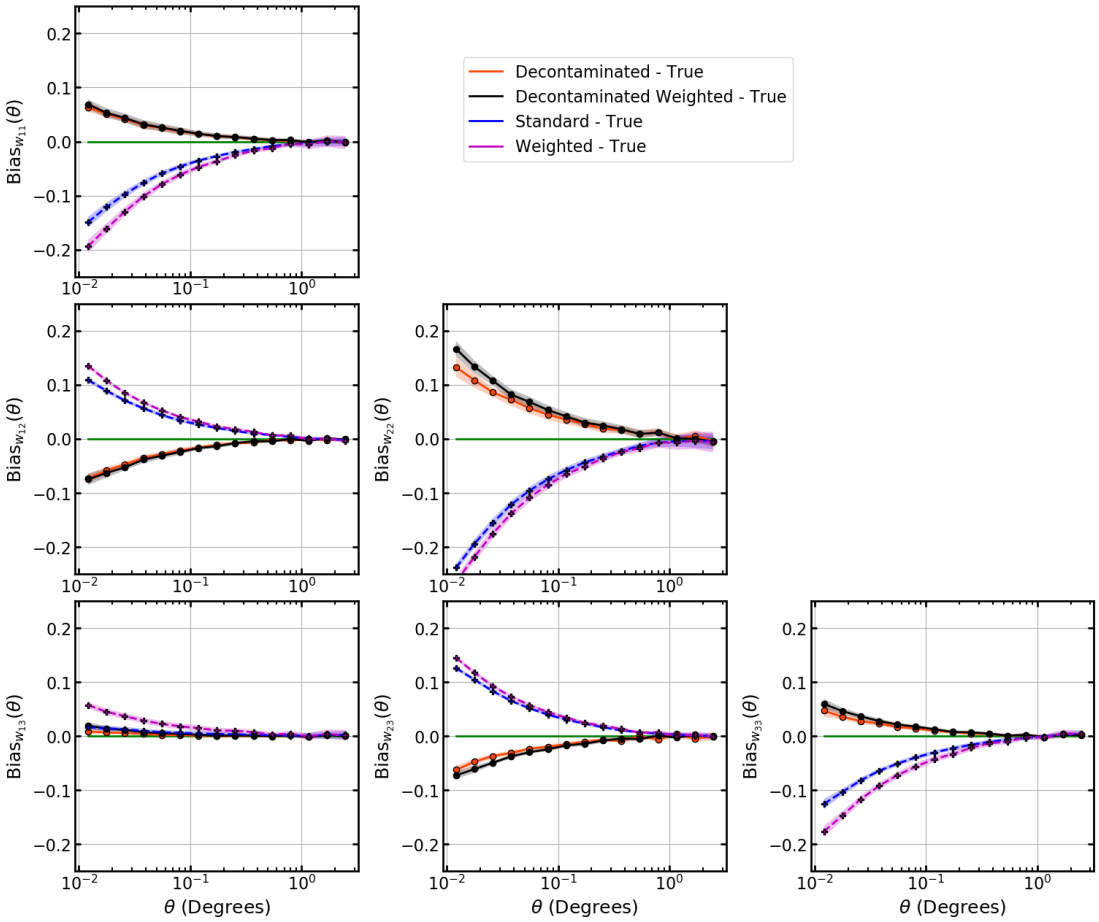

Figure 4.9: Bias in the correlation functions in the three sample case of Section 4.5.2, with 1σ uncertainties in each estimator indicated with the shaded regions. We see that as in the toy example in Section 4.5.1, just as `Decontamination` (red) reduces the bias using the correlations on the contaminated subsamples (blue), the `Decontaminated Weighted` estimator (black) reduces the bias from the `Weighted` correlations on the entire observed sample (magenta), without needing to divide the observed sample into subsamples; the green line indicates zero bias.

bias in the decontaminated estimates, which is likely caused by our simple debiasing of the classification probabilities.

As a more comprehensive metric for comparing the various estimators, we consider the covariances in correlation functions across the three redshift bins for an example θ-bin. Specifically, given that we have access to the truth here, we first calculate the covariances in the estimators without accounting for the LSS sample variance – this we term as the "estimator covariance" and calculate as $\left\langle \{\widehat{w}_i(\theta_k) - w_i^{\rm true}(\theta_k)\} \left\{\widehat{w}_j(\theta_k) - w_j^{\rm true}(\theta_k)\right\}\right\rangle$ where i, j run over all the correlations (both auto and cross) and the expectation value is over all the realizations[4];

[4]We calculate covariances using the `numpy.cov` function, which automatically subtracts off the mean for each variable (which, in this case, is the residual bias for each estimator); the default parameters of the function also account

note here that the diagonal of this covariance matrix is the estimator variance used to generate uncertainties shown in Figures 4.5-4.6 and Figure 4.9. We show the estimator covariances for the mock galaxy sample considered here in Figure 4.10, where we see that without decontamination, the covariances are large, as expected given the strong mixing of the samples. Both decontaminated estimators effectively reduce the covariances, with `Decontaminated Weighted` outperforming `Decontaminated`.

Then we consider the covariances accounting for the LSS sample variance – this we term as the "full covariance" and calculate as $\langle \{\widehat{w}_i(\theta_k) - \langle \widehat{w}_i(\theta_k)\rangle\}\{\widehat{w}_j(\theta_k) - \langle \widehat{w}_j(\theta_k)\rangle\}\rangle$ where i, j again run over all the correlations and the expectation value is over all the realizations; these are shown in Figure 4.11. We see that without decontamination, the clustering information is smeared across the CF-space and is much in contrast from the true covariances. However, both of our decontaminated estimators are able to approximate the true covariances effectively, hence achieving their purpose of correcting for sample contamination. We also note here that decontamination does not simply diagonalize the covariance matrices but instead reduces off-diagonal elements appropriately; diagonalization would not account for true covariances that exist between auto- and cross- CFs for neighboring bins due to shared LSS. Finally, comparing with Figure 4.10, we note that LSS sample variance largely dominates over the estimator variance for the 10x10 patches considered here – a reassuring result; a comparison between the two sources of variance for larger effective survey area is left for future work.

4.5.3 Realistic Example: Pessimistic Case

Now we consider a more pessimistic scenario for the true galaxy sample of Section 4.5.2: instead of having all the galaxies with well-behaved Gaussian photo-z PDFs, we assign half of the galaxies bimodal photo-z PDFs – a scenario where standard $N(z)$ forward modeling might be problematic. Specifically, the Gaussian photo-z PDFs are constructed as described above: by drawing a random number from a Gaussian of width $\sigma = 0.03(1 + z_{\rm true})$, with the observed photo-z PDF being a Gaussian centered at $z_{\rm draw}$ and with width $\sigma = 0.03(1 + z_{\rm draw})$. In contrast, the bimodal photo-z PDFs are constructed with one mode at the true redshift and another randomly chosen to be $\pm$ 0.13 away (while ensuring the second mode remains in the redshift range of 0.7-1.0); 0.13 separation mimics a degeneracy arising from Balmer vs. 4000Å decrement at $\sim$7% separations in $1 + z$. This treatment leads to slightly higher contamination rates: 39-42%, 54-57%, 33-36% in the three tomographic bins, respectively. To illustrate the

for the lost degree-of-freedom (i.e., using $N - 1$ when calculating the average, where N is the number of realizations).

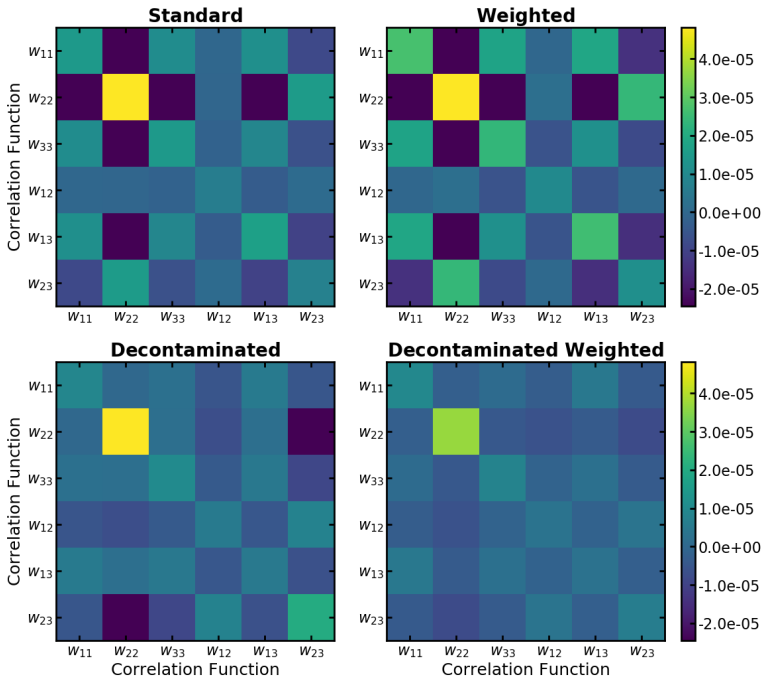

Figure 4.10: Estimator covariances across redshift bins for the case with three target redshift bins of Section 4.5.2 for an example theta-bin (with $\theta = 0.79$ degrees as nominal center of the bin in $\log(\theta)$); these probe the covariances in the estimators without accounting for LSS sample variance. Here, $w_{\alpha\beta}$ refers to the CF between galaxies in redshift bins α and β, and as noted in the text, we estimate the estimator covariance as $\left\langle \left\{ \widehat{w}_i(\theta_k) - w_i^{\rm true}(\theta_k) \right\} \left\{ \widehat{w}_j(\theta_k) - w_j^{\rm true}(\theta_k) \right\} \right\rangle$ for each estimator, where i, j run over all the correlations (both auto and cross) and the expectation value is over all the realizations. Note that this is not sensitive to sample variance since the true CF for each realization is subtracted from the observed CF for that realization. The left column shows estimator covariances in contaminated samples constructed using photo-z point estimates before (top) and after (bottom) decontamination, while the right column shows the estimator covariances in CF estimates using our `Weighted` estimator before (top) and after (left) decontamination. We see that our new decontaminated estimators reduce the covariances, with `Decontaminated Weighted` outperforming `Decontaminated`.

difference between the two cases more explicitly, Figure 4.12 shows an example set of PDFs for the case of all-Gaussian PDFs vs. half-bimodal ones.

Figure 4.13 shows the distributions of the true redshifts and the photo-zs using one of the patches (with 1,095,404 galaxies as before, but now with 40%, 55% and 35% contamination in the three redshift bins, respectively). Comparing it to Figure 4.7, we see that the distribution is slightly more biased, although the middle redshift bin sees a comparable observed redshift distribution; and as before, the bin edges see the impacts of artificially having contamination from only one side.

Figure 4.14 shows the classification probabilities for all the galaxies here; comparing it to Figure 4.8, we see that the classification probabilities are now more varied, with more objects in the edge-bins with larger classification probabilities due to the bimodality in some of the photo-z PDFs. As before, we calculate the various correlations for each of the nine

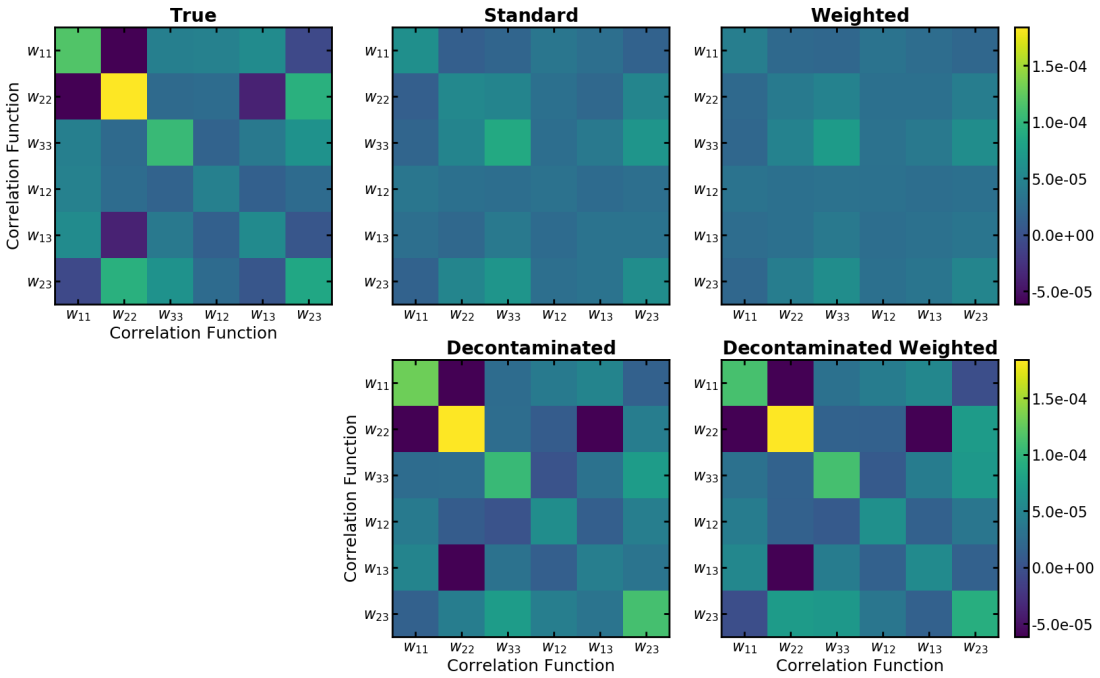

Figure 4.11: Full covariances across redshift bins for the case with three target redshift bins of Section 4.5.2 for an example theta-bin (with $\theta = 0.79$ degrees as the nominal center of the bin in $\log(\theta)$); these probe the covariances in the estimators while accounting for LSS sample variance. Here, $w_{\alpha\beta}$ refers to the CF between galaxies in redshift bins α and β, and e.g., w_{11} and w_{12} are correlated since LSS at the boundary of the two bins makes w_{12} non-zero and contributes to w_{11}. As noted in the text, we calculate these full covariances as $\langle\{\widehat{w}_i(\theta_k) - \langle\widehat{w}_i(\theta_k)\rangle\}\,\{\widehat{w}_j(\theta_k) - \langle\widehat{w}_j(\theta_k)\rangle\}\rangle$ for each estimator, where i, j again run over all the correlations and the expectation value is over all the realizations. The top left panel shows the true covariances across multiple realizations of the LSS, the middle column shows covariances in contaminated samples constructed using photo-z point estimates before (top) and after (bottom) decontamination, while the rightmost column shows the covariances in CF estimates using our `Weighted` estimator before (top) and after (left) decontamination. We see that our new decontaminated estimators approximate the true covariances, successfully accounting for sample contamination arising from photo-z uncertainties.

patches and estimate the mean CFs and the covariances. Figure 4.15 shows the residuals in the CF estimates, and we see that the decontaminated estimators are able to reduce the bias significantly. Figure 4.16 shows the estimator covariance matrices where we see that as in the all-Gaussian case, our decontaminated estimators lead to lower estimator covariances, with `Decontaminated Weighted` outperforming `Decontaminated` slightly more strongly than in Figure 4.10. Finally, Figure 4.17 shows the full covariance matrices. Here too, we see that as in Figure 4.11 for the all-Gaussian case, our decontaminated estimators approximate the true covariances more effectively with those without decontamination.

This completes the demonstration of our new estimators: they provide for a way to decontaminate correlations, while the `Weighted` estimator specifically allows using the full photo-z PDFs and full observed samples, in a framework that can be extended e.g., to minimize variance.

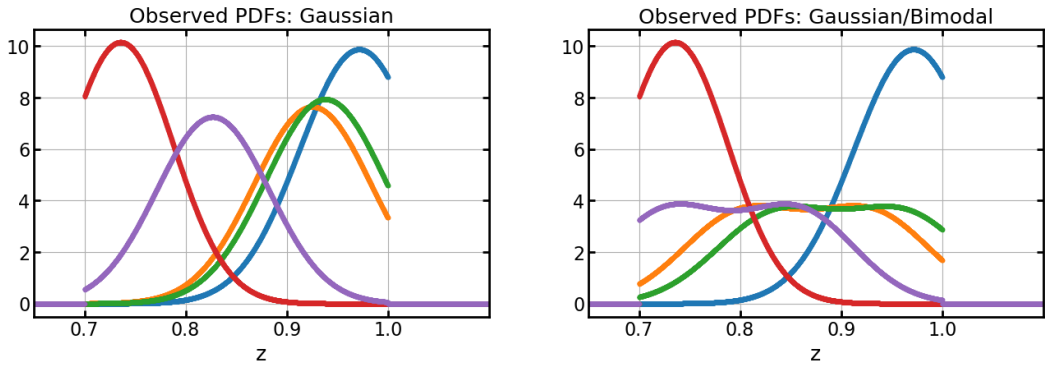

Figure 4.12: An example set of PDFs to compare the case of all-Gaussian PDFs of Section 4.5.2 vs. the case presented in Section 4.5.3 where half of the galaxies have bimodal PDFs. The left panel shows the observed photo-z PDFs for the case of all-Gaussian PDFs while the right panel shows them for the case where half of the galaxies have bimodal PDFs. The colors correspond to the same objects across the panels.

Figure 4.13: True and observed redshift histograms for the mock galaxy sample of Section 4.5.3. As in Figure 4.7, the bin edges shown using the vertical dashed lines. We see that as in Figure 4.7, the photo-z uncertainties lead to a smearing of the redshift information, while the truncation of the edge-bins makes the $N(z)$ biased near the outermost edges.

4.6 Discussion

We have presented a formalism to estimate the ACFs in the presence of sample contamination arising from photo-z uncertainties. We achieve this by a two-fold process: using the information in the contaminated correlations and utilizing the probabilistic information available via each galaxy's photo-z PDF in each target redshift bin. As mentioned in Section 4.1, our method avoids forward modeling the contaminated ACFs based on estimated $N(z)$, which is the standard way to handle the photo-z contamination for cosmological analyses. We note, however, that forward modeling is effective if the contamination can be modeled effectively; a full investigation of measurements using our method vs. those using forward modeling is left for future work. We also note that the BAO signal is washed out by projection and hence its measurement should benefit from our approach.

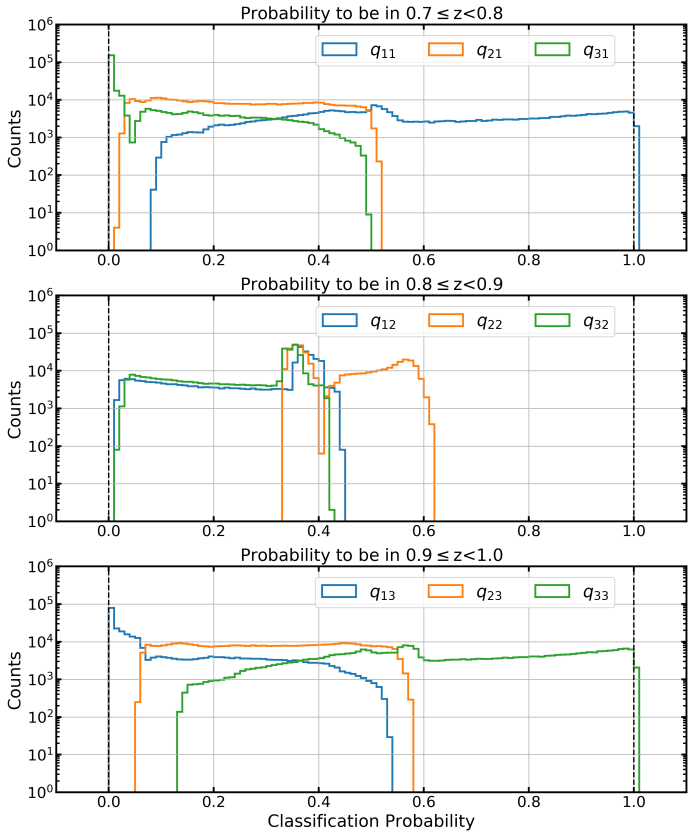

Figure 4.14: Distribution of the classification probabilities to be in the three target redshift bins for the mock galaxy sample of Section 4.5.3. As in Figure 4.8, the middle bin sees the largest contamination and therefore has no objects that have a very high probability to be in any target bin.

Our estimators are distinct from previous work employing weighted correlation functions, specifically on three accounts: 1) our weighted estimator considers all galaxies in the entire observed sample as a part of every photo-z bin, 2) to our knowledge, there is no literature on the usage of a decontamination matrix to correct for correlation function contamination, and our `Decontaminated Weighted` estimator presents a novel way to decontaminate marked correlation functions, and 3) we weight only the data, and not the randoms. As far as we are aware, the only other estimator in the literature that uses weights that are dependent on a galaxy's photo-z PDF in a galaxy clustering analysis is Asorey et al. (2016) but they employ a threshold to determine whether a galaxy contributes to a given redshift bin and do not allow contributions from a single galaxy to more than one bin. In a further comparison with our work, for instance, Ross et al. (2017) employ weights to account for photo-z uncertainty by weighting both the data and random galaxies in the target subsamples by inverse-variance weights. Blake et al. (2019) also weight both the data and random galaxies to increase the precision with which they can measure the BAO by accounting for the dependency on the environment of

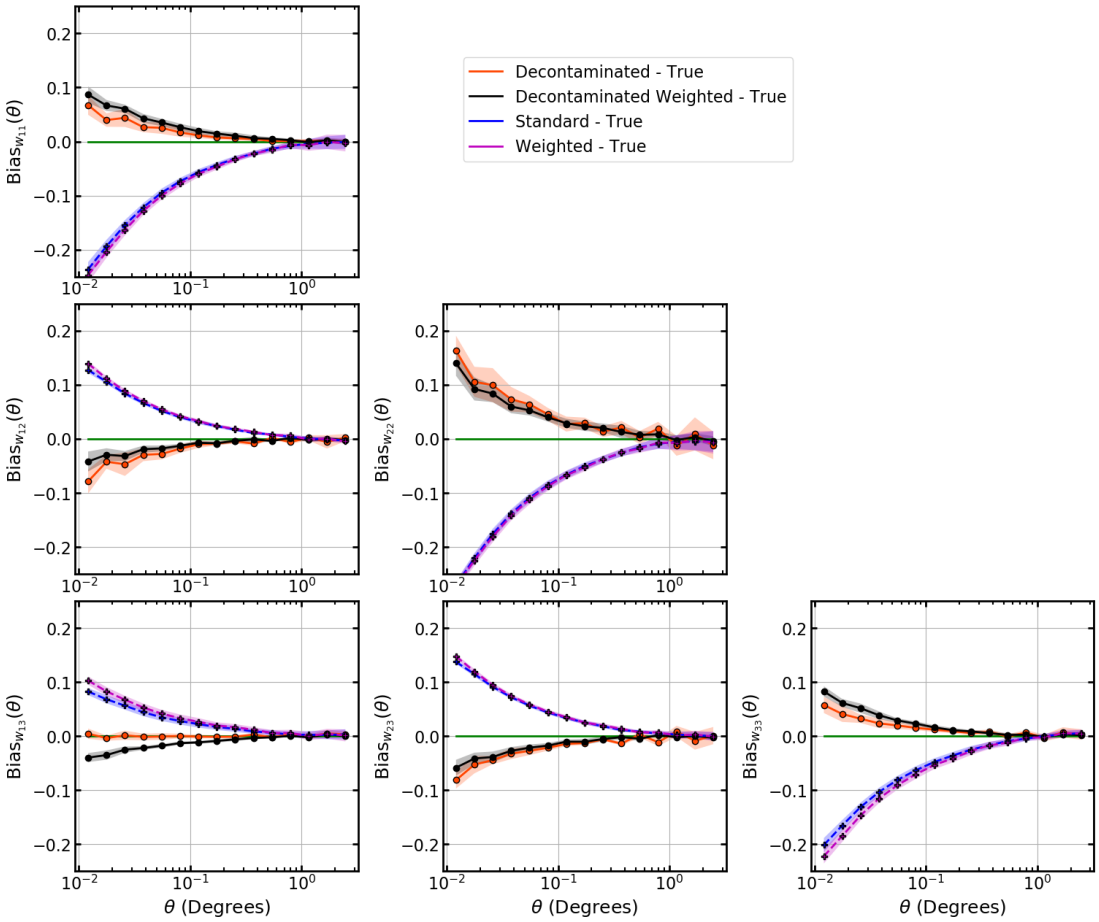

Figure 4.15: Bias in the correlation functions in the three sample case of Section 4.5.3. As in Figure 4.9, the 1σ uncertainties in each estimator are indicated with the shaded regions. We see that as for the all-Gaussian photo-z PDFs case, both decontaminated estimators significantly reduce the bias and lead to estimates closer to the truth.

the measured signal. In somewhat of a contrast, Zhu et al. (2015) use both weighted data and random pairs, and unweighted random pairs for optimized BAO measurements, while Morrison & Hildebrandt (2015) employ weighted randoms to account for mitigating survey systematics. Percival & Bianchi (2017), on the other hand, upweight only their data (data-data, data-random pairs, but not the random-random pairs) for 3D BAO measurements when the spectroscopic data is available only for a subset of the angular sample while Bianchi & Percival (2017) employ a similar weighting to account for missing information.

Since this work introduces a new estimator, we note various avenues for further development. For the 2D case, we can optimize the estimator to be minimum variance by introducing an additional parameter for each pair of galaxies, i.e., $\mathtt{w}^{\alpha\beta}_{ij,\mathrm{opt}} = \Upsilon_{ij}(q,k)\mathtt{w}^{\alpha\beta}_{ij}$, where $\Upsilon_{ij}(q,k)$ are the optimization parameters that minimize the variance of the estimator for each bin k.

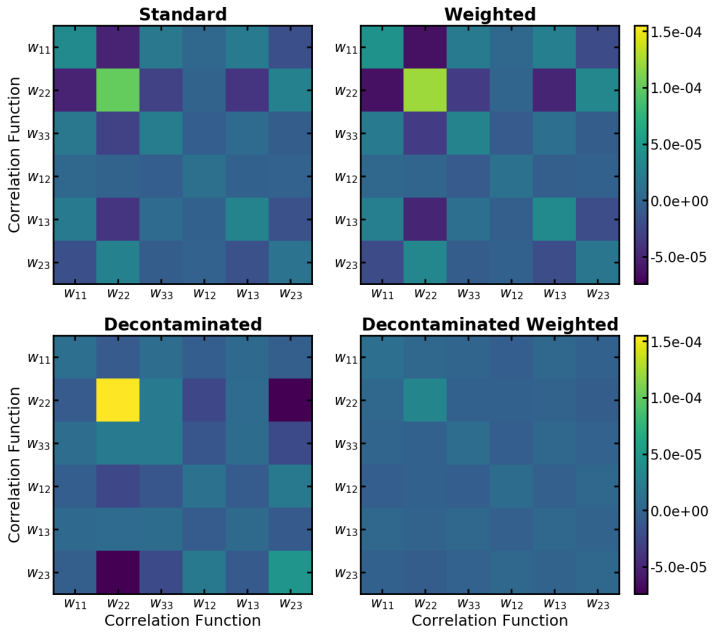

Figure 4.16: Estimator covariances across redshift bins for the case of Section 4.5.3 for the same example theta-bin as in Figure 4.10. As in Figure 4.10, the left column shows estimator covariances in contaminated samples constructed using photo-z point estimates before (top) and after (bottom) decontamination, while the right column shows the estimator covariances in CF estimates using our `Weighted` estimator before (top) and after (left) decontamination. We see that our new decontaminated estimators reduce the covariances, with `Decontaminated Weighted` outperforming `Decontaminated`.

We note again that the `Decontaminated` estimator presented in the text is in fact a special case of the `Decontaminated Weighted` estimator, with the weights set to 1 when the probability is high enough to place an object in a given subsample and 0 otherwise and then with average contamination fractions used to decontaminate instead of the classification probabilities. It is indeed surprising that the `Decontaminated` estimator performs nearly as well as our `Decontaminated` probability-`Weighted` estimator; this implies either a broad range of optimal weights or, more likely, that the optimal weights lie somewhere between these two simplistic approaches. Optimization of the weights will be an important aspect of applying the new estimator. Furthermore, since we have introduced general pair weights, we can incorporate Bayesian priors on the correlation functions, based on current measurements, or when measuring correlation functions for different galaxy types, as then, we can incorporate priors that are dependent on the separations, e.g., accounting for one galaxy sample clustering strongly on smaller scales. This will call for an in-depth analysis of the covariance matrices for the various correlation functions. Also, we can extend the weighting scheme to harmonic space, where it will be relevant for a tomographic analysis for LSST (Awan et al., in prep).

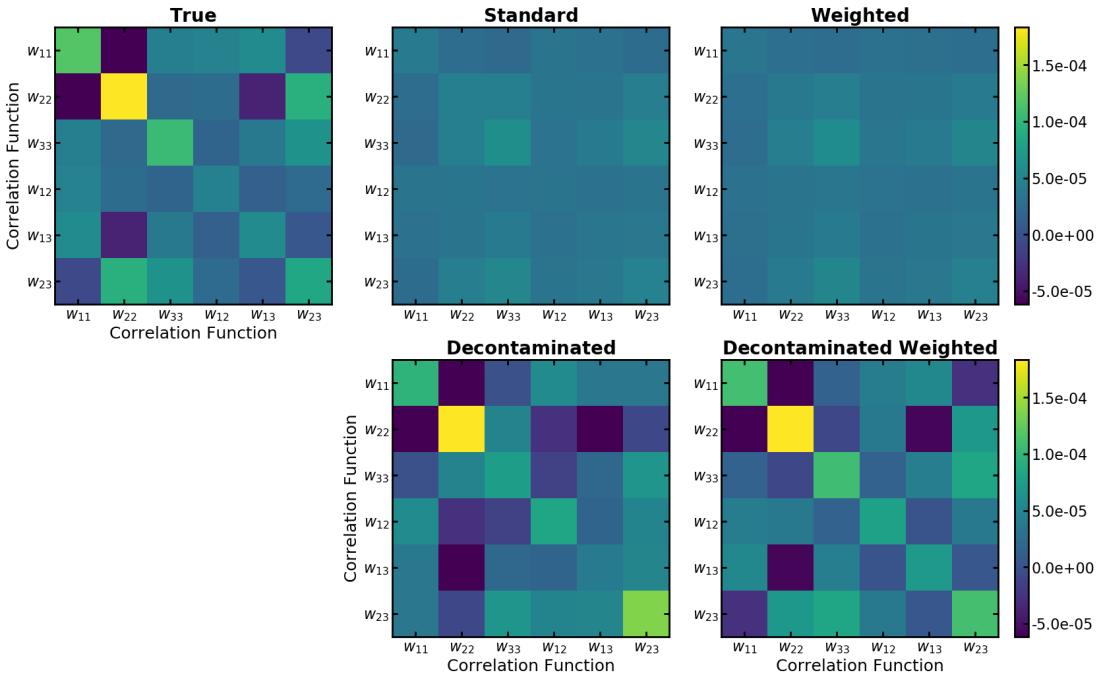

Figure 4.17: Full covariances across redshift bins for the case of Section 4.5.3 for the same example theta-bin as in Figure 4.11. As in Figure 4.11, the top left panel shows the true covariances across multiple realizations of the LSS, the middle column shows covariances in contaminated samples constructed using photo-z point estimates before (top) and after (bottom) decontamination, while the rightmost column shows the covariances in CF estimates using our `Weighted` estimator before (top) and after (left) decontamination. We see that our new decontaminated estimators approximate the true covariances, successfully accounting for sample contamination arising from photo-z uncertainties.

We also note that our method can handle other kinds of contamination, e.g., star-galaxy contamination, where probabilistic models for whether an object is a star or a galaxy can inform the weights for each object in our observed sample; this is possible since neither decontamination nor the pair weights have an explicit redshift dependence, hence allowing decontaminating and weighting any types . Finally, we can also extend the 2D formulation to 3D, where it will be relevant for HETDEX (Hill et al., 2008), Euclid and WFIRST, as they face emission line contaminants, as well as LSST where the projected correlation function will be measurable (without tomographic binning). Note that for the 3D case in real space, we must treat the random catalogs more carefully than in 2D; in the 2D case considered here, we have not made a distinction between random catalogs for the different samples as they are spatially overlapping with the same selection function – a case that does not hold for 3D.

4.7 Conclusions

Cosmology is entering a data-driven era, with several upcoming galaxy surveys opening gateways for huge galaxy catalogs. Given the increased statistical power of our datasets, we face imminent challenges, including the need to account for systematic uncertainties that dominate the uncertainty budget on our measurements. In this paper, we have studied the treatment of contamination arising from photo-z uncertainties when measuring the two-point angular correlation functions. We first introduced a simple formalism: decontamination that uses the correlations in contaminated subsamples to estimate the true correlations. We then introduced a new estimator that accounts for the full photo-z PDF of each galaxy to estimate the true correlations, allowing each galaxy to contribute to all bins (or samples) based on their probabilities. We demonstrated the effectiveness of our method in recovering true CFs and covariance matrix on both a toy example and a realistic scenario that is scaleable for surveys like LSST. We also note that our estimator can correct for contamination when measuring correlation functions of multiple galaxy populations, rather than photo-z bins, alongside other kinds of contamination.

We emphasize the need for more data-driven tools in order to truly utilize the statistical power of the large datasets. Here we have presented an estimator that incorporates the available probabilistic information to reduce the bias and variance in the measured correlation functions; this represents a step in the direction of reducing biases and uncertainties in the measurement of cosmological parameters from upcoming surveys.

4.8 Acknowledgements

We thank David Alonso, Nelson Padilla and Javier Sánchez for their helpful feedback. H. Awan also thanks Kartheik Iyer and Willow Kion-Crosby for insightful discussions through the various stages of this work. H. Awan has been supported by the Rutgers Discovery Informatics Institute (RDI^2) Fellowship of Excellence in Computational and Data Science(AY 2017-2020) and Rutgers University & Bevier Dissertation Completion Fellowship (AY 2019-2020). This work has used resources from RDI^2, which are supported by Rutgers and the State of New Jersey; specifically, our analysis used the `Caliburn` supercomputer (Villalobos et al., 2018). The authors also acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, the State University of New Jersey for providing access to the `Amarel` cluster and associated research computing resources that have contributed to our work. H. Awan also thanks the LSSTC Data Science Fellowship Program, which is funded by LSSTC, NSF Cybertraining Grant #1829740, the Brinson Foundation, and the Moore Foundation, as participation in the program

has benefited this work. This research was also supported by the Department of Energy (grants DE-SC0011636 and DE-SC0010008).

Appendices

4.A `Decontaminated` Estimator: Decontamination, Bias and Variance

4.A.1 Decontamination Derivation

Here, we re-derive the decontamination equation (Equation 4.11) using the definition of angular correlation function. We start with Equation 4.1, rewriting it as

$$dP_{\alpha\beta}(\theta_k) = \eta^{\rm pair}_{\alpha\beta}\left[1+w_{\alpha\beta}(\theta_k)\right]d\Omega_\alpha d\Omega_\beta = \mathcal{N}_{\alpha\beta}\left[1+w_{\alpha\beta}(\theta_k)\right]\frac{d\Omega_\alpha}{V_\alpha}\frac{d\Omega_\beta}{V_\beta} \tag{4.19}$$

where $\eta^{\rm pair}_{\alpha\beta}$ is the observed sky density of Type-$\alpha\beta$ pairs of galaxies while $\mathcal{N}_{\alpha\beta}$ is the observed number of Type-$\alpha\beta$ pairs. Assuming that we work with large surveys such that the integral constraint is nearly zero, we have $\mathcal{N}_{\alpha\beta} \to \langle\mathcal{N}_{\alpha\beta}\rangle$, hence the simplification in the last line in the equation above. Since we consider samples in the same volume, $V_\alpha = V_\beta = V$ and $d\Omega_\alpha = d\Omega_\beta = d\Omega$. Therefore, for the `Standard` estimator, for the case where we have the correlations measured in the contaminated subsamples, we have

$$dP_{\alpha\beta}(\theta_k) = \mathcal{N}_{\alpha\beta,\rm obs}\left[1+w^{\rm obs}_{\alpha\beta}(\theta_k)\right]\frac{d\Omega}{V}\frac{d\Omega}{V} = \sum_{\gamma,\delta}\mathcal{N}^{\gamma\delta,\rm true}_{\alpha\beta,\rm obs}[1+w^{\rm true}_{\gamma,\delta}(\theta_k)]\frac{d\Omega}{V}\frac{d\Omega}{V} \tag{4.20}$$

where $w^{\rm obs}_{\alpha\beta}(\theta_k)$ is the biased correlation function, measured using contaminated samples. Expanding the sum on the right hand side, we have

$$\begin{aligned}\mathcal{N}^{\rm tot}_{\alpha\beta,\rm obs}\left[1+w^{\rm obs}_{\alpha\beta}(\theta_k)\right] = {}& \mathcal{N}^{11,\rm true}_{\alpha\beta,\rm obs}\left[1+w^{\rm true}_{11}(\theta_k)\right] + \mathcal{N}^{12,\rm true}_{\alpha\beta,\rm obs}\left[1+w^{\rm true}_{12}(\theta_k)\right] \\ & + \mathcal{N}^{21,\rm true}_{\alpha\beta,\rm obs}\left[1+w^{\rm true}_{21}(\theta_k)\right] + \mathcal{N}^{22,\rm true}_{\alpha\beta,\rm obs}\left[1+w^{\rm true}_{22}(\theta_k)\right]\end{aligned} \tag{4.21}$$

Since we have

$$\frac{\mathcal{N}^{\gamma\delta,\rm true}_{\alpha\beta,\rm obs}}{\mathcal{N}^{\rm tot}_{\alpha\beta,\rm obs}} = f_{\alpha\gamma}f_{\beta\delta} \tag{4.22}$$

$$\begin{aligned}\Rightarrow \left[1+w^{\rm obs}_{\alpha\beta}(\theta_k)\right] = {}& f_{\alpha1}f_{\beta1}\left[1+w^{\rm true}_{11}(\theta_k)\right] + \left\{f_{\alpha1}f_{\beta2}+f_{\alpha2}f_{\beta1}\right\}\left[1+w^{\rm true}_{12}(\theta_k)\right] \\ & + f_{\alpha2}f_{\beta2}\left[1+w^{\rm true}_{22}(\theta_k)\right]\end{aligned} \tag{4.23}$$

Therefore, for $\alpha, \beta = 1, 2$, Equation 4.23 becomes

$$\begin{aligned}\left[1 + w_{12}^{\text{obs}}(\theta_k)\right] = f_{11}f_{21}\left[1 + w_{11}^{\text{true}}(\theta_k)\right] + \left\{f_{11}f_{22} + f_{12}f_{21}\right\}\left[1 + w_{12}^{\text{true}}(\theta_k)\right] \\ + f_{12}f_{22}\left[1 + w_{22}^{\text{true}}(\theta_k)\right]\end{aligned} \tag{4.24}$$

Now, since

$$f_{11}f_{21} + \left\{f_{11}f_{22} + f_{12}f_{21}\right\} + f_{12}f_{22} = f_{11}\left[f_{21} + f_{22}\right] + f_{12}\left[f_{21} + f_{22}\right] = 1, \tag{4.25}$$

we have

$$w_{12}^{\text{obs}}(\theta_k) = f_{11}f_{21}w_{11}^{\text{true}}(\theta_k) + \left\{f_{11}f_{22} + f_{12}f_{21}\right\}w_{12}^{\text{true}}(\theta_k) + f_{12}f_{22}w_{22}^{\text{true}}(\theta_k) \tag{4.26}$$

which agrees with Equation 4.11. Similar results follow for (α, β) = (1,1), (2, 2).

4.A.2 Estimator Bias

We expect that the `Decontaminated` estimators are unbiased given their construction (i.e., Equation 4.10). However, for brevity, we formally show that they are indeed unbiased. By definition, an unbiased estimator is such that

$$\langle \widehat{w} \rangle = w_{\text{true}} \tag{4.27}$$

where the expectation value is over many realizations of the survey. Then, using Equations 4.11 and 4.12, we have

$$\begin{aligned}\left\langle \left[\widehat{w}_{AA}(\theta_k)\ \ \widehat{w}_{AB}(\theta_k)\ \ \widehat{w}_{BB}(\theta_k)\right]^T \right\rangle &= \left\langle [D_{\text{S}}]^{-1}\left[w_{AA}^{\text{obs}}(\theta_k)\ \ w_{AB}^{\text{obs}}(\theta_k)\ \ w_{BB}^{\text{obs}}(\theta_k)\right]^T \right\rangle \\ &= [D_{\text{S}}]^{-1}[D_{\text{S}}]\left[w_{AA}^{\text{true}}(\theta_k)\ \ w_{AB}^{\text{true}}(\theta_k)\ \ w_{BB}^{\text{true}}(\theta_k)\right]^T \\ &= \left[w_{AA}^{\text{true}}(\theta_k)\ \ w_{AB}^{\text{true}}(\theta_k)\ \ w_{BB}^{\text{true}}(\theta_k)\right]^T\end{aligned} \tag{4.28}$$

where the second equality follows by substituting Equation 4.11. Hence, the `Decontaminated` estimators are unbiased. We note here that $[D_{\text{S}}]$ in Equation 4.12 is effectively a decontamination matrix: it removes the contamination from the biased estimates, $w_{\alpha\beta}^{\text{obs}}$, in the presence of sample contamination. A similar argument follows for the case where we have M target samples, using Equation 4.108. We also note that Equation 4.28 is valid only when $f_{\alpha\beta}$ are accurate averages of the classification probabilities.

4.A.3 Estimator Variance

As for the variance of the `Decontaminated` estimators, we can calculate it by using the variance in our observed correlations. That is, given Equation 4.12, we have

$$\left[\sigma^2_{\widehat{w}_{AA}}(\theta_k)\ \ \sigma^2_{\widehat{w}_{AB}}(\theta_k)\ \ \sigma^2_{\widehat{w}_{BB}}(\theta_k)\right]^T = \left\{[D_{\rm S}]^{-1}\right\}^2_{ij}\left[\sigma^2_{w^{\rm obs}_{AA}}(\theta_k)\ \ \sigma^2_{w^{\rm obs}_{AB}}(\theta_k)\ \ \sigma^2_{w^{\rm obs}_{BB}}(\theta_k)\right]^T \tag{4.29}$$

where $\left\{[D_{\rm S}]^{-1}\right\}^2_{ij}$ denotes that matrix resulting from squaring each individual coefficient in the matrix $[D_{\rm S}]^{-1}$. We also note that the above derivation assumes no covariance between the observed correlations (i.e., $w^{\rm obs}_{\alpha\beta}$), which is incorrect for the case of neighboring redshift bin given the shared LSS between them; this is discussed in more detail when we discuss the covariance matrices in Section 4.5.2. To consider the covariance matrix for the `Decontaminated` estimators, we start with Equation 4.12, which is reproduced here:

$$\left[\widehat{w}_{AA}(\theta_k)\ \ \widehat{w}_{AB}(\theta_k)\ \ \widehat{w}_{BB}(\theta_k)\right]^T = [D_{\rm S}]^{-1}\left[w^{\rm obs}_{AA}(\theta_k)\ \ w^{\rm obs}_{AB}(\theta_k)\ \ w^{\rm obs}_{BB}(\theta_k)\right]^T \tag{4.30}$$

Given Equation 4.28, we therefore have

$$\left\langle\left[\widehat{w}_{AA}(\theta_k)\ \ \widehat{w}_{AB}(\theta_k)\ \ \widehat{w}_{BB}(\theta_k)\right]^T\right\rangle = [D_{\rm S}]^{-1}\left\langle\left[w^{\rm obs}_{AA}(\theta_k)\ \ w^{\rm obs}_{AB}(\theta_k)\ \ w^{\rm obs}_{BB}(\theta_k)\right]^T\right\rangle \tag{4.31}$$

where we assume that $[D_{\rm S}]$ is constant across the samples over which the expectation value is calculated. Now, using the above equations, we can write the variations in the estimators from their expectation value ($\equiv \Delta w \equiv w - \langle w\rangle$) as

$$\left[\Delta\widehat{w}_{AA}(\theta_k)\ \ \Delta\widehat{w}_{AB}(\theta_k)\ \ \Delta\widehat{w}_{BB}(\theta_k)\right]^T = [D_{\rm S}]^{-1}\left[\Delta w^{\rm obs}_{AA}(\theta_k)\ \ \Delta w^{\rm obs}_{AB}(\theta_k)\ \ \Delta w^{\rm obs}_{BB}(\theta_k)\right]^T \tag{4.32}$$

Now defining $C_{\widehat{w}}(\theta_k)$ as the covariance matrix for the `Decontaminated` estimators $\widehat{w}_{\alpha\beta}(\theta_k)$, we have

$$C_{\widehat{w}}(\theta_k) = \left\langle\left[\Delta\widehat{w}_{AA}(\theta_k)\ \ \Delta\widehat{w}_{AB}(\theta_k)\ \ \Delta\widehat{w}_{BB}(\theta_k)\right]^T\left[\Delta\widehat{w}_{AA}(\theta_k)\ \ \Delta\widehat{w}_{AB}(\theta_k)\ \ \Delta\widehat{w}_{BB}(\theta_k)\right]\right\rangle \tag{4.33}$$

Using Equation 4.32 and its transpose, we then have

$$\begin{aligned} C_{\widehat{w}}(\theta_k) &= \Big\langle [D_{\rm S}]^{-1} \begin{bmatrix} \Delta w^{\rm obs}_{AA}(\theta_k) & \Delta w^{\rm obs}_{AB}(\theta_k) & \Delta w^{\rm obs}_{BB}(\theta_k) \end{bmatrix}^T \cdot \\ &\qquad \begin{bmatrix} \Delta w^{\rm obs}_{AA}(\theta_k) & \Delta w^{\rm obs}_{AB}(\theta_k) & \Delta w^{\rm obs}_{BB}(\theta_k) \end{bmatrix} \left[[D_{\rm S}]^{-1} \right]^T \Big\rangle \\ &= [D_{\rm S}]^{-1} \Big\langle \begin{bmatrix} \Delta w^{\rm obs}_{AA}(\theta_k) & \Delta w^{\rm obs}_{AB}(\theta_k) & \Delta w^{\rm obs}_{BB}(\theta_k) \end{bmatrix}^T \cdot \\ &\qquad \begin{bmatrix} \Delta w^{\rm obs}_{AA}(\theta_k) & \Delta w^{\rm obs}_{AB}(\theta_k) & \Delta w^{\rm obs}_{BB}(\theta_k) \end{bmatrix} \Big\rangle \left[[D_{\rm S}]^{-1} \right]^T \\ &= [D_{\rm S}]^{-1} C_{w^{\rm obs}}(\theta_k) \left[[D_{\rm S}]^{-1} \right]^T \end{aligned} \tag{4.34}$$

where $C_{w^{\rm obs}}$ is covariance matrix for the observed correlations, $w^{\rm obs}_{\alpha\beta}$. Note that the second equality is valid only under the assumption that $[D_{\rm S}]$ is constant.

Both $C_{w^{\rm obs}}(\theta_k)$ and $C_{\widehat{w}}(\theta_k)$ can be determined by bootstrap, as done for the example considered in Section 4.5.2, with the estimated covariance matrices presented in Figures 4.11 and 4.17. We note that $C_{\widehat{w}}(\theta_k)$ may be calculated using $C_{w^{\rm obs}}(\theta_k)$ given Equation 4.34, assuming that $[D_{\rm S}]$ is constant across the bootstrapped samples. We also that one can construct covariance matrices for both $w^{\rm obs}$ and $\widehat{w}$ spanning all θ-bins via a block combination of the θ-dependent matrices presented here; these larger matrices are only block diagonal to the extent that individual CFs are uncorrelated between neighboring θ-bins. Finally, as a simple check of the expression in Equation 4.34, we note that if $C_{w^{\rm obs}}(\theta_k)$ is diagonal, i.e., there are no covariances in the observed correlations, Equation 4.34 leads to the variance in the `Decontaminated` estimators as given by Equation 4.29.

4.B Decontamination: From `Decontaminated` with Full Sample to Weighted

Here, we present the methodology to decontaminate the `Weighted` correlation function introduced in Equation 4.13, using the formalism introduced in 4.A.1. To develop intuition, we first extend the methodology in 4.A.1 to consider an unweighted full observed sample, followed by considering the weighted full sample.

4.B.1 `Decontaminated`: Full Sample

We extend the treatment in 4.A.1 to consider an unweighted full sample. Then, the analog of Equation 4.20 is

$$dP(\theta_k) = \mathcal{N}_{\rm tot_{obs}} \left[1 + w^{\rm full}(\theta_k)\right] \frac{d\Omega}{V} \frac{d\Omega}{V} = \sum_{\gamma,\delta} \mathcal{N}^{\gamma\delta,\rm true}_{\rm tot_{obs}} [1 + w^{\rm true}_{\gamma,\delta}(\theta_k)] \frac{d\Omega}{V} \frac{d\Omega}{V} \tag{4.35}$$

Note that we have dropped the α, β markers since there is only one correlation that can be measured for the unweighted full sample. Expanding the sum, we have

$$\begin{aligned}\mathcal{N}_{\mathrm{tot}_{\mathrm{obs}}}\left[1+w^{\mathrm{full}}(\theta_k)\right] = \mathcal{N}^{11,\mathrm{true}}_{\mathrm{tot}_{\mathrm{obs}}}\left[1+w^{\mathrm{true}}_{11}(\theta_k)\right] + \mathcal{N}^{12,\mathrm{true}}_{\mathrm{tot}_{\mathrm{obs}}}\left[1+w^{\mathrm{true}}_{12}(\theta_k)\right] \\ + \mathcal{N}^{21,\mathrm{true}}_{\mathrm{tot}_{\mathrm{obs}}}\left[1+w^{\mathrm{true}}_{21}(\theta_k)\right] + \mathcal{N}^{22,\mathrm{true}}_{\mathrm{tot}_{\mathrm{obs}}}\left[1+w^{\mathrm{true}}_{22}(\theta_k)\right]\end{aligned} \tag{4.36}$$

Now if we assume that our classification probabilities are unbiased, we can write

$$\sum_{i}^{N^{\gamma}_{\mathrm{tot}_{\mathrm{obs}}}} \sum_{j\neq i}^{N^{\delta}_{\mathrm{tot}_{\mathrm{obs}}}} \mathfrak{q}^{\gamma}_i \mathfrak{q}^{\delta}_j = \widehat{\mathcal{N}}^{\gamma\delta,\mathrm{true}}_{\mathrm{tot}_{\mathrm{obs}}} \tag{4.37}$$

Note that technically $N^{\gamma}_{\mathrm{tot}_{\mathrm{obs}}} = N^{\delta}_{\mathrm{tot}_{\mathrm{obs}}} = N_{\mathrm{tot}_{\mathrm{obs}}}$ but we keep γ, δ tags just to keep track of samples when reducing to `Decontaminated`. Now, simplifying the equation above, we have

$$\begin{aligned}\mathcal{N}_{\mathrm{tot}_{\mathrm{obs}}}\left[1+w^{\mathrm{full}}(\theta_k)\right] = &\sum_{i}^{N^{1}_{\mathrm{tot}_{\mathrm{obs}}}} \sum_{j\neq i}^{N^{1}_{\mathrm{tot}_{\mathrm{obs}}}} \mathfrak{q}^{1}_i \mathfrak{q}^{1}_j \left[1+w^{\mathrm{true}}_{11}(\theta_k)\right] \\ &+ \sum_{i}^{N^{1}_{\mathrm{tot}_{\mathrm{obs}}}} \sum_{j\neq i}^{N^{2}_{\mathrm{tot}_{\mathrm{obs}}}} \mathfrak{q}^{1}_i \mathfrak{q}^{2}_j \left[1+w^{\mathrm{true}}_{12}(\theta_k)\right] \\ &+ \sum_{i}^{N^{2}_{\mathrm{tot}_{\mathrm{obs}}}} \sum_{j\neq i}^{N^{1}_{\mathrm{tot}_{\mathrm{obs}}}} \mathfrak{q}^{2}_i \mathfrak{q}^{1}_j \left[1+w^{\mathrm{true}}_{21}(\theta_k)\right] \\ &+ \sum_{i}^{N^{2}_{\mathrm{tot}_{\mathrm{obs}}}} \sum_{j\neq i}^{N^{2}_{\mathrm{tot}_{\mathrm{obs}}}} \mathfrak{q}^{2}_i \mathfrak{q}^{2}_j \left[1+w^{\mathrm{true}}_{22}(\theta_k)\right]\end{aligned} \tag{4.38}$$

We now check what happens when we reduce the above equation to `Decontaminated`, i.e., we consider not the full sample but the target subsamples, while all the probabilities are

represented by their averages. Then, for $\alpha, \beta = 1, 2$, Equation 4.38 becomes

$$
\begin{aligned}
&N_{1,\text{obs}} N_{2,\text{obs}} \left[1 + w_{11}^{\text{obs}}(\theta_k)\right] \\
&= \left(\sum_i^{N_{1,\text{obs}}} \sum_{j \neq i}^{N_{2,\text{obs}}} \mathfrak{q}_i^1 \mathfrak{q}_j^1 \right) \left[1 + w_{11}^{\text{true}}(\theta_k)\right] + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j \neq i}^{N_{2,\text{obs}}} \mathfrak{q}_i^1 \mathfrak{q}_j^2 \right) \left[1 + w_{12}^{\text{true}}(\theta_k)\right] \\
&\quad + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j \neq i}^{N_{2,\text{obs}}} \mathfrak{q}_i^2 \mathfrak{q}_j^1 \right) \left[1 + w_{21}^{\text{true}}(\theta_k)\right] + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j \neq i}^{N_{2,\text{obs}}} \mathfrak{q}_i^2 \mathfrak{q}_j^2 \right) \left[1 + w_{22}^{\text{true}}(\theta_k)\right] \\
&= \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \mathfrak{q}_i^1 \mathfrak{q}_j^1 \right) \left[1 + w_{11}^{\text{true}}(\theta_k)\right] + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \mathfrak{q}_i^1 \mathfrak{q}_j^2 \right) \left[1 + w_{12}^{\text{true}}(\theta_k)\right] \\
&\quad + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \mathfrak{q}_i^2 \mathfrak{q}_j^1 \right) \left[1 + w_{21}^{\text{true}}(\theta_k)\right] + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \mathfrak{q}_i^2 \mathfrak{q}_j^2 \right) \left[1 + w_{22}^{\text{true}}(\theta_k)\right] \\
&\xrightarrow{\text{simplify } q\text{s}} \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} q_{i,11} q_{j,12} \right) \left[1 + w_{11}^{\text{true}}(\theta_k)\right] \\
&\qquad + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} q_{i,11} q_{j,22} \right) \left[1 + w_{12}^{\text{true}}(\theta_k)\right] \\
&\qquad\quad + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} q_{i,12} q_{j,21} \right) \left[1 + w_{21}^{\text{true}}(\theta_k)\right] \\
&\qquad\qquad + \left(\sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} q_{i,12} q_{j,22} \right) \left[1 + w_{22}^{\text{true}}(\theta_k)\right] \\
&\xrightarrow{q\text{s=}f\text{s}} \left(f_{11} f_{21} \sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \right) \left[1 + w_{11}^{\text{true}}(\theta_k)\right] + \left(f_{11} f_{22} \sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \right) \left[1 + w_{12}^{\text{true}}(\theta_k)\right] \\
&\quad + \left(f_{12} f_{21} \sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \right) \left[1 + w_{21}^{\text{true}}(\theta_k)\right] + \left(f_{12} f_{22} \sum_i^{N_{1,\text{obs}}} \sum_{j}^{N_{2,\text{obs}}} \right) \left[1 + w_{22}^{\text{true}}(\theta_k)\right] \\
&= f_{11} f_{21} N_{1,\text{obs}} N_{2,\text{obs}} \left[1 + w_{11}^{\text{true}}(\theta_k)\right] + f_{11} f_{22} N_{1,\text{obs}} N_{2,\text{obs}} \left[1 + w_{12}^{\text{true}}(\theta_k)\right] \\
&\quad + f_{12} f_{21} N_{1,\text{obs}} N_{2,\text{obs}} \left[1 + w_{21}^{\text{true}}(\theta_k)\right] + f_{12} f_{22} N_{1,\text{obs}} N_{2,\text{obs}} \left[1 + w_{22}^{\text{true}}(\theta_k)\right]
\end{aligned}
\tag{4.39}
$$

$$
\begin{aligned}
\Rightarrow \left[1 + w_{12}^{\text{obs}}(\theta_k)\right] &= f_{11} f_{21} \left[1 + w_{11}^{\text{true}}(\theta_k)\right] + \left\{f_{11} f_{22} + f_{12} f_{21}\right\} \left[1 + w_{12}^{\text{true}}(\theta_k)\right] \\
&\quad + f_{12} f_{22} \left[1 + w_{22}^{\text{true}}(\theta_k)\right]
\end{aligned}
\tag{4.40}
$$

which agrees with Equation 4.26. Similar results follow for $(\alpha, \beta) = (1, 1) = (2, 2)$.

4.B.2 `Weighted`: Full Sample

We now extend the analysis above further for the weighted (biased) estimator:

$$d\widetilde{P}_{\alpha\beta}(\theta_k) = \widetilde{\mathcal{N}}^{\alpha\beta,\text{obs}}_{\text{tot}_{\text{obs}}} \left[1 + \widetilde{w}_{\alpha\beta}(\theta_k)\right] \frac{d\Omega}{V} \frac{d\Omega}{V} \tag{4.41}$$

where we introduce $\widetilde{\mathcal{N}}$ to account for the weighted pair counts which we define as

$$\widetilde{\mathcal{N}}^{\alpha\beta,\text{obs}}_{\text{tot}_{\text{obs}}} = \sum_{i}^{N^{\alpha}_{\text{tot}_{\text{obs}}}} \sum_{j\neq i}^{N^{\beta}_{\text{tot}_{\text{obs}}}} \mathtt{w}^{\alpha\beta}_{ij} \tag{4.42}$$

Now, when writing the analog of Equations 4.20-4.35, we need to account for pair weights, leading us to

$$d\widetilde{P}_{\alpha\beta}(\theta_k) = \widetilde{\mathcal{N}}^{\alpha\beta,\text{obs}}_{\text{tot}_{\text{obs}}} \left[1 + \widetilde{w}_{\alpha\beta}(\theta_k)\right] \frac{d\Omega}{V} \frac{d\Omega}{V} = \sum_{\gamma,\delta} \widetilde{\mathcal{N}}^{\gamma\delta,\text{true}}_{\text{tot}_{\text{obs}}} [1 + w^{\text{true}}_{\gamma,\delta}(\theta_k)] \frac{d\Omega}{V} \frac{d\Omega}{V} \tag{4.43}$$

where we have the analog of Equation 4.37:

$$\sum_{i}^{N^{\alpha}_{\text{tot}_{\text{obs}}}} \sum_{j\neq i}^{N^{\beta}_{\text{tot}_{\text{obs}}}} \mathtt{w}^{\alpha\beta}_{ij} \mathtt{q}^{\alpha}_{i} \mathtt{q}^{\beta}_{j} = \widehat{\widetilde{\mathcal{N}}}^{\alpha\beta,\text{true}}_{\text{tot}_{\text{obs}}} \tag{4.44}$$

Now, expanding the sum in Equation 4.43, we have

$$\begin{aligned} \widetilde{\mathcal{N}}^{\alpha\beta,\text{obs}}_{\text{tot}_{\text{obs}}} \left[1 + \widetilde{w}_{\alpha\beta}(\theta_k)\right] = \widetilde{\mathcal{N}}^{11,\text{true}}_{\text{tot}_{\text{obs}}} \left[1 + w^{\text{true}}_{11}(\theta_k)\right] + \widetilde{\mathcal{N}}^{12,\text{true}}_{\text{tot}_{\text{obs}}} \left[1 + w^{\text{true}}_{12}(\theta_k)\right] \\ + \widetilde{\mathcal{N}}^{21,\text{true}}_{\text{tot}_{\text{obs}}} \left[1 + w^{\text{true}}_{21}(\theta_k)\right] + \widetilde{\mathcal{N}}^{22,\text{true}}_{\text{tot}_{\text{obs}}} \left[1 + w^{\text{true}}_{22}(\theta_k)\right] \end{aligned} \tag{4.45}$$

Substituting Equation 4.37 to estimate the true counts, we have

$$\begin{aligned} &\left(\sum_{i}^{N^{\alpha}_{\text{tot}_{\text{obs}}}} \sum_{j\neq i}^{N^{\beta}_{\text{tot}_{\text{obs}}}} \mathtt{w}^{\alpha\beta}_{ij} \right) \left[1 + \widetilde{w}^{\text{full}}_{\alpha\beta}(\theta_k)\right] \\ &= \left(\sum_{i}^{N^{\alpha}_{\text{tot}_{\text{obs}}}} \sum_{j\neq i}^{N^{\beta}_{\text{tot}_{\text{obs}}}} \mathtt{w}^{\alpha\beta}_{ij} \mathtt{q}^{1}_{i} \mathtt{q}^{1}_{j} \right) \left[1 + w^{\text{true}}_{11}(\theta_k)\right] + \left(\sum_{i}^{N^{\alpha}_{\text{tot}_{\text{obs}}}} \sum_{j\neq i}^{N^{\beta}_{\text{tot}_{\text{obs}}}} \mathtt{w}^{\alpha\beta}_{ij} \mathtt{q}^{1}_{i} \mathtt{q}^{2}_{j} \right) \left[1 + w^{\text{true}}_{12}(\theta_k)\right] \\ &+ \left(\sum_{i}^{N^{\alpha}_{\text{tot}_{\text{obs}}}} \sum_{j\neq i}^{N^{\beta}_{\text{tot}_{\text{obs}}}} \mathtt{w}^{\alpha\beta}_{ij} \mathtt{q}^{2}_{i} \mathtt{q}^{1}_{j} \right) \left[1 + w^{\text{true}}_{21}(\theta_k)\right] + \left(\sum_{i}^{N^{\alpha}_{\text{tot}_{\text{obs}}}} \sum_{j\neq i}^{N^{\beta}_{\text{tot}_{\text{obs}}}} \mathtt{w}^{\alpha\beta}_{ij} \mathtt{q}^{2}_{i} \mathtt{q}^{2}_{j} \right) \left[1 + w^{\text{true}}_{22}(\theta_k)\right] \end{aligned} \tag{4.46}$$

Note that, this equation reduces to `Decontaminated` as in Equation 4.39 when weights

are set to 1 for target subsample and 0 for the rest; and we basically have theta-independent decontamination.

4.C `Weighted` Estimator: Variance and Practical Notes

4.C.1 `Weighted` Estimator: Variance

Here, we follow the procedure in LS93 to estimate the variance of the `Weighted` estimator introduced in Equation 4.13, filling in additional details while accounting for the weights in the data-data pair counts. While the details may be of value to the interested reader, we note that the derivation is lengthy, culminating in the analytical expression for the variance in 4.C.1.6. Specifically, we write the pair counts, i.e., the unnormalized $\overline{DD}$, $\overline{RR}$ histograms in terms of the fluctuations about their means, i.e., we have

$$\begin{aligned}(\overline{DD})_{\alpha\beta}(\theta_k) &= \left\langle (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle (1 + \eta(\theta_k)) \\ (\overline{RR})(\theta_k) &= \left\langle (\overline{RR})(\theta_k) \right\rangle (1 + \gamma(\theta_k))\end{aligned} \tag{4.47}$$

where we use the overline to distinguish the *unnormalized* histograms from the normalized ones (denoted with a tilde). Here, η and γ are the fluctuations in the histograms about their means, which follows

$$\langle \eta(\theta_k) \rangle = \langle \gamma(\theta_k) \rangle = 0 \tag{4.48}$$

and hence, we have

$$\begin{aligned}\sigma^2_\eta(\theta_k) &= \left\langle \eta^2(\theta_k) \right\rangle - \cancelto{0}{\langle \eta(\theta_k) \rangle^2} = \left\langle \eta^2(\theta_k) \right\rangle \\ \sigma^2_\gamma(\theta_k) &= \left\langle \gamma^2(\theta_k) \right\rangle - \cancelto{0}{\langle \gamma(\theta_k) \rangle^2} = \left\langle \gamma^2(\theta_k) \right\rangle \\ \mathrm{cov}(\eta, \gamma)(\theta_\mathrm{k}) &= \langle \eta(\theta_k)\gamma(\theta_k) \rangle - \cancelto{0}{\langle \eta(\theta_k) \rangle} \cancelto{0}{\langle \gamma(\theta_k) \rangle} = 0\end{aligned} \tag{4.49}$$

where $\langle \eta(\theta_k)\gamma(\theta_k) \rangle = 0$ since the data and random catalogs are not correlated. Note that η here is the same as α in LS93; we choose the former given that the latter letter is already in use here. Then, given Equation 4.13 and Equation 4.47, we have

$$\begin{aligned}1 + \widetilde{w}_{\alpha\beta}(\theta_k) &= \frac{(\widetilde{DD})_{\alpha\beta}(\theta_k)}{RR(\theta_k)} = \frac{(\overline{DD})_{\alpha\beta}(\theta_k)}{\sum_{j\neq i}^{N_\mathrm{tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{N_r(N_r-1)/2}{(\overline{RR})(\theta_k)} \\ &= \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_\mathrm{tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\left\langle (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle (1 + \eta(\theta_k))}{\left\langle (\overline{RR})(\theta_k) \right\rangle (1 + \gamma(\theta_k))}\end{aligned} \tag{4.50}$$

where we have collapsed the double sums for brevity, and have defined

$$RR(\theta_k) = \frac{\sum_i^{N_r}\sum_{j>i}^{N_r}\bar{\Theta}_{ij,k}}{\sum_i^{N_r}\sum_{j>i}^{N_r}} = \frac{\sum_i^{N_r}\sum_{j>i}^{N_r}\bar{\Theta}_{ij,k}}{N_r(N_r-1)/2} \tag{4.51}$$

$$\begin{aligned}
\Rightarrow 1 + \langle\widetilde{w}_{\alpha\beta}(\theta_k)\rangle &= \left\langle \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle\,(1+\eta(\theta_k))}{\langle(\overline{RR})\rangle\,(\theta_k)(1+\gamma(\theta_k))} \right\rangle \\
&= \frac{N_r(N_r-1)}{2} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \left\langle \frac{1}{\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \right\rangle \left\langle \frac{(1+\eta(\theta_k))}{(1+\gamma(\theta_k))} \right\rangle \\
&\approx \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \left\langle (1+\eta(\theta_k))(1-\gamma(\theta_k)+\gamma^2(\theta_k)) \right\rangle \\
&= \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \big\langle 1-\gamma(\theta_k)+\gamma^2(\theta_k)+\eta(\theta_k) \\
&\qquad\qquad -\eta(\theta_k)\gamma(\theta_k)+\eta(\theta_k)\gamma^2(\theta_k) \big\rangle
\end{aligned} \tag{4.52}$$

where we only keep the terms up to the second order in fluctuations. Note that the second equality is justified since the weights for individual galaxies are fixed across the different realizations. Now, we calculate the variance of the estimator as

$$\begin{aligned}
\mathrm{var}\left[\widetilde{w}_{\alpha\beta}\right](\theta_k) = \sigma^2_{\widetilde{w}_{\alpha\beta}}(\theta_k) &= \mathrm{var}\left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \times \right. \\
&\qquad \left. \left\{1-\gamma(\theta_k)+\gamma^2(\theta_k)+\eta(\theta_k)-\eta(\theta_k)\gamma(\theta_k)+\eta(\theta_k)\gamma^2(\theta_k)\right\} \right] \\
&\approx \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \right]^2 \mathrm{var}\left[1-\gamma(\theta_k)+\eta(\theta_k)\right] \\
&= \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \right]^2 \times \\
&\qquad \left[\sigma^2_\gamma(\theta_k) + \sigma^2_\eta(\theta_k) - 2\,\cancelto{0}{\mathrm{cov}(\eta(\theta_k),\gamma(\theta_k))} \right] \\
&= \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \right]^2 \left[\langle\gamma^2(\theta_k)\rangle + \langle\eta^2(\theta_k)\rangle\right]
\end{aligned} \tag{4.53}$$

where, again, we only keep the terms up to the second order in fluctuations. Here, as derived from Equation 4.47, we have the second moments of the fluctuations defined as

$$\langle\eta^2(\theta_k)\rangle = \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle - \langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle^2}{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle^2} \tag{4.54}$$

$$\left\langle \gamma^2(\theta_k) \right\rangle = \frac{\left\langle (\overline{RR})(\theta_k) \cdot (\overline{RR})(\theta_k) \right\rangle - \left\langle (\overline{RR})(\theta_k) \right\rangle^2}{\left\langle (\overline{RR})(\theta_k) \right\rangle^2} \tag{4.55}$$

In order to evaluate the variance, we calculate the second moments of the fluctuations using the first and second moments of the pair counts. Specifically, we only need $\left\langle (\overline{RR})(\theta_k) \right\rangle$, $\left\langle (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle$, and $\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle$; we do not need the second moment of the random pair counts, since $\left\langle \gamma^2 \right\rangle$ is simply the variance of the random data and hence the variance of the Poisson distribution.

4.C.1.1 Pair Counts: First and Second Moments

As in Section 2 in LS93, we consider counts in cells in order to write out the first and second moments of the pair counts. We calculate first moment of random pairs in 4.C.1.2; random pairs are uncorrelated in the limit of large N_r and hence present a simpler case. Then, we calculate the first moment of correlated data pairs in 4.C.1.3, followed by the second moment for the correlated data pairs in 4.C.1.4.

4.C.1.2 Random Pairs: First Moment

Here, we consider N_r points distributed randomly over the survey area, which we divide into K cells. The probability of finding the ith random point in any cell is the continuum probability, $\langle \rho_j \rangle = N_r/K$, in the limit of large enough K that we essentially have either zero or one point in each cell. This follows that the number of random pairs is

$$\left\langle (\overline{RR})(\theta_k) \right\rangle = \left\langle \sum_{j<i}^{K} \rho_i \rho_j \bar{\Theta}_{ij,k} \right\rangle = \frac{1}{2} \sum_{i \neq j}^{K} \langle \rho_i \rho_j \rangle \, \bar{\Theta}_{ij,k} \tag{4.56}$$

where we have borrowed the notation introduced in Equation 4.5 to express the heavisides. Now, the probability of finding two random points in two cells, chosen without replacement, is

$$\langle \rho_i \rho_j \rangle = \frac{N_r(N_r-1)}{K(K-1)} \tag{4.57}$$

and, similar to LS93 Equation 10, we have

$$\sum_{i \neq j}^{K} \bar{\Theta}_{ij,k} = K(K-1) G_p(\theta_k) \tag{4.58}$$

where $G_p(\theta_k)$ is the probability of finding two random points at separations $\theta_k \pm d\theta_k/2$. Hence $\sum_{i\neq j}^{K} \bar{\Theta}_{ij,k}$ is just the total number of random points with separations between $\theta_{\min,k}$, $\theta_{\max,k}$ as we have $K(K-1)$ cells. Substituting Equations 4.57-4.58 into Equation 4.56, we have

$$\left\langle (\overline{RR})(\theta_k) \right\rangle = \frac{1}{2}\frac{N_r(N_r-1)}{K(K-1)}\left[K(K-1)G_p(\theta_k)\right] = \frac{N_r(N_r-1)}{2}G_p(\theta_k) \tag{4.59}$$

4.C.1.3 Data Pairs: First Moment

Here, we have $N_{\rm tot}$ points distributed randomly over the survey area. As in 4.C.1.2, the probability of finding a galaxy in any cell is $\langle\nu\rangle = N_{\rm tot}/K$, in the limit of large enough K that we essentially have either no galaxy or one galaxy in each cell. Furthermore, we assign the pair weight to the cells in which the pair falls. This follows, given Equation 4.14, that

$$\left\langle (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle = C_\Omega \left\langle \sum_{i\neq j}^{K} \mathtt{w}_{ij}^{\alpha\beta}\nu_i\nu_j\bar{\Theta}_{ij,k} \right\rangle = C_\Omega \sum_{i\neq j}^{K} \left\langle \mathtt{w}_{ij}^{\alpha\beta} \right\rangle \langle\nu_i\nu_j\rangle\, \bar{\Theta}_{ij,k} \tag{4.60}$$

where C_Ω is a normalization constant to ensure that we recover the correct number of pairs, $\sum_{i\neq j}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}$, when integrating over all angles. Here, the pair weights are assumed to be uncorrelated with the probability of finding galaxies in a particular pair of cells, allowing us to separate their expectation values in the second equality; this assumption is valid since we are assigning pair weights based upon galaxy properties rather than their locations. Now, since data pairs are generally correlated, we must account for the correlation explicitly when considering the probabilities of finding a pair of galaxies in any two cells, chosen without replacement. That is, we have the probability of finding two galaxies in two cells separated by θ_k, chosen without replacement, as

$$\langle\nu_i\nu_j\rangle = \frac{N_{\rm tot}(N_{\rm tot}-1)}{K(K-1)}\left[1+w_{\alpha\beta}(\theta_k)\right] \tag{4.61}$$

Therefore, using Equations 4.58 and 4.61, Equation 4.60 becomes

$$\begin{aligned}
\left\langle (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle &= C_\Omega \left\langle \mathtt{w}_{ij}^{\alpha\beta} \right\rangle |_{i\neq j} \frac{N_{\rm tot}(N_{\rm tot}-1)}{K(K-1)}\left[1+w_{\alpha\beta}(\theta_k)\right]\left[K(K-1)G_p(\theta_k)\right] \\
&= C_\Omega \left[\frac{\sum_{i\neq j}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}}{N_{\rm tot}(N_{\rm tot}-1)}\right]\left[1+w_{\alpha\beta}(\theta_k)\right] G_p(\theta_k) N_{\rm tot}(N_{\rm tot}-1) \\
&= C_\Omega \left[1+w_{\alpha\beta}(\theta_k)\right] G_p(\theta_k) \sum_{i\neq j}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}
\end{aligned} \tag{4.62}$$

Now, before finding the normalization constant, we define w_Ω as the mean of $w_{\alpha\beta}(\theta_k)$ over the sampling geometry, i.e.,

$$w_\Omega \equiv \int_\Omega G_p(\theta_k) w_{\alpha\beta}(\theta_k) d\Omega \tag{4.63}$$

with $G_p(\theta_k)$ normalized to unity, i.e.,

$$\int_\Omega G_p(\theta_k) d\Omega = 1 \tag{4.64}$$

Therefore, we have

$$\begin{aligned} \int_\Omega \left\langle (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle d\Omega &= \sum_{i\neq j}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \\ \Rightarrow \int_\Omega C_\Omega G_p(\theta_k) \left[1 + w_{\alpha\beta}(\theta_k)\right] \sum_{i\neq j}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} &= \sum_{i\neq j}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \\ \Rightarrow C_\Omega &= \frac{1}{1+w_\Omega} \end{aligned} \tag{4.65}$$

where we make use of Equation 4.64. Therefore, Equation 4.62 becomes

$$\left\langle (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle = G_p(\theta_k) \left[\frac{1 + w_{\alpha\beta}(\theta_k)}{1 + w_\Omega}\right] \sum_{i\neq j}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \tag{4.66}$$

4.C.1.4 Data-Data Pairs

As in LS93, using counts in cells, the second moment is defined as

$$\begin{aligned} \left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle &= \left\langle \sum_{j\neq i}^{K} \mathtt{w}_{ij}^{\alpha\beta} \nu_i \nu_j \bar{\Theta}_{ij,k} \sum_{l\neq m}^{K} \mathtt{w}_{ml}^{\alpha\beta} \nu_m \nu_l \bar{\Theta}_{ml,k} \right\rangle \\ &= \sum_{j\neq i}^{K} \sum_{l\neq m}^{K} \left\langle \nu_i \nu_j \nu_m \nu_l \right\rangle \left\langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \right\rangle \bar{\Theta}_{ij,k} \bar{\Theta}_{ml,k} \end{aligned} \tag{4.67}$$

Now, there are three cases to consider, each of which needs to be normalized to give the right total weight from each case (as done in 4.C.1.3):

1. No indices overlap: there are $K(K-1)(K-2)(K-3)$ cases of the sort since we choose each of the four cells without replacement. Since the data pairs are correlated, the probability of

finding each of the four galaxies in the four cells, chosen without replacement, is given by

$$\begin{aligned}\langle \nu_i \nu_j \nu_m \nu_l \rangle = \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)}{K(K-1)(K-2)(K-3)} \left[1 + w_{ij}(\theta_k) + w_{im}(\theta_k) + w_{il}(\theta_k) \right. \\ \left. + w_{jm}(\theta_k) + w_{jl}(\theta_k) + w_{ml}(\theta_k)\right]\end{aligned} \tag{4.68}$$

Here, since pairs i, j and m, l are separated by $\theta_k \pm d\theta_k/2$, $w_{ij}(\theta_k) = w_{ml}(\theta_k) = w_{\alpha\beta}(\theta_k)$ while the rest of the correlations can be approximated as w_Ω. Therefore,

$$\langle \nu_i \nu_j \nu_m \nu_l \rangle = \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)}{K(K-1)(K-2)(K-3)} \left[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega\right] \tag{4.69}$$

Also, as in LS93, we introduce $G_q(\theta_k)$ as the probability of finding quadrilaterals, i.e., pairs i, j and m, l separated by $\theta_k \pm d\theta_k/2$. Then, the total number of quadrilaterals is

$$\sum_{\rm unique\{i,j,l,m\}}^{K} \bar{\Theta}_{ij,k}\bar{\Theta}_{ml,k} = K(K-1)(K-2)(K-3)G_q(\theta_k), \quad i \neq j, m \neq l \tag{4.70}$$

Note that as in Equation 4.64, $G_q(\theta_k)$ is also normalized to unity, i.e.,

$$\int_\Omega G_q(\theta_k) d\Omega = 1 \tag{4.71}$$

Therefore, the contribution to the second moment of the pair counts by the quadrilaterals is given by

$$\begin{aligned}\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm quad} &= C_{\rm quad} \sum_{j\neq i\neq l\neq m}^{K} \langle \nu_i \nu_j \nu_m \nu_l \rangle \left\langle \mathtt{W}_{ij}^{\alpha\beta} \mathtt{W}_{ml}^{\alpha\beta} \right\rangle \bar{\Theta}_{ij,k}\bar{\Theta}_{ml,k} \\ &= C_{\rm quad} N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3) \left[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega\right] \times \\ &\qquad G_q(\theta_k) \left\langle \mathtt{W}_{ij}^{\alpha\beta} \mathtt{W}_{ml}^{\alpha\beta} \right\rangle_{i\neq j\neq m\neq l} \\ &= C_{\rm quad} N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3) \left[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega\right] \times \\ &\qquad G_q(\theta_k) \left[\frac{\sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{W}_{ij}^{\alpha\beta} \mathtt{W}_{ml}^{\alpha\beta}}{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)} \right] \\ &= C_{\rm quad} \left[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega\right] G_q(\theta_k) \sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{W}_{ij}^{\alpha\beta} \mathtt{W}_{ml}^{\alpha\beta}\end{aligned} \tag{4.72}$$

where $C_{\rm quad}$ is the normalization constant so that we get the correct weight for the quadrilaterals when integrating over all angles, i.e.,

$$\begin{aligned}\int \left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm quad} d\Omega &= \sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \\ \Rightarrow \int \left\{ C_{\rm quad} \left[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega\right] G_q(\theta_k) \right\} d\Omega &= 1 \\ \Rightarrow C_{\rm quad} &= \frac{1}{1 + 2\int w_{\alpha\beta}(\theta_k) G_q(\theta_k) d\Omega + 4w_\Omega} \\ &= \frac{1}{1 + 2w_{\Omega,q} + 4w_\Omega}\end{aligned} \tag{4.73}$$

where we have used Equation 4.71 and have defined a new mean:

$$w_{\Omega,q} \equiv \int w_{\alpha\beta}(\theta_k) G_q(\theta_k) d\Omega \tag{4.74}$$

Therefore,

$$\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm quad} = \left[\frac{1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega}{1 + 2w_{\Omega,q} + 4w_\Omega} \right] G_q(\theta_k) \sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \tag{4.75}$$

2. One of the indices is repeated: there are $K(K-1)(K-2)$ cases of the sort, since we choose only three cells without replacement, i.e., we choose two cells for the first $(\overline{DD})$ and one for the second $(\overline{DD})$. Note that we do not have to account for m, l swap since we consider the two cases explicitly when calculating $\langle \nu_i \nu_j \nu_m \nu_l \rangle$ (needed since the swap carries different meaning for the pair weights). As for the probabilities of finding the data points in the

chosen cells, we have

$$\begin{aligned}
\langle \nu_i \nu_j \nu_m \nu_l \rangle \,|_{i=m} &= \left\langle \nu_i^2 \nu_j \nu_l \right\rangle = \langle \nu_i \nu_j \nu_l \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + w_{ij}(\theta_k) + w_{il}(\theta_k) + w_{jl}(\theta_k)\right] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + 3w_{\alpha\beta}(\theta_k)\right] \\
\langle \nu_i \nu_j \nu_m \nu_l \rangle \,|_{i=l} &= \left\langle \nu_i^2 \nu_l \nu_m \right\rangle = \langle \nu_i \nu_l \nu_m \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + w_{il}(\theta_k) + w_{im}(\theta_k) + w_{lm}(\theta_k)\right] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + 3w_{\alpha\beta}(\theta_k)\right] \\
\langle \nu_i \nu_j \nu_m \nu_l \rangle \,|_{j=m} &= \left\langle \nu_i \nu_j^2 \nu_l \right\rangle = \langle \nu_i \nu_j \nu_l \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + w_{ij}(\theta_k) + w_{il}(\theta_k) + w_{jl}(\theta_k)\right] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + 3w_{\alpha\beta}(\theta_k)\right] \\
\langle \nu_i \nu_j \nu_m \nu_l \rangle \,|_{j=l} &= \left\langle \nu_i \nu_j^2 \nu_m \right\rangle = \langle \nu_i \nu_j \nu_m \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + w_{ij}(\theta_k) + w_{im}(\theta_k) + w_{jm}(\theta_k)\right] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)} \left[1 + 3w_{\alpha\beta}(\theta_k)\right]
\end{aligned} \tag{4.76}$$

where we note that $\langle \nu \rangle = \left\langle \nu^2 \right\rangle = N_{\rm tot}/K$ since we are working in the large-K regime where there is only 0 or 1 galaxy in each cell. Also, as in LS93, we introduce $G_t(\theta_k)$ as the probability of finding triangles, i.e., two galaxies within $\theta_k \pm d\theta_k/2$ of a given galaxy. Then, the total number of triangles is

$$\sum_{\text{unique}\{i,j,m\};l=i}^{K} \bar{\Theta}_{ij,k} \bar{\Theta}_{ml,k} = K(K-1)(K-2) G_t(\theta_k), \quad i \neq j, m \neq i \tag{4.77}$$

where $G_t(\theta_k)$ is also normalized to unity:

$$\int_\Omega G_t(\theta_k) d\Omega = 1 \tag{4.78}$$

Therefore, the contribution to the second moment of the pair counts by the triangles is given

by

$$\begin{aligned}\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm tri} &= C_{\rm tri} N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2) G_t(\theta_k)\left[1+3w_{\alpha\beta}(\theta_k)\right] \times \\ &\left\{ \left\langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta}\right\rangle_{i=m\neq j\neq l} + \left\langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta}\right\rangle_{i=l\neq j\neq m} \right. \\ &\left. + \left\langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta}\right\rangle_{i\neq j=m\neq l} + \left\langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta}\right\rangle_{i\neq j=l\neq m} \right\} \\ &= C_{\rm tri} G_t(\theta_k)\left[1+3w_{\alpha\beta}(\theta_k)\right] \sum_{i\neq j\neq l}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{il}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{li}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{jl}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{lj}^{\alpha\beta} \right\}\end{aligned} \tag{4.79}$$

where $C_{\rm tri}$ is the normalization constant so that we get the correct weight for the triangles when integrating over all angles, i.e.,

$$\begin{aligned}\int \left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm tri} d\Omega &= \sum_{i\neq j\neq l}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{il}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{li}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{jl}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{lj}^{\alpha\beta} \right\} \\ \Rightarrow \int \left\{ C_{\rm tri}\left[1+3w_{\alpha\beta}(\theta_k)\right] G_t(\theta_k) \right\} d\Omega &= 1 \\ \Rightarrow C_{\rm tri} &= \frac{1}{1+3\int w_{\alpha\beta}(\theta_k) G_t(\theta_k) d\Omega + 3w_\Omega} \\ &= \frac{1}{1+3w_{\Omega,t}}\end{aligned} \tag{4.80}$$

where we have used Equation 4.78 and have defined a new mean:

$$w_{\Omega,t} \equiv \int w_{\alpha\beta}(\theta_k) G_t(\theta_k) d\Omega \tag{4.81}$$

Therefore,

$$\begin{aligned}\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm tri} &= \left[\frac{1+3w_{\alpha\beta}(\theta_k)}{1+3w_{\Omega,t}} \right] G_t(\theta_k) \times \\ &\sum_{i\neq j\neq l}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{il}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{li}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{jl}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{lj}^{\alpha\beta} \right\}\end{aligned} \tag{4.82}$$

3. Two of the indices overlap: there are $K(K-1)$ cases, since we choose only two cells. This follows that the probability of finding two galaxies in the chosen cells is

$$\begin{aligned}\langle \nu_i\nu_j\nu_m\nu_l \rangle_{i=m,j=l} &= \langle \nu_i\nu_j\nu_i\nu_j \rangle = \left\langle \nu_i^2\nu_j^2 \right\rangle = \langle \nu_i\nu_j \rangle = \frac{N_{\rm tot}(N_{\rm tot}-1)}{K(K-1)}\left[1+w_{\alpha\beta}(\theta_k)\right] \\ \langle \nu_i\nu_j\nu_m\nu_l \rangle_{i=l,j=m} &= \langle \nu_i\nu_j\nu_j\nu_i \rangle = \left\langle \nu_i^2\nu_j^2 \right\rangle = \langle \nu_i\nu_j \rangle = \frac{N_{\rm tot}(N_{\rm tot}-1)}{K(K-1)}\left[1+w_{\alpha\beta}(\theta_k)\right]\end{aligned} \tag{4.83}$$

Here, Equation 4.58 applies, giving us the contribution to the second moment of the pair counts by the pairs as

$$\begin{aligned}
&\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm pairs} \\
&= C_{\rm pairs} N_{\rm tot}(N_{\rm tot}-1) G_p(\theta_k) \left[1 + w_{\alpha\beta}(\theta_k)\right] \left\{ \left\langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \right\rangle_{i=m\neq j=l} + \left\langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \right\rangle_{i=l\neq j=m} \right\} \\
&= C_{\rm pairs} N_{\rm tot}(N_{\rm tot}-1) G_p(\theta_k) \left[1 + w_{\alpha\beta}(\theta_k)\right] \left\{ \left\langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ij}^{\alpha\beta} \right\rangle_{i\neq j} + \left\langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ji}^{\alpha\beta} \right\rangle_{i\neq j} \right\} \\
&= C_{\rm pairs} G_p(\theta_k) \left[1 + w_{\alpha\beta}(\theta_k)\right] \sum_{i\neq j}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ij}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ji}^{\alpha\beta} \right\}
\end{aligned} \tag{4.84}$$

where $C_{\rm pairs}$ is the normalization constant so that we get the correct weight for the pairs when integrating over all angles, i.e.,

$$\begin{aligned}
\int \left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm pairs} d\Omega &= \sum_{i\neq j}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ij}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ji}^{\alpha\beta} \right\} \\
\Rightarrow \int \left\{ C_{\rm pairs} \left[1 + w_{\alpha\beta}(\theta_k)\right] G_p(\theta_k) \right\} d\Omega &= 1 \\
\Rightarrow C_{\rm pairs} &= \frac{1}{1 + w_\Omega}
\end{aligned} \tag{4.85}$$

where we have used Equation 4.64; this results matches with Equation 4.65 as it should. Therefore,

$$\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle_{\rm pairs} = G_p(\theta_k) \left[\frac{1 + w_{\alpha\beta}(\theta_k)}{1 + w_\Omega} \right] \sum_{i\neq j}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ij}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ji}^{\alpha\beta} \right\} \tag{4.86}$$

Combining the three cases, i.e., Equations 4.75, 4.82 and 4.86, Equation 4.67 becomes

$$\begin{aligned}
\left\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \right\rangle &= \sum_{j\neq i}^{K} \sum_{l\neq m}^{K} \left\langle \nu_i \nu_j \nu_m \nu_l \right\rangle \left\langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \right\rangle \bar{\Theta}_{ij,k} \bar{\Theta}_{ml,k} \\
&= \left[\frac{1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega}{1 + 2w_{\Omega,q} + 4w_\Omega} \right] G_p(\theta_k)^2 \sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \\
&\quad + \left[\frac{1 + 3w_{\alpha\beta}(\theta_k)}{1 + 3w_{\Omega,t}} \right] G_t(\theta_k) \sum_{i\neq j\neq l}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{il}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{li}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{jl}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{lj}^{\alpha\beta} \right\} \\
&\quad + G_p(\theta_k) \left[\frac{1 + w_{\alpha\beta}(\theta_k)}{1 + w_\Omega} \right] \sum_{i\neq j}^{N_{\rm tot}} \left\{ \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ij}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ji}^{\alpha\beta} \right\}
\end{aligned} \tag{4.87}$$

where we have used the result $G_q(\theta_k) = G_p^2(\theta_k)$ from LS93, valid in the large-K limit.

4.C.1.5 Fluctuations

Now, substituting Equations 4.66, 4.87 in Equation 4.54, we have

$$\begin{aligned}\left\langle \eta^2(\theta_k)\right\rangle &= \frac{\begin{array}{r}\left[\frac{1+2w_{\alpha\beta}(\theta_k)+4w_\Omega}{1+2w_{\Omega,q}+4w_\Omega}\right] G_p(\theta_k)^2 \sum_{i\neq j\neq m\neq l}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ml}^{\alpha\beta} \\ + \left[\frac{1+3w_{\alpha\beta}(\theta_k)}{1+3w_{\Omega,t}}\right] G_t(\theta_k) \sum_{i\neq j\neq l}^{N_{\text{tot}}} \left\{\mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{il}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{li}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{jl}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{lj}^{\alpha\beta}\right\} \\ + G_p(\theta_k)\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right] \sum_{i\neq j}^{N_{\text{tot}}} \left\{\mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ij}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ji}^{\alpha\beta}\right\}\end{array}}{\left(G_p(\theta_k)\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}\right)^2} - 1 \\ &= \frac{\begin{array}{r}\left[\frac{1+2w_{\alpha\beta}(\theta_k)+4w_\Omega}{1+2w_{\Omega,q}+4w_\Omega}\right] \sum_{i\neq j\neq m\neq l}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ml}^{\alpha\beta} \\ + \left[\frac{1+3w_{\alpha\beta}(\theta_k)}{1+3w_{\Omega,t}}\right] \frac{G_t(\theta_k)}{G_p^2(\theta_k)} \sum_{i\neq j\neq l}^{N_{\text{tot}}} \left\{\mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{il}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{li}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{jl}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{lj}^{\alpha\beta}\right\} \\ + \frac{1}{G_p(\theta_k)}\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right] \sum_{i\neq j}^{N_{\text{tot}}} \left\{\mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ij}^{\alpha\beta} + \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ji}^{\alpha\beta}\right\}\end{array}}{\left(\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}\right)^2} - 1\end{aligned} \tag{4.88}$$

As for $\left\langle \gamma^2(\theta_k)\right\rangle$, given Equation 4.59, it takes the form

$$\left\langle \gamma^2(\theta_k)\right\rangle = \frac{2}{N_r(N_r-1)G_p(\theta_k)} \tag{4.89}$$

4.C.1.6 Variance

We now go back to Equation 4.53, and attempt to evaluate it. First, substituting Equations 4.66 and 4.59, we have

$$\begin{aligned}\sigma^2_{\widetilde{w}_{\alpha\beta}}(\theta_k) &= \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}} \frac{G_p(\theta_k)\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}}{\frac{N_r(N_r-1)}{2}G_p(\theta_k)}\right]^2 \left[\left\langle \gamma^2(\theta_k)\right\rangle + \left\langle \eta^2(\theta_k)\right\rangle\right] \\ &= \left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]^2 \left[\left\langle \gamma^2(\theta_k)\right\rangle + \left\langle \eta^2(\theta_k)\right\rangle\right]\end{aligned} \tag{4.90}$$

Now, in the limit of large N_r, i.e., $\left\langle \gamma^2\right\rangle \to 0$, we have

$$\sigma^2_{\widetilde{w}_{\alpha\beta}}(\theta_k) \xrightarrow[\text{large } N_r]{} \left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]^2 \left\langle \eta^2(\theta_k)\right\rangle \tag{4.91}$$

where $\left\langle \eta^2(\theta_k)\right\rangle$ is given by Equation 4.88. The expression can be simplified: we first look at leading order term, i.e., the quadrilateral contribution:

$$\sigma^2_{\widetilde{w}_{\alpha\beta}}(\theta_k) \xrightarrow[\text{order}]{\text{leading}} \frac{\left[\frac{1+2w_{\alpha\beta}(\theta_k)+4w_\Omega}{1+2w_{\Omega,q}+4w_\Omega}\right] \sum_{i\neq j\neq m\neq l}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ml}^{\alpha\beta}}{\left(\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\text{tot}}} \mathtt{W}_{ij}^{\alpha\beta}\right)^2} - 1 \tag{4.92}$$

Then, in the limit of weak correlations as then $1 << w_{\alpha\beta}(\theta_k) \sim w_\Omega < w_{\Omega,t} < w_{\Omega,q}$, we have

$$\sigma^2_{\widetilde{w}_{\alpha\beta}}(\theta_k) \xrightarrow[\text{correlations}]{\text{weak}} \frac{\sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{W}_{ij}^{\alpha\beta}\mathtt{W}_{ml}^{\alpha\beta}}{\left(\sum_{i\neq j}^{N_{\rm tot}} \mathtt{W}_{ij}^{\alpha\beta}\right)^2} - 1 \tag{4.93}$$

where we note that $\mathtt{W}_{ij}^{\alpha\beta} = \mathtt{W}_{ji}^{\beta\alpha}$.

Now, in order to get the analytical expression for the variance of the unbiased estimator, i.e., the `Decontaminated Weighted` estimator, we must consider not only the variance of each of the biased correlations but also the covariances. As an example, based on Equation 4.18 which is valid for when there are two galaxy types in our observed sample, we essentially have the unbiased estimator for the AA auto-correlation function as

$$\widehat{w}_{AA}(\theta_k) = C_{AA}(\theta_k)\widetilde{w}^{\rm obs}_{AA}(\theta_k) + C_{AB}(\theta_k)\widetilde{w}^{\rm obs}_{AB}(\theta_k) + C_{BB}(\theta_k)\widetilde{w}^{\rm obs}_{BB}(\theta_k) \tag{4.94}$$

where $C_{AA}(\theta_k), C_{AB}(\theta_k), C_{BB}(\theta_k)$ are the elements of the first row of the inverse matrix in Equation 4.18. Given the dependency of all terms and factors on the pair weights, we have the variance of the unbiased estimator as

$$\begin{aligned}
\sigma^2_{\widehat{w}_{AA}}(\theta_k) = {} & C^2_{AA}(\theta_k)\sigma^2_{\widetilde{w}_{AA}}(\theta_k) + C^2_{AB}(\theta_k)\sigma^2_{\widetilde{w}_{AB}}(\theta_k) + C^2_{BB}(\theta_k)\sigma^2_{\widetilde{w}_{BB}}(\theta_k) \\
& - 2\text{cov}\left[C_{AA}(\theta_k), \widetilde{w}^{\rm obs}_{AA}(\theta_k)\right] - 2\text{cov}\left[C_{AB}(\theta_k), \widetilde{w}^{\rm obs}_{AB}(\theta_k)\right] \\
& - 2\text{cov}\left[C_{BB}(\theta_k), \widetilde{w}^{\rm obs}_{BB}(\theta_k)\right] - 2\widetilde{w}^{\rm obs}_{AA}(\theta_k)\widetilde{w}^{\rm obs}_{AB}(\theta_k)\text{cov}\left[C_{AA}(\theta_k), C_{AB}(\theta_k)\right] \\
& - 2\widetilde{w}^{\rm obs}_{AA}(\theta_k)\widetilde{w}^{\rm obs}_{BB}(\theta_k)\text{cov}\left[C_{AA}(\theta_k), C_{BB}(\theta_k)\right] \\
& - 2\widetilde{w}^{\rm obs}_{AB}(\theta_k)\widetilde{w}^{\rm obs}_{BB}(\theta_k)\text{cov}\left[C_{AB}(\theta_k), C_{BB}(\theta_k)\right]
\end{aligned} \tag{4.95}$$

This expression is unwieldy to evaluate for the general case, even if when we use the leading-order, weak-correlation approximation as in Equation 4.93. Therefore, we resort to numerical estimation of the variance.

4.C.2 `Weighted` Estimator: Practical Notes

4.C.2.1 Weighted Data-Data Pair Counts

Here, we note some points that are important when it comes to implementing the `Weighted` estimator proposed in Equation 4.13. Specifically considering Equation 4.14 for the auto correlation, we have

$$(\widetilde{DD})_{AA}(\theta_k) = \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathtt{W}_{ij}^{AA}\bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathtt{W}_{ij}^{AA}} \tag{4.96}$$

while for the cross, we have

$$(\widetilde{DD})_{AB}(\theta_k) = \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB} \bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{AB}} \tag{4.97}$$

It might appear that $(\widetilde{DD})_{AB} \neq (\widetilde{DD})_{BA}$ since $\mathtt{w}_{ij}^{AB} \neq \mathtt{w}_{ij}^{BA}$ but we must realize that

$$\mathtt{w}_{ij}^{AB} = \mathtt{w}_{ji}^{BA} \tag{4.98}$$

and since the sums are re-indexable, we have

$$(\widetilde{DD})_{BA}(\theta_k) = \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BA} \bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{BA}} = \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ji}^{AB} \bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ji}^{AB}} = (\widetilde{DD})_{AB}(\theta_k) \tag{4.99}$$

Therefore, when implementing the weighted data-data histogram, we can work with either $\mathtt{w}_{ij}^{\alpha\beta}$ or $\mathtt{w}_{ij}^{\beta\alpha}$, even though $\mathtt{w}_{ij}^{\alpha\beta} \neq \mathtt{w}_{ij}^{\beta\alpha}$ when $\alpha \neq \beta$.

4.C.2.2 Pair Weights

While we have used simple pair weights in this work, i.e., $\mathtt{w}_{ij}^{\alpha\beta} = \mathtt{q}_i^{\alpha}\mathtt{q}_j^{\beta}$, the `Weighted` estimator presented in Equation 4.13 works with general pair weights. In the case where the pair weights are not separable (e.g., they account for a theta-dependence), we must circumvent the problem presented by the normalization of the data-data histogram in Equation 4.14: it requires summing over all the pair weights – a task that is computationally prohibitive when working with large datasets where standard correlation function algorithms focus on a specified range of separations to reduce compute time. We can address the challenge by two methods: 1) estimating the number of pairs and the average weights for the larger θ-bins, and hence still being able to use the all-pairs normalization, and 2) introducing a new, exact normalization, which can be achieved by considering Equation 4.13 with its full details, i.e.,

$$\begin{aligned} \widetilde{w}_{\alpha\beta}^{\rm obs}(\theta_k) + 1 = \frac{(\widetilde{DD})_{\alpha\beta}(\theta_k)}{RR(\theta_k)} &= \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \frac{\sum_i^{N_r} \sum_{j\neq i}^{N_r}}{\sum_i^{N_r} \sum_{j\neq i}^{N_r} \bar{\Theta}_{ij,k}} \\ &= \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \bar{\Theta}_{ij,k}}{\sum_i^{N_r} \sum_{j\neq i}^{N_r} \bar{\Theta}_{ij,k}} \frac{\sum_i^{N_r} \sum_{j\neq i}^{N_r}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta}} \end{aligned} \tag{4.100}$$

where the first fraction in the last line compares the data-data pair weight in bin k with the random-random pairs in the same bins, while the second fraction normalizes the total data-data pair weight with the total random-random pair counts. Now, since exact numerical calculation

of the total data-data pair weight is prohibitive and affects only the overall normalization, we can normalize *both* the total data-data pair weight and the total random pair counts in a less computationally challenging way, i.e.,

$$\widetilde{w}^{\rm obs}_{\alpha\beta}(\theta_k)+1=\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{\alpha\beta}\bar{\Theta}_{ij,k}}{\sum_i^{N_r}\sum_{j\neq i}^{N_r}\bar{\Theta}_{ij,k}}\frac{\sum_m^{N_{\rm bin}}\sum_i^{N_r}\sum_{j\neq i}^{N_r}\bar{\Theta}_{ij,m}}{\sum_m^{N_{\rm bin}}\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{\alpha\beta}\bar{\Theta}_{ij,m}} \tag{4.101}$$

where have replaced the total counts over all possible scales to those in only the scales of interest.

4.C.3 Direct Decontamination

Here we attempt to find weights that allow us to decontaminate *while* estimating the correlations – a step towards optimal weights. To achieve this, we consider Equation 4.17 which is reproduced here for convenience:

$$\begin{bmatrix}\langle\widetilde{w}^{\rm obs}_{AA}(\theta_k)\rangle\\ \langle\widetilde{w}^{\rm obs}_{AB}(\theta_k)\rangle\\ \langle\widetilde{w}^{\rm obs}_{BB}(\theta_k)\rangle\end{bmatrix}=\begin{bmatrix}\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}\mathtt{q}_i^A\mathtt{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}\{\mathtt{q}_i^A\mathtt{q}_j^B+\mathtt{q}_i^B\mathtt{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}\mathtt{q}_i^B\mathtt{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}}\\ \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}\mathtt{q}_i^A\mathtt{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}\{\mathtt{q}_i^A\mathtt{q}_j^B+\mathtt{q}_i^B\mathtt{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}\mathtt{q}_i^B\mathtt{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}}\\ \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}\mathtt{q}_i^A\mathtt{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}\{\mathtt{q}_i^A\mathtt{q}_j^B+\mathtt{q}_i^B\mathtt{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}\mathtt{q}_i^B\mathtt{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}}\end{bmatrix}\begin{bmatrix}w^{\rm true}_{AA}(\theta_k)\\ w^{\rm true}_{AB}(\theta_k)\\ w^{\rm true}_{BB}(\theta_k)\end{bmatrix} \tag{4.102}$$

In order to achieve our goal, we would like to find weights $\mathtt{w}^{\alpha\beta}_{ij,{\rm opt}}$ such that we can write the above equation as

$$\begin{bmatrix}\langle\widetilde{w}^{\rm obs}_{AA}(\theta_k)\rangle\\ \langle\widetilde{w}^{\rm obs}_{AB}(\theta_k)\rangle\\ \langle\widetilde{w}^{\rm obs}_{BB}(\theta_k)\rangle\end{bmatrix}=\begin{bmatrix}\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}\mathtt{q}_i^A\mathtt{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AA}} & 0 & 0\\ 0 & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}\{\mathtt{q}_i^A\mathtt{q}_j^B+\mathtt{q}_i^B\mathtt{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{AB}} & 0\\ 0 & 0 & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}\mathtt{q}_i^B\mathtt{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathtt{w}_{ij}^{BB}}\end{bmatrix}\begin{bmatrix}w^{\rm true}_{AA}(\theta_k)\\ w^{\rm true}_{AB}(\theta_k)\\ w^{\rm true}_{BB}(\theta_k)\end{bmatrix} \tag{4.103}$$

To consider a simple scenario, we first assume that the pair weights are a linear product of the weights of individual weights, i.e., $\mathtt{w}^{\alpha\beta}_{ij,\mathrm{opt}} = \mathtt{w}^{\alpha}_{i,\mathrm{opt}}\mathtt{w}^{\beta}_{j,\mathrm{opt}}$, which follows that we only need to find $\mathtt{w}^{\alpha}_{i,\mathrm{opt}}$ and $\mathtt{w}^{\beta}_{i,\mathrm{opt}}$ (where we note α, β can be either A or B). Then, we must have the non-diagonal terms in Equation 4.102 be zero, leading us to specific constraints on the pair weights. To demonstrate the method, we achieved the optimization by assuming a functional form for the optimized weights:

$$\mathtt{w}^{\alpha}_{i,\mathrm{opt}} = \mu^{\alpha} + \nu^{\alpha} q^{\alpha}_{i} \tag{4.104}$$

where μ, ν are the optimization parameters and are allowed to be negative (which is what allows this method to mimic `Decontaminated` by automatically subtracting off pairs in which one contributor is likely a contaminant). Using this method, we were able to decontaminate as effectively as `Decontaminated` for the 2-sample case, but without reducing the variance. We note that the equivalence between this direct decontamination with optimized weights and `Decontaminated` is not guaranteed for larger numbers of samples or for weights that are non-linear functions of probability, meriting further investigation as part of a larger investigation of optimizing the weights.

4.D Generalized Estimators

4.D.1 `Decontaminated` Estimator

As an extension of our derivation for two samples in Section 4.3.1, we now consider three samples, with galaxies of Types A, B, C present in our sample. For instance, we have

$$\begin{aligned} w^{\mathrm{obs}}_{AB}(\theta_k) = {} & f_{AA}f_{BA}w^{\mathrm{true}}_{AA}(\theta_k) + \{f_{AA}f_{BB} + f_{AB}f_{BA}\}\, w^{\mathrm{true}}_{AB}(\theta_k) + f_{AB}f_{BB}w^{\mathrm{true}}_{BB}(\theta_k) \\ & + \{f_{AB}f_{BC} + f_{AC}f_{BB}\}\, w^{\mathrm{true}}_{BC}(\theta_k) + f_{AC}f_{BC}w^{\mathrm{true}}_{CC}(\theta_k) \\ & + \{f_{AA}f_{BC} + f_{AC}f_{BA}\}\, w^{\mathrm{true}}_{CA}(\theta_k) \end{aligned} \tag{4.105}$$

Therefore, similar to the construction of Equation 4.12, we have

$$\begin{bmatrix} \widehat{w}_{AA}(\theta_k) \\ \widehat{w}_{AB}(\theta_k) \\ \widehat{w}_{BB}(\theta_k) \\ \widehat{w}_{BC}(\theta_k) \\ \widehat{w}_{CC}(\theta_k) \\ \widehat{w}_{CA}(\theta_k) \end{bmatrix} = \begin{bmatrix} \varsigma_{AA}^{AA} & 2\varsigma_{AB}^{AA} & \varsigma_{AB}^{AB} & 2\varsigma_{AC}^{AB} & \varsigma_{AC}^{AC} & 2\varsigma_{AC}^{AA} \\ \varsigma_{BA}^{AA} & \varsigma_{BB}^{AA} + \varsigma_{AB}^{BA} & \varsigma_{AB}^{BB} & \varsigma_{BC}^{AB} + \varsigma_{AC}^{BB} & \varsigma_{BC}^{AC} & \varsigma_{BC}^{AA} + \varsigma_{AC}^{BA} \\ \varsigma_{BA}^{BA} & 2\varsigma_{BA}^{BB} & \varsigma_{BB}^{BB} & 2\varsigma_{BC}^{BB} & \varsigma_{BC}^{BC} & 2\varsigma_{BC}^{BA} \\ \varsigma_{CB}^{BA} & \varsigma_{CB}^{BA} + \varsigma_{BB}^{CA} & \varsigma_{CB}^{BB} & \varsigma_{CC}^{BB} + \varsigma_{BC}^{CB} & \varsigma_{CC}^{BC} & \varsigma_{BC}^{BA} + \varsigma_{BC}^{BA} \\ \varsigma_{CA}^{CA} & 2\varsigma_{CB}^{CA} & \varsigma_{CB}^{CB} & 2\varsigma_{CC}^{CB} & \varsigma_{CC}^{CC} & 2\varsigma_{CC}^{CA} \\ \varsigma_{CA}^{AA} & \varsigma_{CB}^{AA} + \varsigma_{AB}^{CA} & \varsigma_{CB}^{AB} & \varsigma_{CC}^{AB} + \varsigma_{AC}^{CB} & \varsigma_{CC}^{AC} & \varsigma_{CC}^{AA} + \varsigma_{AC}^{CA} \end{bmatrix}^{-1} \begin{bmatrix} w_{AA}^{\text{obs}}(\theta_k) \\ w_{AB}^{\text{obs}}(\theta_k) \\ w_{BB}^{\text{obs}}(\theta_k) \\ w_{BC}^{\text{obs}}(\theta_k) \\ w_{CC}^{\text{obs}}(\theta_k) \\ w_{CA}^{\text{obs}}(\theta_k) \end{bmatrix} \tag{4.106}$$

where we have defined the following for brevity:

$$\varsigma_{mn}^{ij} = f_{A_i A_j} f_{A_m A_n} = \varsigma_{ij}^{mn} \tag{4.107}$$

Extending the idea to M samples, we can write the analog of the unbiased estimator for `Decontamination`, given by Equation 4.12, as

$$\begin{bmatrix} \widehat{w}_{11}(\theta_k) \\ \widehat{w}_{12}(\theta_k) \\ \vdots \\ \widehat{w}_{\gamma\gamma}(\theta_k) \\ \widehat{w}_{\gamma(\gamma+1)}(\theta_k) \\ \vdots \\ \widehat{w}_{MM}(\theta_k) \\ \widehat{w}_{M1}(\theta_k) \end{bmatrix} = \begin{bmatrix} \varsigma_{11}^{11} & 2\varsigma_{12}^{11} & \cdots & \varsigma_{1\gamma}^{1\gamma} & 2\varsigma_{1(\gamma+1)}^{1\gamma} & \cdots & \varsigma_{1M}^{1M} & 2\varsigma_{1M}^{11} \\ \varsigma_{21}^{11} & \varsigma_{22}^{11} + \varsigma_{12}^{21} & \cdots & \varsigma_{2\gamma}^{1\gamma} & \varsigma_{2(\gamma+1)}^{1\gamma} + \varsigma_{1(\gamma+1)}^{2\gamma} & \cdots & \varsigma_{2M}^{1M} & \varsigma_{21}^{1M} + \varsigma_{11}^{2M} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \varsigma_{\gamma 1}^{\gamma 1} & 2\varsigma_{\gamma 2}^{\gamma 1} & \cdots & \varsigma_{\gamma\gamma}^{\gamma\gamma} & 2\varsigma_{\gamma(\gamma+1)}^{\gamma\gamma} & \cdots & \varsigma_{\gamma M}^{\gamma M} & 2\varsigma_{\gamma 1}^{\gamma M} \\ \varsigma_{(\gamma+1)1}^{\gamma 1} & \varsigma_{(\gamma+1)2}^{\gamma 1} + \varsigma_{\gamma 2}^{(\gamma+1)1} & \cdots & \varsigma_{(\gamma+1)\gamma}^{\gamma\gamma} & \varsigma_{(\gamma+1)(\gamma+1)}^{\gamma\gamma} + \varsigma_{\gamma(\gamma+1)}^{(\gamma+1)\gamma} & \cdots & \varsigma_{(\gamma+1)M}^{\gamma M} & \varsigma_{(\gamma+1)1}^{\gamma M} + \varsigma_{\gamma 1}^{(\gamma+1)M} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \varsigma_{M1}^{M1} & 2\varsigma_{M2}^{M1} & \cdots & \varsigma_{M\gamma}^{M\gamma} & 2\varsigma_{M(\gamma+1)}^{M\gamma} & \cdots & \varsigma_{MM}^{MM} & 2\varsigma_{M1}^{MM} \\ \varsigma_{11}^{M1} & \varsigma_{12}^{M1} + \varsigma_{M2}^{11} & \cdots & \varsigma_{1\gamma}^{M\gamma} & \varsigma_{1(\gamma+1)}^{M\gamma} + \varsigma_{M(\gamma+1)}^{1\gamma} & \cdots & \varsigma_{1M}^{MM} & \varsigma_{11}^{MM} \end{bmatrix}^{-1} \begin{bmatrix} w_{11}^{\text{obs}}(\theta_k) \\ w_{12}^{\text{obs}}(\theta_k) \\ \vdots \\ w_{\gamma\gamma}^{\text{obs}}(\theta_k) \\ w_{\gamma(\gamma+1)}^{\text{obs}}(\theta_k) \\ \vdots \\ w_{MM}^{\text{obs}}(\theta_k) \\ w_{M1}^{\text{obs}}(\theta_k) \end{bmatrix} \tag{4.108}$$

As for the 2-sample case, we can get the variance of the estimators for M target samples as

$$\begin{aligned} &\begin{bmatrix} \sigma^2_{\widehat{w}_{A_1A_1}} & \sigma^2_{\widehat{w}_{A_1A_2}} & \dots \sigma^2_{\widehat{w}_{A_\gamma A_\gamma}} & \sigma^2_{\widehat{w}_{A_\gamma A_{\gamma+1}}} & \dots \sigma^2_{\widehat{w}_{A_M A_M}} & \sigma^2_{\widehat{w}_{A_M A_1}} \end{bmatrix}^T \\ &= \left\{ [D_{\text{S}}^{\text{gen}}]^{-1} \right\}^2_{ij} \begin{bmatrix} \sigma^2_{w^{\text{obs}}_{A_1A_1}} & \sigma^2_{w^{\text{obs}}_{A_1A_2}} & \dots \sigma^2_{w^{\text{obs}}_{A_\gamma A_\gamma}} & \sigma^2_{w^{\text{obs}}_{A_\gamma A_{\gamma+1}}} & \dots \sigma^2_{w^{\text{obs}}_{A_M A_M}} & \sigma^2_{w^{\text{obs}}_{A_M A_1}} \end{bmatrix}^T \end{aligned} \tag{4.109}$$

where $[D_{\text{S}}^{\text{gen}}]$ is the square matrix in Equation 4.108 and as in 4.A.3, $\left\{ [D_{\text{S}}^{\text{gen}}]^{-1} \right\}^2_{ij}$ denotes that matrix resulting from squaring each individual coefficient in the matrix $[D_{\text{S}}^{\text{gen}}]^{-1}$. The covariance matrix for the M-samples case follows the derivation in Equation 4.34, with all of its assumptions.

4.D.2 `Decontaminated Weighted` Estimator

Expanding our derivation for two samples to three samples, with galaxies of Types A, B, C present in our sample, we have

$$\begin{bmatrix} \widehat{w}_{AA}(\theta_k) \\ \widehat{w}_{AB}(\theta_k) \\ \widehat{w}_{BB}(\theta_k) \\ \widehat{w}_{BC}(\theta_k) \\ \widehat{w}_{CC}(\theta_k) \\ \widehat{w}_{CA}(\theta_k) \end{bmatrix} = \begin{bmatrix} \varkappa_{AA}^{AA} & 2\varkappa_{AB}^{AA} & \varkappa_{AB}^{AB} & 2\varkappa_{AC}^{AB} & \varkappa_{AC}^{AC} & 2\varkappa_{AC}^{AA} \\ \varkappa_{BA}^{AA} & \varkappa_{BB}^{AA} + \varkappa_{AB}^{BA} & \varkappa_{AB}^{BB} & \varkappa_{BC}^{AB} + \varkappa_{AC}^{BB} & \varkappa_{BC}^{AC} & \varkappa_{BC}^{AA} + \varkappa_{AC}^{BA} \\ \varkappa_{BA}^{BA} & 2\varkappa_{BA}^{BB} & \varkappa_{BB}^{BB} & 2\varkappa_{BC}^{BB} & \varkappa_{BC}^{BC} & 2\varkappa_{BC}^{BA} \\ \varkappa_{CB}^{BA} & \varkappa_{CB}^{BA} + \varkappa_{BB}^{CA} & \varkappa_{CB}^{BB} & \varkappa_{CC}^{BB} + \varkappa_{BC}^{CB} & \varkappa_{CC}^{BC} & \varkappa_{BC}^{BA} + \varkappa_{BC}^{BA} \\ \varkappa_{CA}^{CA} & 2\varkappa_{CB}^{CA} & \varkappa_{CB}^{CB} & 2\varkappa_{CC}^{CB} & \varkappa_{CC}^{CC} & 2\varkappa_{CC}^{CA} \\ \varkappa_{CA}^{AA} & \varkappa_{CB}^{AA} + \varkappa_{AB}^{CA} & \varkappa_{CB}^{AB} & \varkappa_{CC}^{AB} + \varkappa_{AC}^{CB} & \varkappa_{CC}^{AC} & \varkappa_{CC}^{AA} + \varkappa_{AC}^{CA} \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{w}_{AA}^{\rm obs}(\theta_k) \\ \widetilde{w}_{AB}^{\rm obs}(\theta_k) \\ \widetilde{w}_{BB}^{\rm obs}(\theta_k) \\ \widetilde{w}_{BC}^{\rm obs}(\theta_k) \\ \widetilde{w}_{CC}^{\rm obs}(\theta_k) \\ \widetilde{w}_{CA}^{\rm obs}(\theta_k) \end{bmatrix} \tag{4.110}$$

where we have defined the following for brevity:

$$\varkappa_{mn}^{uv} = \frac{\sum_{i}^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{A_u A_v} \mathtt{q}_i^{A_m} \mathtt{q}_j^{A_n}}{\sum_{i}^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathtt{w}_{ij}^{A_u A_v}} \tag{4.111}$$

Extending the idea to M samples, we can write the analog of our unbiased estimator for `Decontaminated Weighted`, given by Equation 4.18, as

$$\begin{bmatrix} \widehat{w}_{11}(\theta_k) \\ \widehat{w}_{12}(\theta_k) \\ \vdots \\ \widehat{w}_{\gamma\gamma}(\theta_k) \\ \widehat{w}_{\gamma(\gamma+1)}(\theta_k) \\ \vdots \\ \widehat{w}_{MM}(\theta_k) \\ \widehat{w}_{M1}(\theta_k) \end{bmatrix} = \begin{bmatrix} \varkappa_{11}^{11} & 2\varkappa_{12}^{11} & \dots & \varkappa_{1\gamma}^{1\gamma} & 2\varkappa_{1(\gamma+1)}^{1\gamma} & \dots & \varkappa_{1M}^{1M} & 2\varkappa_{1M}^{11} \\ \varkappa_{21}^{11} & \varkappa_{22}^{11} + \varkappa_{12}^{21} & \dots & \varkappa_{2\gamma}^{1\gamma} & \varkappa_{2(\gamma+1)}^{1\gamma} + \varkappa_{1(\gamma+1)}^{2\gamma} & \dots & \varkappa_{2M}^{1M} & \varkappa_{21}^{1M} + \varkappa_{11}^{2M} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots \\ \varkappa_{\gamma 1}^{\gamma 1} & 2\varkappa_{\gamma 2}^{\gamma 1} & \dots & \varkappa_{\gamma\gamma}^{\gamma\gamma} & 2\varkappa_{\gamma(\gamma+1)}^{\gamma\gamma} & \dots & \varkappa_{\gamma M}^{\gamma M} & 2\varkappa_{\gamma 1}^{\gamma M} \\ \varkappa_{(\gamma+1)1}^{\gamma 1} & \varkappa_{(\gamma+1)2}^{\gamma 1} + \varkappa_{\gamma 2}^{(\gamma+1)1} & \dots & \varkappa_{(\gamma+1)\gamma}^{\gamma\gamma} & \varkappa_{(\gamma+1)(\gamma+1)}^{\gamma\gamma} + \varkappa_{\gamma(\gamma+1)}^{(\gamma+1)\gamma} & \dots & \varkappa_{(\gamma+1)M}^{\gamma M} & \varkappa_{(\gamma+1)1}^{\gamma M} + \varkappa_{\gamma 1}^{(\gamma+1)M} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots \\ \varkappa_{M1}^{M1} & 2\varkappa_{M2}^{M1} & \dots & \varkappa_{M\gamma}^{M\gamma} & 2\varkappa_{M(\gamma+1)}^{M\gamma} & \dots & \varkappa_{MM}^{MM} & 2\varkappa_{M1}^{MM} \\ \varkappa_{11}^{M1} & \varkappa_{12}^{M1} + \varkappa_{M2}^{11} & \dots & \varkappa_{1\gamma}^{M\gamma} & \varkappa_{1(\gamma+1)}^{M\gamma} + \varkappa_{M(\gamma+1)}^{1\gamma} & \dots & \varkappa_{1M}^{MM} & \varkappa_{11}^{MM} \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{w}_{11}^{\rm obs}(\theta_k) \\ \widetilde{w}_{12}^{\rm obs}(\theta_k) \\ \vdots \\ \widetilde{w}_{\gamma\gamma}^{\rm obs}(\theta_k) \\ \widetilde{w}_{\gamma(\gamma+1)}^{\rm obs}(\theta_k) \\ \vdots \\ \widetilde{w}_{MM}^{\rm obs}(\theta_k) \\ \widetilde{w}_{M1}^{\rm obs}(\theta_k) \end{bmatrix} \tag{4.112}$$

Chapter 5

Summary & Future Work

Studying the evolution of the large scale structure (LSS) in the universe is a strong probe of dark energy – the leading theory to explain the cosmic acceleration. As discussed in this thesis, LSS studies require some important features: survey uniformity over an area that maximizes the usable survey footprint with deep photometric data. Focusing on one of the largest surveys of the next decade – the Legacy Survey of Space and Time (LSST) – we studied the first feature in Chapter 2 and explored the impacts of translational dithers in increasing survey uniformity; while the latter is discussed in Chapter 3 where we investigated the survey area resulting from selection cuts needed to yield deep photometric galaxy samples. We also studied the impacts of redshift contamination in Chapter 4 and developed tools to make use of all the probabilistic distance information available for each photometric galaxy when measuring the two-point correlation function.

While there are various avenues for further development to understand and handle systematics in cosmological analyses, there are some direct next-steps:

5.1 Application and Optimization of the New Estimators to Simulated and Pre-Cursor LSST Data

As discussed in 4.6, the new estimator for galaxy correlation functions (CFs) can be optimized to yield low-variance measurements of cosmological parameters. This can be achieved specifically using the outputs of the Data Challenges (DCs) from the LSST Dark Energy Science Collaborations (DESC) – organized efforts for an end-to-end simulation of the LSST-like data, aimed at testing and validating various analysis pipelines (The LSST Dark Energy Science Collaboration et al., 2018). Specifically, DC1 simulation yielded 40 deg^2 in just one photometric filter (Sánchez et al., 2020); Figure 5.1 shows the DC1 survey footprint with and without translational dithering. DC2 simulations (LSST DESC et al., 2020, in prep) are currently underway[1],

[1]DC2 image simulation is expected to be completed by June 2020, based on plans in the LSST DESC Science Roadmap.

providing us with 300 deg^2 in all six LSST photometric filters; Figure 5.2 shows the footprint of the simulation. Both these simulations have used translationally dithered visits, deriving from our work in Chapter 2, as well as rotational dithers (based on preliminary dithering scheme discussed in Section 5.3).

DC2 is lucrative for the optimization of the new CF estimators since we will have access to the truth catalog corresponding to the photometric catalog, allowing a robust test of the method and an easy platform for a direct comparison of results with those from forward modeling. The optimized estimator can then be applied to data from Hyper Suprime-Cam (HSC) (Aihara et al., 2018) as well as the Dark Energy Survey (DES) (Flaugher, 2005) to obtain improved constraints on cosmological parameters.

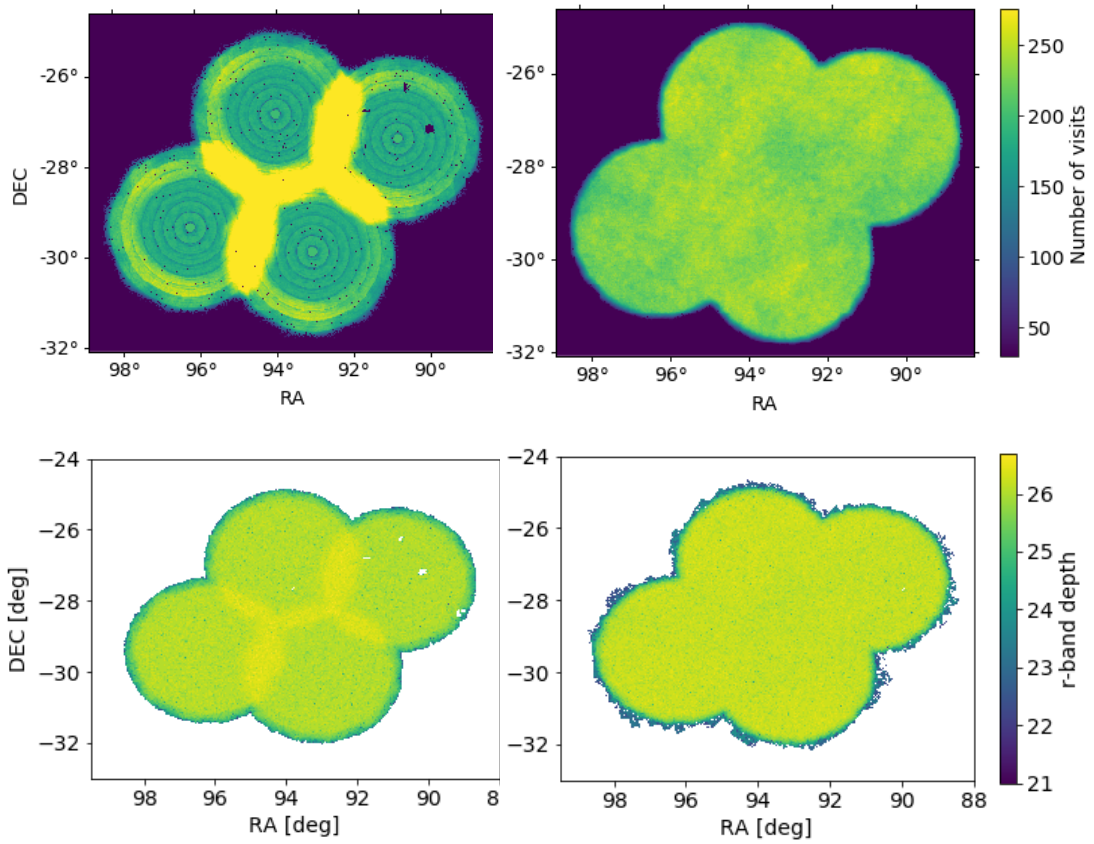

Figure 5.1: Figure adapted from Figures 16, 20, 21 in Sánchez et al. (2020) (with permission), showing the number of visits to the DC1 survey area in the top row when translationally undithered (left) and dithered (right), leading to the survey depth as shown in the corresponding panels in the bottom row. We see that as expected, translational dithering leads to a more uniform coverage, as it redistributes the visits to the overlapping regions of the undithered fields to the rest of the survey.

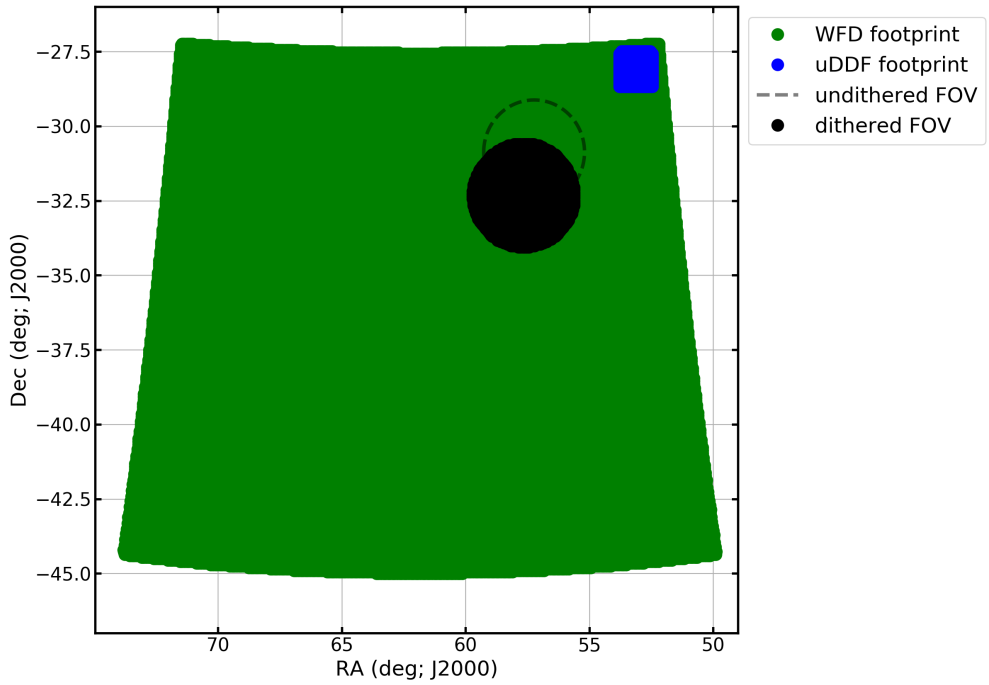

Figure 5.2: Figure adapted from LSST DESC et al., 2020, in prep., showing the footprint of the simulated data from the LSST-DESC Data Challenge 2, yielding 300 deg^2 of the wide-fast-deep survey area (green) along side a 25 deg^2 deep drilling field (blue); figure included here with permission. The simulated visits are translationally dithered.

5.2 Impacts of LSS Systematics on Cosmological Parameter Estimation

Taking our work in Chapters 2-3, we can not only quantify the uncertainties/biases induced in the power spectra but also those induced in the cosmological parameters. This can be done by creating a modular framework to assess the impacts of systematics induced by the observing strategy and Milky Way dust uncertainties. Figure 5.3 shows the workflow for this framework. Taking the simulated cadences and the resulting survey statistics, the pipeline built for the analysis Chapter 2 yields the fluctuations in galaxy density resulting from the two systematics; this can then be combined with theoretical power spectra given a specified cosmology using the `Core Cosmology Library`, leading us to the observed galaxy density maps. Then, using `NaMaster` – a fast power spectrum calculation code – and `LSSLike` – the cosmological parameters likelihood code, we can quantify the impacts of the systematic uncertainties and determine the need for rigorous mitigation schemes to counter the added systematic uncertainties. Note that this can also be applied to the simulated data from DESC DC2, discussed in Section 5.1, to test the current systematics mitigation schemes.

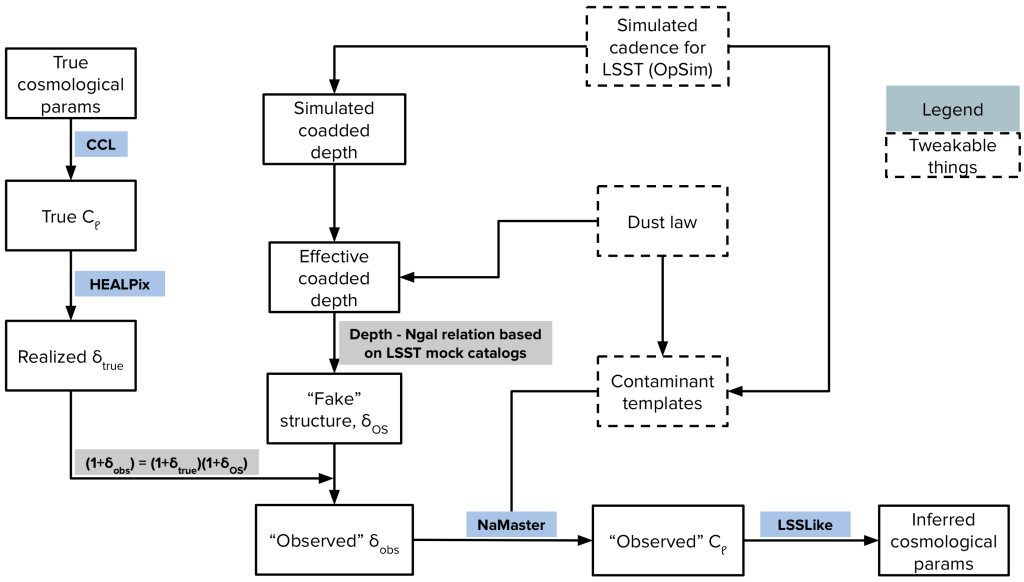

Figure 5.3: Workflow for the modular framework that would allow us to quantify the impacts of the systematic uncertainties on cosmological parameter estimations using galaxy power spectra.

5.3 Rotational Uniformity

Just as translational uniformity of telescope coverage is critical for LSS studies with LSST, rotational uniformity is important for Weak Lensing studies and hence for cosmological analyses with photometric surveys like the LSST: the standard tool for competitive constraints from joint dark energy probes (e.g., Krause et al., 2017) is the 3x2pt analysis which entails joint constraints from three two-point correlations – galaxy-galaxy, galaxy-shear, and shear-shear. To address rotational uniformity, we investigate the impacts of rotational dithers, with the goal of making the distribution of rotational angles more uniform. Figure 5.4 shows an example result, demonstrating that rotational dithers would be critical for ensuring uniform rotational distributions.

5.4 Impacts of Milky Way Dust Uncertainties on CMB Lensing $\times$ LSS Studies

Cross-correlations between CMB lensing and LSS provides yet another probe to study cosmology but suffers from a correlated systematic arising from the Milky Way dust: Galactic dust leads to extinction in observation of extragalactic sources alongside contributing foreground emission that affects CMB lensing. Given the infrastructure set up for work discussed in Section 5.2, we can investigate the impacts of Milky Way dust on the cross-correlation, specifically

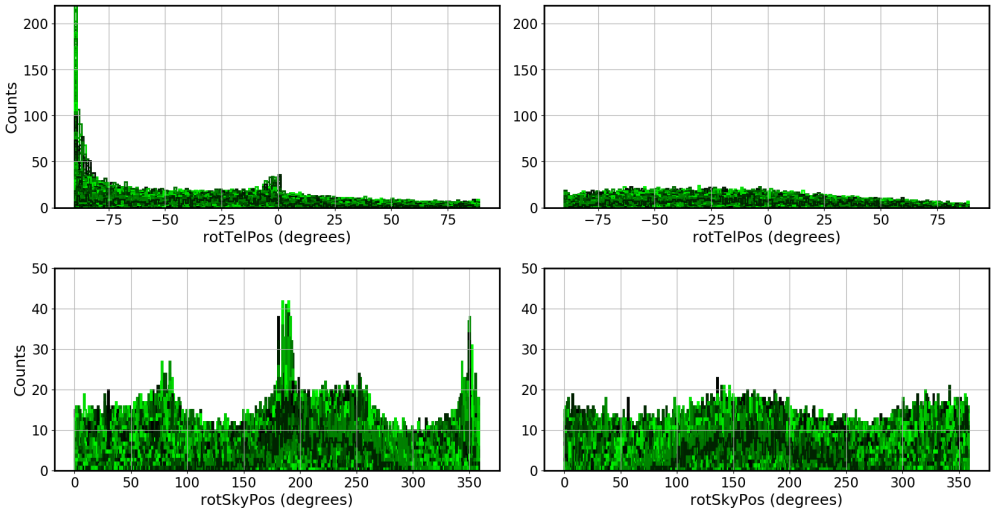

Figure 5.4: Distribution of two rotational angles with and without rotational dithers for all the pointings of the survey. We see that our simple rotational dithering scheme makes the distribution of the rotational angles more uniform, with room for further improvement.

focusing on CMB observations from the Simons Observatory and LSS from the LSST, while taking into account survey-specific systematics and specifications. Figure 5.5 shows some preliminary results: the residual power spectra for the lensing-galaxy density cross-correlation without any systematics (blue), with observing strategy systematics in the galaxy density field (orange), and with both observing strategy systematics in the galaxy density field and dust systematics in lensing and galaxy density fields (black). We see a modest bias in the power spectrum due to the dust systematics, and the next step is to quantify the impacts of the residual uncertainties on the cosmological parameters.

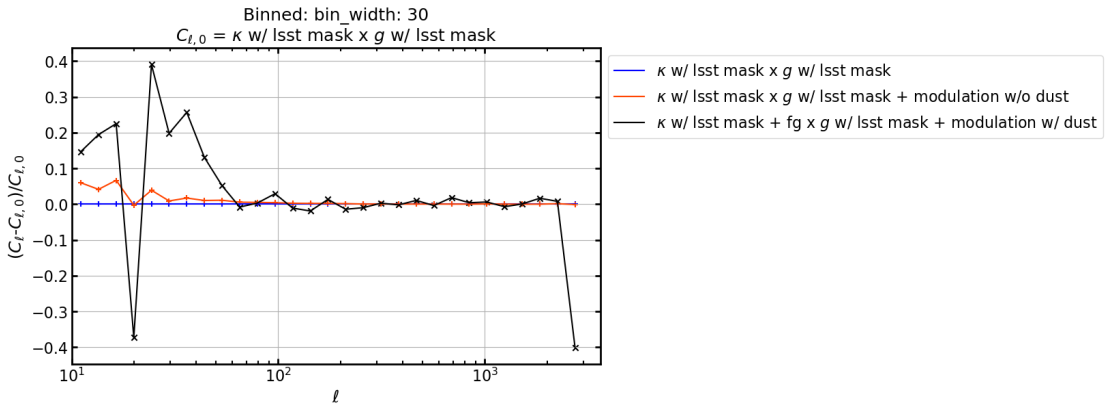

Figure 5.5: Residual power spectra for the lensing-galaxy density cross correlation without any systematics (blue), with observing strategy systematics in the galaxy density field (orange), and with both observing strategy systematics in the galaxy density field and dust systematics in lensing and galaxy density fields (black).

5.5 Conclusions

With various large astronomical surveys coming online in the next decade, it is imperative to understand and reduce the dominant source of uncertainties in our measurements of the cosmological parameters in the Big Data cosmology era: systematic uncertainties. Only with an enhanced understanding of the sources of systematic uncertainties and better mitigation of their impacts can we expect to utilize the statistical power of the large datasets. In this thesis, we focused on three specific sources of systematic unceratainties.

In Chapter 2, we considered the impact of the telescope observing strategy for the Legacy Survey of Space and Time (LSST) and demonstrated that fixed telescope pointings lead to strong artifacts that induce large systematic uncertainties at the scale of interest for cosmology. To mitigate these artifacts, we considered translational dithers and showed them to be effective in reducing the added systematic uncertainties to be subdominant to the statistical uncertainties – the regime where we can truly utilize the statistical power of the large datasets. Our recommendations have now been incorporated into the baseline LSST observing strategy, as the LSST scheduler no longer uses a static gridding of the sky but one that randomly changes every night, effectively implementing random per night translational dithers; the re-gridding also removes the shallow borders in the survey area that resulted from translational dithering on a fixed grid of the sky. We plan to extend our investigation to rotational dithers, as discussed in Section 5.3, in order to ensure rotational uniformity that will be critical for photometric calibration as well as Weak Lensing systematics.

In Chapter 3, we studied the impacts of Milky Way dust extinction and demonstrate that nearly 25% of the default LSST survey area does not pass the selection cuts needed to yield a deep photometric sample. To address this, we proposed a reconfiguration of the LSST survey area to avoid high extinction regions of the sky. We also considered the evolution of the median depth at various points in the 10-year survey from different simulations, and demonstrate that it is critical to pay attention to not just Y1, Y10 statistics but also the intermediate years to better assess the optimization of the survey.

Considering Milky Way dust uncertainties, we can incorporate them into the joint-probes analysis as discussed in Section 5.2; this will allow us to not only figure out the dominant source of systematic uncertainties but fully incorporate their impacts in the planned analysis pipelines. We can also consider the impact of dust uncertainties on the cross-correlation between CMB lensing and LSS, as discussed in Section 5.4.

In Chapter 4, we focused on the two-point angular correlation estimators that are used

to quantify galaxy clustering. We discussed how the standard estimators do not handle distance uncertainties directly, and hence are limited in utilizing the statistical power of the large datasets. We presented a formalism to correct for any kind of contamination and a new weighted estimator that incorporates the full redshift probabilistic density function for each galaxy when measuring the correlation function. We demonstrated that our estimators lead to smaller estimator covariances, and recover the true correlation functions and their covariance matrices. Applying the weighted estimator to simulated data, as discussed in Section 5.1, we can compare our results directly with those resulting from forward modeling the redshift contamination as well as optimized the estimator to apply to real data.

The work presented here is only a step in the right direction. In order to fully realize the statistical power of the large datasets, it is critical to reframe our treatment of data: to fully accept the uncertainties associated with our measurements and to develop tools that incorporate them more carefully into our analyses. With a better understanding of the systematic uncertainties and new tools to mitigate and address their impacts – the focus of this work – we can ensure robust measurements of dark energy using large surveys like the LSST and unveil the nature of our universe.

Bibliography

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., et al. 2018, Physical Review D, 98, 043526

Abbott, T. M. C., et al. 2019, Mon. Not. Roy. Astron. Soc., 483, 4866

Addison, G. E., Bennett, C. L., Jeong, D., Komatsu, E., & Weiland, J. L. 2019, Astrophysical Journal, 879, 15

Aihara, H., Armstrong, R., Bickerton, S., et al. 2018, Publications of the Astronomical Society of Japan, 70, S8

Albrecht, A., Bernstein, G., Cahn, R., et al. 2006, Report of the Dark Energy Task Force, arXiv:astro-ph/0609591

Armijo, J., Cai, Y.-C., Padilla, N., Li, B., & Peacock, J. A. 2018, Monthly Notices of the Royal Astronomical Society, 478, 3627

Asorey, J., Carrasco Kind, M., Sevilla-Noarbe, I., Brunner, R. J., & Thaler, J. 2016, Monthly Notices of the Royal Astronomical Society, 459, 1293

Awan, H., & Gawiser, E. 2020, The Astrophysical Journal, 890, 78

Awan, H., Gawiser, E., Kurczynski, P., et al. 2016, Astrophysical Journal, 829, 50

Bailoni, A., Spurio Mancini, A., & Amendola, L. 2017, Monthly Notices of the Royal Astronomical Society, 470, 688

Balaguera-Antolínez, A., Bilicki, M., Branchini, E., & Postiglione, A. 2018, Monthly Notices of the Royal Astronomical Society, 476, 1050

Barboza, E. M., & Alcaniz, J. S. 2008, Physics Letters B, 666, 415

Beisbart, C., & Kerscher, M. 2000, Astrophysical Journal, 545, 6

Benjamin, J., van Waerbeke, L., Ménard, B., & Kilbinger, M. 2010, Monthly Notices of the Royal Astronomical Society, 408, 1168

Bernstein, G. M. 1994, Astrophysical Journal, 424, 569

Bianchi, D., & Percival, W. J. 2017, Monthly Notices of the Royal Astronomical Society, 472, 1106

Bianchi, D., Burden, A., Percival, W. J., et al. 2018, Monthly Notices of the Royal Astronomical Society, 481, 2338

Blake, C., Achitouv, I., Burden, A., & Rasera, Y. 2019, Monthly Notices of the Royal Astronomical Society, 482, 578

Branch, D., & Tammann, G. A. 1992, Annual Review of Astronomy and Astrophysics, 30, 359

Bunn, E. F., & Hogg, D. W. 2009, American Journal of Physics, 77, 688

Camacho, H., Kokron, N., Andrade-Oliveira, F., et al. 2019, Monthly Notices of the Royal Astronomical Society, 487, 3870

Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., & Fosalba, P. 2015, Monthly Notices of the Royal Astronomical Society, 447, 646

Carretero, J., Tallada, P., Casals, J., et al. 2017, PoS, EPS-HEP2017, 488

Carroll, C. M., Gawiser, E., Kurczynski, P. L., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9149, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series

Carroll, S. M. 2001, Living Reviews in Relativity, 4, 1

Carroll, S. M. 2019, Spacetime and Geometry: An Introduction to General Relativity (Cambridge University Press), doi:10.1017/9781108770385

Chaves-Montero, J., Angulo, R. E., & Hernández-Monteagudo, C. 2018, Monthly Notices of the Royal Astronomical Society, 477, 3892

Connolly, A. J., Scranton, R., Johnston, D., et al. 2002, Astrophysical Journal, 579, 42

Cooray, A., & Sheth, R. 2002, Physics Reports, 372, 1

Cora, S. A. 2006, Monthly Notices of the Royal Astronomical Society, 368, 1540

Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., & Carretero, J. 2015, Monthly Notices of the Royal Astronomical Society, 453, 1513

Crocce, M., Carretero, J., Bauer, A. H., et al. 2016, Monthly Notices of the Royal Astronomical Society, 455, 4301

de Jong, R. 2011, The Messenger, 145, 14

Delgado, F., Saha, A., Chandrasekharan, S., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9150, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 15

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv e-prints, arXiv:1611.00036

Dodelson, S. 2003, Modern Cosmology, Academic Press (Academic Press)

Eisenstein, D. J., & Hu, W. 1998, Astrophysical Journal, 496, 605

Eisenstein, D. J., Hu, W., Silk, J., & Szalay, A. S. 1998a, Astrophysical Journal, Letters, 494, L1

Eisenstein, D. J., Hu, W., & Tegmark, M. 1998b, Astrophysical Journal, Letters, 504, L57

Eisenstein, D. J., Seo, H.-J., & White, M. 2007, Astrophysical Journal, 664, 660

Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, Astrophysical Journal, 633, 560

Elsner, F., Leistedt, B., & Peiris, H. V. 2016, Monthly Notices of the Royal Astronomical Society, 456, 2095

Fall, S. M. 1978, Monthly Notices of the Royal Astronomical Society, 185, 165

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, Astrophysical Journal, 426, 23

Flaugher, B. 2005, International Journal of Modern Physics A, 20, 3121

Fleming, D. E. B., Harris, W. E., Pritchet, C. J., & Hanes, D. A. 1995, Astronomical Journal, 109, 1044

Fosalba, P., Crocce, M., Gaztañaga, E., & Castand er, F. J. 2015a, Monthly Notices of the Royal Astronomical Society, 448, 2987

Fosalba, P., Gaztañaga, E., Castander, F. J., & Crocce, M. 2015b, Monthly Notices of the Royal Astronomical Society, 447, 1319

Gargiulo, I. D., Cora, S. A., Padilla, N. D., et al. 2015, Monthly Notices of the Royal Astronomical Society, 446, 3820

Gawiser, E., van Dokkum, P. G., Herrera, D., et al. 2006, The Astrophysical Journal Supplement Series, 162, 1

Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, Astrophysical Journal, 622, 759

Grasshorn Gebhardt, H. S., Jeong, D., Awan, H., et al. 2019, Astrophysical Journal, 876, 32

Gwyn, S. D. J. 2008, Publications of the Astronomical Society of the Pacific, 120, 212

Hamilton, A. J. S. 1998, in Astrophysics and Space Science Library, Vol. 231, The Evolving Universe, ed. D. Hamilton, 185

Harker, G., Cole, S., Helly, J., Frenk, C., & Jenkins, A. 2006, Monthly Notices of the Royal Astronomical Society, 367, 1039

Hernández-Aguayo, C., Baugh, C. M., & Li, B. 2018, Monthly Notices of the Royal Astronomical Society, 479, 4824

Hill, G. J., Gebhardt, K., Komatsu, E., et al. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 399, Panoramic Views of Galaxy Formation and Evolution, ed. T. Kodama, T. Yamada, & K. Aoki, 115

Hoekstra, H., Mellier, Y., van Waerbeke, L., et al. 2006, Astrophysical Journal, 647, 116

Hoffmann, K., Bel, J., Gaztañaga, E., et al. 2015, Monthly Notices of the Royal Astronomical Society, 447, 1724

Holmes, R., Hogg, D. W., & Rix, H.-W. 2012, Publications of the Astronomical Society of the Pacific, 124, 1219

Hubble, E. 1929, Proceedings of the National Academy of Science, 15, 168

Huterer, D., Cunha, C. E., & Fang, W. 2013, Monthly Notices of the Royal Astronomical Society, 432, 2945

Huterer, D., Knox, L., & Nichol, R. C. 2001, Astrophysical Journal, 555, 547

Ivezic, Z., Tyson, J., Acosta, E., et al. 2008, arXiv preprint arXiv:0805.2366

Jain, B., & Zhang, P. 2008, Physical Review D, 78, 063503

Jarosik, N., Bennett, C. L., Dunkley, J., et al. 2011, Astrophysical Journal, Supplement, 192, 14

Jarvis, M., Bernstein, G., & Jain, B. 2004, Monthly Notices of the Royal Astronomical Society, 352, 338

Jing, Y. P. 2005, Astrophysical Journal, 620, 559

Jones, R. L., Yoachim, P., Chandrasekharan, S., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9149, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 0

Kazin, E. A., Sánchez, A. G., & Blanton, M. R. 2012, Monthly Notices of the Royal Astronomical Society, 419, 3223

Kerscher, M., Szapudi, I., & Szalay, A. S. 2000, The Astrophysical Journal, 535, L13

Koehler, R. 2009, PhD thesis, Ludwig-Maximilians-Universität

Krause, E., Eifler, T. F., Zuntz, J., et al. 2017, arXiv e-prints, arXiv:1706.09359

Krughoff, K. 2016, Hexagonal Dithering for LSST, doi:10.5281/zenodo.55701

Lagos, C. D. P., Cora, S. A., & Padilla, N. D. 2008, Monthly Notices of the Royal Astronomical Society, 388, 587

Landy, S. D., & Szalay, A. S. 1993, Astrophysical Journal, 412, 64

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193

Leistedt, B., Peiris, H. V., Elsner, F., et al. 2016, Astrophysical Journal, Supplement, 226, 24

Leistedt, B., Peiris, H. V., Elsner, F., et al. 2016, The Astrophysical Journal Supplement Series, 226, 24

Leung, A. S., Acquaviva, V., Gawiser, E., et al. 2017, Astrophysical Journal, 843, 130

Lochner, M., Scolnic, D. M., Awan, H., et al. 2018, ArXiv e-prints, arXiv:1812.00515

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv e-prints, arXiv:0912.0201

LSST Science Collaboration, Marshall, P., Anguita, T., et al. 2017, arXiv e-prints, arXiv:1708.04058

McLean, I. 2008, Electronic Imaging in Astronomy: Detectors and Instrumentation (Springer)

Mo, H., van den Bosch, F., & White, S. 2010, Galaxy Formation and Evolution, Galaxy Formation and Evolution (Cambridge University Press)

Morrison, C. B., & Hildebrandt, H. 2015, Monthly Notices of the Royal Astronomical Society, 454, 3121

Muñoz, R. P., Puzia, T. H., Lançon, A., et al. 2014, Astrophysical Journal, Supplement, 210, 4

Muñoz Arancibia, A. M., Navarrete, F. P., Padilla, N. D., et al. 2015, Monthly Notices of the Royal Astronomical Society, 446, 2291

Newman, J. A. 2008, Astrophysical Journal, 684, 88

Nicola, A., Alonso, D., Sánchez, J., et al. 2019, arXiv e-prints, arXiv:1912.08209

Noh, Y. 2013, PhD thesis, UC Berkeley: Astrophysics

Olsen, K., Di Criscienzo, M., Jones, R. L., et al. 2018, arXiv e-prints, arXiv:1812.02204

Orsi, Á., Padilla, N., Groves, B., et al. 2014, Monthly Notices of the Royal Astronomical Society, 443, 799

Peacock, J. A., Percival, W. J., & Verde, L. 2004, Monthly Notices of the Royal Astronomical Society, 347, 645

Pearson, D. W., Samushia, L., & Gagrani, P. 2016, Monthly Notices of the Royal Astronomical Society, 463, 2708

Peebles, P. 1993, Principles of Physical Cosmology, Princeton series in physics (Princeton University Press)

Peebles, P. J., & Ratra, B. 2003, Reviews of Modern Physics, 75, 559

Penzias, A. A., & Wilson, R. W. 1965, Astrophysical Journal, 142, 419

Percival, W. J., & Bianchi, D. 2017, Monthly Notices of the Royal Astronomical Society, 472, L40

Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, Astrophysical Journal, 517, 565

Phillips, M. M., Lira, P., Suntzeff, N. B., et al. 1999, The Astronomical Journal, 118, 1766

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, Astronomy and Astrophysics, 594, A13

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, Astronomical Journal, 116, 1009

Riess, A. G., Macri, L. M., Hoffmann, S. L., et al. 2016, Astrophysical Journal, 826, 56

Robaina, A. R., & Bell, E. F. 2012, Monthly Notices of the Royal Astronomical Society, 427, 901

Ross, A. J., Percival, W. J., Sánchez, A. G., et al. 2012, Monthly Notices of the Royal Astronomical Society, 424, 564

Ross, A. J., Banik, N., Avila, S., et al. 2017, Monthly Notices of the Royal Astronomical Society, 472, 4456

Rubin, V. C. 1965, Astrophysical Journal, 142, 934

Rubin, V. C., Burley, J., Kiasatpoor, A., et al. 1962, Astronomical Journal, 67, 491

Ryden, B. 2003, Introduction to cosmology (Addison-Wesley)

Sahni, V. 2002, Classical and Quantum Gravity, 19, 3435

Sato, T., Hütsi, G., Nakamura, G., & Yamamoto, K. 2013, International Journal of Astronomy and Astrophysics, 3, 243

Schlafly, E. F., & Finkbeiner, D. P. 2011, Astrophysical Journal, 737, 103

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, Astrophysical Journal, 500, 525

Scranton, R., Johnston, D., Dodelson, S., et al. 2002, Astrophysical Journal, 579, 48

Shafer, D. L., & Huterer, D. 2015a, Monthly Notices of the Royal Astronomical Society, 447, 2961

—. 2015b, Monthly Notices of the Royal Astronomical Society, 447, 2961

Sheth, R. K., Connolly, A. J., & Skibba, R. 2005, arXiv e-prints, astro

Sheth, R. K., & Tormen, G. 2004, Monthly Notices of the Royal Astronomical Society, 350, 1385

Skibba, R., Sheth, R. K., Connolly, A. J., & Scranton, R. 2006, Monthly Notices of the Royal Astronomical Society, 369, 68

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757

Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, Monthly Notices of the Royal Astronomical Society, 328, 726

Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Nature, 435, 629

Sánchez, F. J., Walter, C. W., Awan, H., et al. 2020, The LSST DESC Data Challenge 1: Generation and Analysis of Synthetic Images for Next Generation Surveys, arXiv:2001.00941

Tecce, T. E., Cora, S. A., Tissera, P. B., Abadi, M. G., & Lagos, C. D. P. 2010, Monthly Notices of the Royal Astronomical Society, 408, 2008

The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., et al. 2018, ArXiv e-prints, arXiv:1809.01669

Vargas-Magaña, M., Bautista, J. E., Hamilton, J. C., et al. 2013, Astronomy and Astrophysics, 554, A131

Villalobos, J. J., Parashar, M., Rodero, I., & Brennan-Tonetta, M. 2018, High Performance Computing at the Rutgers Discovery Informatics Institute, doi:10.13140/RG.2.2.11579.87846

White, M. 2016, Journal of Cosmology and Astroparticle Physics, 2016, 057

White, M., & Padmanabhan, N. 2009, Monthly Notices of the Royal Astronomical Society, 395, 2381

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, Astronomical Journal, 120, 1579

Zehavi, I., Blanton, M. R., Frieman, J. A., et al. 2002, The Astrophysical Journal, 571, 172

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, Astrophysical Journal, 736, 59

Zhan, H. 2006, Journal of Cosmology and Astroparticle Physics, 2006, 008

Zhu, F., Padmanabhan, N., & White, M. 2015, Monthly Notices of the Royal Astronomical Society, 451, 236