The Astrophysical Journal, 890:78 (32pp), 2020 February 10 https://doi.org/10.3847/1538-4357/ab63c8

Angular Correlation Function Estimators Accounting for Contamination from Probabilistic Distance Measurements

Humna Awan[1] and Eric Gawiser[1,2]
[1] Department of Physics & Astronomy, Rutgers University, 136 Frelinghuysen Rd., Piscataway, NJ 08554, USA; awan@physics.rutgers.edu
[2] Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave., New York, NY 10010, USA
Received 2019 July 5; revised 2019 December 1; accepted 2019 December 17; published 2020 February 13

Abstract

With the advent of surveys containing millions to billions of galaxies, it is imperative to develop analysis techniques that utilize the available statistical power. In galaxy clustering, even small sample contamination arising from distance uncertainties can lead to large artifacts, which the standard estimator for two-point correlation functions does not account for. We first introduce a formalism, termed decontamination, that corrects for sample contamination by utilizing the observed cross-correlations in the contaminated samples; this corrects any correlation function estimator for contamination. Using this formalism, we present a new estimator that uses the standard estimator to measure correlation functions in the contaminated samples but then corrects for contamination. We also introduce a weighted estimator that assigns each galaxy a weight in each redshift bin based on its probability of being in that bin. We demonstrate that these estimators effectively recover the true correlation functions and their covariance matrices. Our estimators can correct for sample contamination caused by misclassification between object types as well as photometric redshifts; they should be particularly helpful for studies of galaxy evolution and baryonic acoustic oscillations, where forward modeling the clustering signal using the contaminated redshift distribution is undesirable.

Unified Astronomy Thesaurus concepts: Large-scale structure of the universe (902); Cosmology (343); Two-point correlation function (1951)

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 Large Synoptic Survey Telescope (LSST)[3] (LSST Science Collaboration et al. 2009), Dark Energy Spectroscopic Instrument[4] (DESI Collaboration et al. 2016), *Euclid*[5] (Laureijs et al. 2011), and *WFIRST*[6] (Spergel et al. 2015). The large data sets, 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 (2015) 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 examples of 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 (e.g., Sheth & Tormen 2004; Sheth et al. 2005; Harker et al. 2006; Skibba et al. 2006; White & Padmanabhan 2009; Robaina & Bell 2012; White 2016; Hernández-Aguayo et al. 2018), extract characteristic-dependent correlations (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 ratio (S/N) 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 fiber 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 also to improve clustering measurements.

[3] https://www.lsst.org/
[4] https://www.desi.lbl.gov/
[5] http://sci.esa.int/euclid/
[6] https://wfirst.gsfc.nasa.gov/

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., Zehavi et al. 2002, 2011) or the ACFs in redshift bins (e.g., Crocce et al. 2016; Abbott et al. 2018; Balaguera-Antolínez et al. 2018). Note that one can also measure the ACFs without the tomographic binning (e.g., 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; see, e.g., Bailoni et al. (2017) for a study on the impacts of bin cross-correlations on cosmological parameters. These uncertainties also smear out valuable cosmological information, including the BAO (e.g., 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; Abbott et al. 2018; Balaguera-Antolínez et al. 2018); see also, 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. Next, 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 data set, 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 2, we formally introduce the ACF and its standard estimator. In Section 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, 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 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 6, and conclude in Section 7.

2. The 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):

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

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:

$$w_{\alpha\beta}(\theta_k) = \frac{(DD)_{\alpha\beta}(\theta_k)}{(RR)_{\alpha\beta}(\theta_k)} - 1, \tag{2}$$

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 (2) leads to an autocorrelation 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

$$\Theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geqslant 0. \end{cases} \tag{3}$$

For instance, for the autocorrelation, we have

$$\begin{aligned} &(DD)_{11}(\theta_k) \\ &= \frac{\sum_i^{N_1}\sum_{j>i}^{N_1} \Theta(\theta_{ij} - \theta_{\min,k})[1 - \Theta(\theta_{ij} - \theta_{\max,k})]}{\sum_i^{N_1}\sum_{j>i}^{N_1}} \\ &\equiv \frac{\sum_i^{N_1}\sum_{j>i}^{N_1} \bar{\Theta}_{ij,k}}{\sum_i^{N_1}\sum_{j>i}^{N_1}} = \frac{\sum_i^{N_1}\sum_{j\neq i}^{N_1} \bar{\Theta}_{ij,k}}{\sum_i^{N_1}\sum_{j\neq i}^{N_1}}, \end{aligned} \tag{4}$$

where

$$\bar{\Theta}_{ij,k} \equiv \Theta(\theta_{ij} - \theta_{\min,k})[1 - \Theta(\theta_{ij} - \theta_{\max,k})]. \tag{5}$$

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 $\theta_{\min,k} \leqslant \theta_{ij} < \theta_{\max,k}$. 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), we can write the histogram for the cross-correlation function as

$$(DD)_{12}(\theta_k) = \frac{\sum_i^{N_1}\sum_j^{N_2}\bar{\Theta}_{ij,k}}{\sum_i^{N_1}\sum_j^{N_2}}, \tag{6}$$

where sample α contains N_α galaxies.

We note here that the estimator in Equation (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) and (48) in LS93). We refer to the DD/RR estimator as the `Standard` estimator, when comparing with the new estimators.

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 of which 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, q_i^A, or Type-B, 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 $q_i^A + q_i^B = 1$, 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 two-sample case, the observed Type-A and Type-B galaxies.[7] Our classifier then 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

$$q_i^{AA} + q_i^{AB} = q_j^{BA} + q_j^{BB} = 1, \tag{7}$$

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

$$f_{AA} = \langle q_i^{AA}\rangle; f_{AB} = \langle q_i^{AB}\rangle, \tag{8}$$

where the average is over the observed Type-A sample. Equation (7) translates into the expected identities on the fractions:

$$f_{AA} + f_{AB} = f_{BA} + f_{BB} = 1. \tag{9}$$

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 $q_{A_1}, q_{A_2}, \ldots, q_{A_M}$. 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 $q_i^{A_jA_m}$ and the fraction of observed Type-A_j galaxies that are Type-A_m galaxies given by $f_{A_jA_m}$.

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 (8); see, e.g., Grasshorn Gebhardt et al. (2018) and Addison et al. (2018) 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 D.1.

Since we have two types of galaxies, we aim to calculate two autocorrelations 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 $w_{AA}^{\rm obs}(\theta_k)$, $w_{AB}^{\rm obs}(\theta_k)$, $w_{BB}^{\rm obs}(\theta_k)$, which differ from the true correlations due to sample contamination. To construct the relation between the two, let us 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

$$\begin{aligned} w_{AB}^{\rm obs}(\theta_k) = {} & f_{AA}f_{BA}w_{AA}^{\rm true}(\theta_k) + \{f_{AA}f_{BB} + f_{BA}f_{AB}\} \\ & \times w_{AB}^{\rm true}(\theta_k) + f_{AB}f_{BB}w_{BB}^{\rm true}(\theta_k). \end{aligned} \tag{10}$$

The autocorrelations follow similarly, leading us to

$$\begin{bmatrix} w_{AA}^{\rm obs}(\theta_k) \\ w_{AB}^{\rm obs}(\theta_k) \\ w_{BB}^{\rm obs}(\theta_k) \end{bmatrix} = \begin{bmatrix} f_{AA}^2 & 2f_{AA}f_{AB} & f_{AB}^2 \\ f_{AA}f_{BA} & f_{AA}f_{BB} + f_{AB}f_{BA} & f_{AB}f_{BB} \\ f_{BA}^2 & 2f_{BB}f_{BA} & f_{BB}^2 \end{bmatrix} \begin{bmatrix} w_{AA}^{\rm true}(\theta_k) \\ w_{AB}^{\rm true}(\theta_k) \\ w_{BB}^{\rm true}(\theta_k) \end{bmatrix}, \tag{11}$$

where we note that the contribution from the true cross-correlation to the observed autocorrelations simplifies (as opposed for that to the observed cross-correlation). We also present a formal derivation of the result above using Equation (1) in Appendix A.1. Now, using these equations, we can construct the `Decontaminated` estimators

[7] A simple way to do this would be to assign all galaxies with $q_i^A > 0.5$ to target sample A and the rest to target sample B.

$\widehat{w}_{AA}(\theta_k)$, $\widehat{w}_{BB}(\theta_k)$, $\widehat{w}_{AB}(\theta_k)$ 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

$$[\widehat{w}_{AA}(\theta_k)\ \ \widehat{w}_{AB}(\theta_k)\ \ \widehat{w}_{BB}(\theta_k)]^T = [D_{\rm S}]^{-1}[w_{AA}^{\rm obs}(\theta_k)\ \ w_{AB}^{\rm obs}(\theta_k)\ \ w_{BB}^{\rm obs}(\theta_k)]^T, \quad (12)$$

where $[D_{\rm S}]$ is the square matrix in Equation (11), which must be invertible.[8] Appendix 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); see Appendix 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 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. 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:

$$\widetilde{w}_{\alpha\beta}^{\rm obs}(\theta_k) = \frac{(\widetilde{DD})_{\alpha\beta}(\theta_k)}{RR(\theta_k)} - 1, \quad (13)$$

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

$$(\widetilde{DD})_{\alpha\beta}(\theta_k) = \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}\bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}}, \quad (14)$$

where $\mathrm{w}_{ij}^{\alpha\beta}$ 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), (6)). Equation (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 C.2.

In the simplest scenario, the pair weight could be linearly dependent on the probabilities of ith and jth objects being respectively of Type α, β, i.e., $\mathrm{w}_{ij}^{\alpha\beta} = \mathrm{w}_i^{\alpha}\mathrm{w}_j^{\beta} = \mathrm{q}_i^{\alpha}\mathrm{q}_j^{\beta}$. 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 3, $\mathrm{w}_i^A = q_i^{AA}$ for observed Type-*A* while $\mathrm{w}_i^A = q_i^{BA}$ for observed Type-*B* galaxies. Similarly, $\mathrm{w}_i^B = q_i^{AB}$ for observed Type-*A*, while $\mathrm{w}_i^B = q_i^{BB}$ 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 w_i^{α} is set to 1 for observed Type-*A* galaxies and to 0 for observed Type-*B* galaxies, and w_i^{β} is set to 0 for observed Type-*A* galaxies and to 1 for observed Type-*B*.

4.1. Estimator Bias and Variance

The estimator in Equation (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

$$\left\langle \begin{bmatrix} \widehat{w}_{AA}(\theta_k) \\ \widehat{w}_{AB}(\theta_k) \\ \widehat{w}_{BB}(\theta_k) \end{bmatrix} \right\rangle = \left\langle [D_{\rm W}] \begin{bmatrix} \widetilde{w}_{AA}^{\rm obs}(\theta_k) \\ \widetilde{w}_{AB}^{\rm obs}(\theta_k) \\ \widetilde{w}_{BB}^{\rm obs}(\theta_k) \end{bmatrix} \right\rangle = \begin{bmatrix} w_{AA}^{\rm true}(\theta_k) \\ w_{AB}^{\rm true}(\theta_k) \\ w_{BB}^{\rm true}(\theta_k) \end{bmatrix}, \quad (15)$$

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 (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 3.1, we calculate the contributions that are coming from each of the true correlation functions to any given

[8] For the matrix to be noninvertible, 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 (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.

weighted correlation function. That is, we have

$$\langle \widehat{w}^{\rm obs}_{\alpha\beta}(\theta_k)\rangle = \frac{\left(\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}\mathrm{q}_i^A\mathrm{q}_j^A\right) w_{AA}^{\rm true}(\theta_k) + \left(\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}\{\mathrm{q}_i^A\mathrm{q}_j^B + \mathrm{q}_i^B\mathrm{q}_j^A\}\right) w_{AB}^{\rm true}(\theta_k) + \left(\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}\mathrm{q}_i^B\mathrm{q}_j^B\right) w_{BB}^{\rm true}(\theta_k)}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}}. \quad (16)$$

We present the full derivation of Equation (16) in Appendix B. Consolidating the terms as done in Equation (11), we have

$$\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}} \mathrm{w}_{ij}^{AA}\mathrm{q}_i^A\mathrm{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AA}\{\mathrm{q}_i^A\mathrm{q}_j^B + \mathrm{q}_i^B\mathrm{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AA}\mathrm{q}_i^B\mathrm{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AA}} \\ \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AB}\mathrm{q}_i^A\mathrm{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AB}\{\mathrm{q}_i^A\mathrm{q}_j^B + \mathrm{q}_i^B\mathrm{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AB}\mathrm{q}_i^B\mathrm{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{AB}} \\ \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{BB}\mathrm{q}_i^A\mathrm{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{BB}\{\mathrm{q}_i^A\mathrm{q}_j^B + \mathrm{q}_i^B\mathrm{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{BB}\mathrm{q}_i^B\mathrm{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{BB}} \end{bmatrix} \begin{bmatrix} w_{AA}^{\rm true}(\theta_k) \\ w_{AB}^{\rm true}(\theta_k) \\ w_{BB}^{\rm true}(\theta_k) \end{bmatrix}. \quad (17)$$

Therefore, the `Decontaminated Weighted` estimators are given by

$$[\widehat{w}_{AA}(\theta_k)\ \ \widehat{w}_{AB}(\theta_k)\ \ \widehat{w}_{BB}(\theta_k)]^T = [D_{\rm W}]^{-1}[\widetilde{w}^{\rm obs}_{AA}(\theta_k)\ \ \widetilde{w}^{\rm obs}_{AB}(\theta_k)\ \ \widetilde{w}^{\rm obs}_{BB}(\theta_k)]^T, \quad (18)$$

where $[D_{\rm W}]$ is the square matrix in Equation (17). We note that each row in Equation (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 C.3.

We have checked Equation (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 (18) does indeed reduce to Equation (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, whereas 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 C.1; since the variance is a function of a four-point sum and depends nontrivially on the pair weights, we choose to estimate the variance numerically using the bootstrap method as described in Section 5.1. Finally, we present the generalized estimator, i.e., applicable to M target samples, in Appendix D.2.

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 3–4, 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 the 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 $0 \leqslant z \leqslant 1.4$. 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, 2015b), Crocce et al. (2015), Carretero et al. (2015), and Hoffmann et al. (2015). We query the catalog using CosmoHub[9] (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 ad hoc 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.[10] Specifically, we model the photo-z PDF for each galaxy as a Gaussian with its true redshift as the mean and σ_z as the standard deviation. Next, 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 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

[9] https://cosmohub.pic.es/home

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

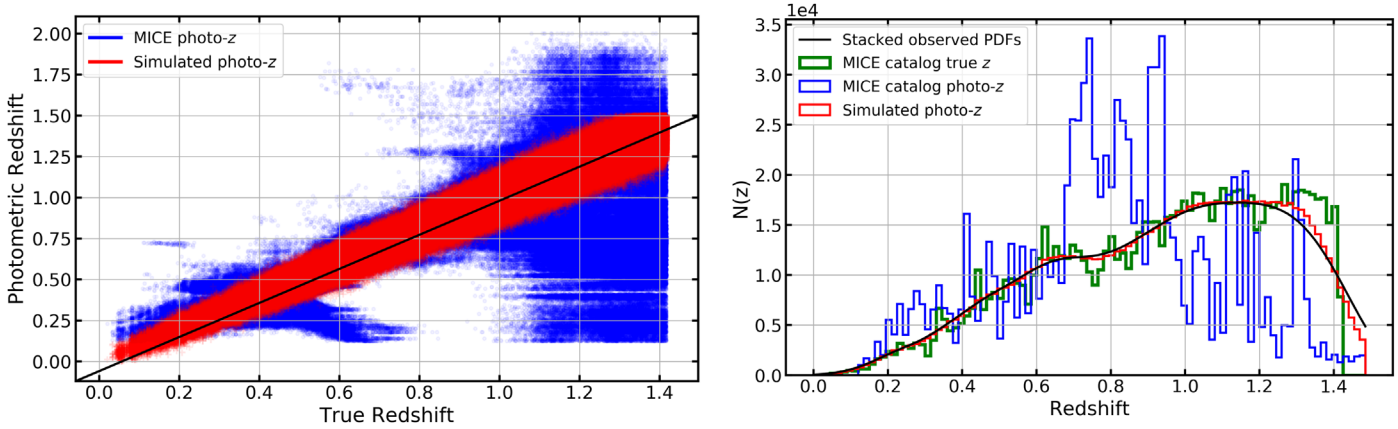

Figure 1. Illustration of the simulated photo-zs. Left: comparison between true redshift and MICE catalog photo-zs (blue) vs. those simulated here (red). Right: comparison between the different $N(z)$ distributions: true $N(z)$; those based on MICE catalog photo-zs vs. those simulated assuming Gaussian PDFs with $\sigma_z = 0.03(1+z)$. The red, blue, and green curves 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 $0 \leqslant$ R.A. $\leqslant 5$ deg, $0 \leqslant$ decl. $\leqslant 5$ deg, yielding 994,863 galaxies at $0 \leqslant z \leqslant 1.4$.

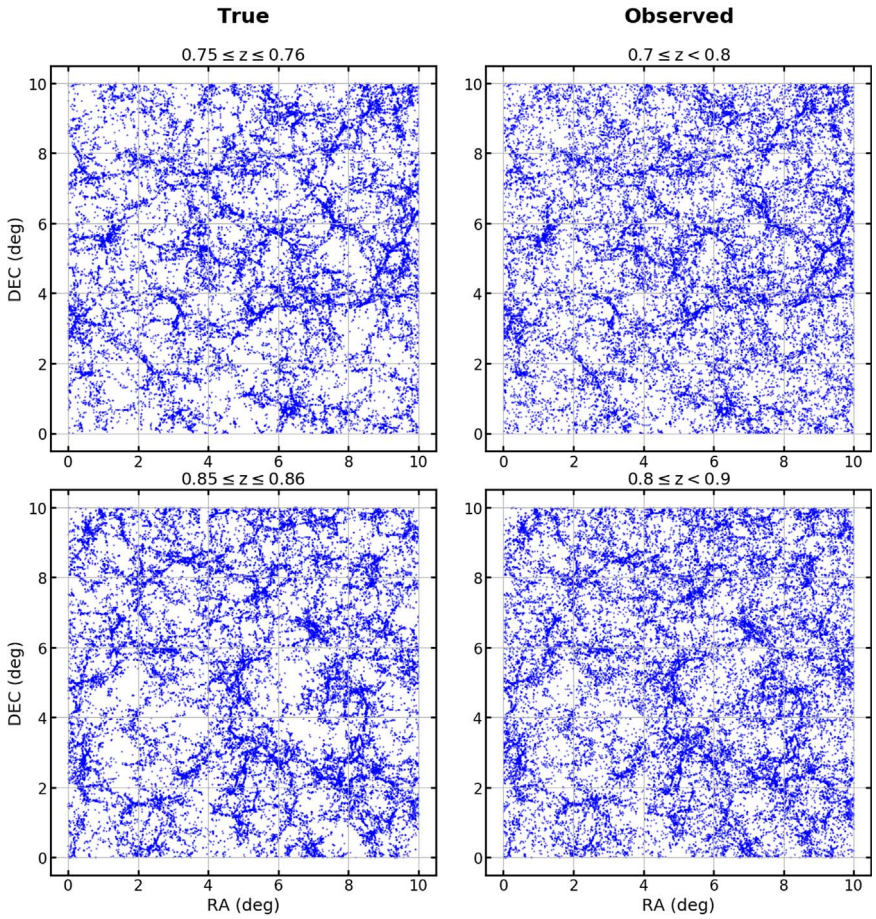

Figure 2. True and observed positions of galaxies for the idealized galaxy sample of Section 5.1, where all the true galaxies lie at $0.75 \leqslant z \leqslant 0.76$, $0.85 \leqslant z \leqslant 0.86$. We see that redshift binning of galaxies based on photo-z point estimates modifies the LSS due to the redshift contamination.

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 (R.A., decl.) and the true redshift, while the observed catalog consists of the R.A., decl., and photo-zs.

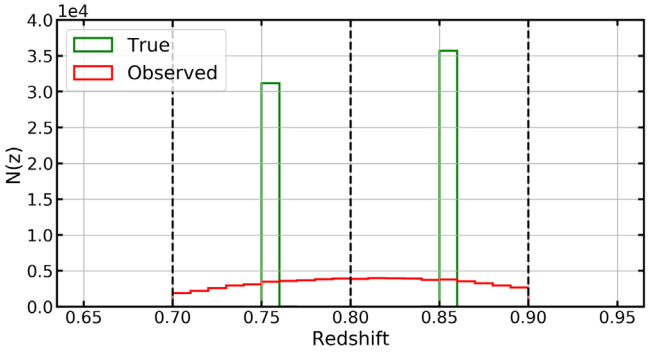

Figure 3. True and observed redshift histograms for the idealized galaxy sample of Section 5.1, with redshift bin edges shown using the vertical dashed lines. We see that photo-z uncertainties lead to a smearing of the redshift information.

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, which only needs the photo-z PDFs for all the observed galaxies. We use `TreeCorr` (Jarvis et al. 2004) to calculate the correlation functions.

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 ($0.7 \leqslant z < 0.8$, $0.8 \leqslant z < 0.9$) with the true galaxy sample having galaxies only at $0.75 \leqslant z \leqslant 0.76$, $0.85 \leqslant z \leqslant 0.86$, but with the photo-$z$ scatter as mentioned before, i.e., $\sigma_z = 0.03(1+z)$. We query the true galaxies in nine 10×10 deg^2 patches along decl. $= 0$; all patches have a similar number of galaxies (66–78 K) and face similar photo-z contamination rates (22–25% and 18–21% in the two tomographic bins, respectively). To demonstrate 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 2, where we can see that the two distributions are different, with photo-z uncertainties mixing the LSS between the two bins. Figure 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).

Next, 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., $\langle q_i^{AB} \rangle = f_{AB}$, where the average is checked over redshift intervals with $\Delta z = 0.02$, while ensuring the debiased probabilities remain in the range 0–1. For real data, this debiasing should be possible utilizing a limited set of spectroscopic redshifts. Figure 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 nine patches: the mean across the nine samples gives us the mean estimate of the respective correlation function while we

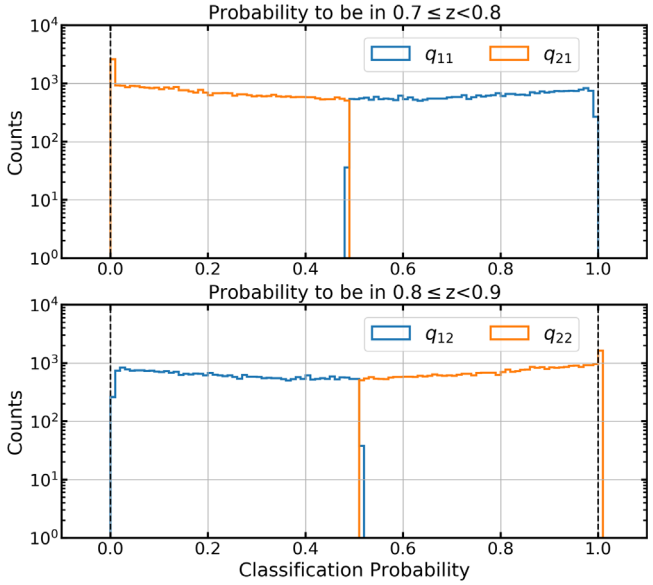

Figure 4. Distribution of the classification probabilities to be in bin 1 (upper panel) or bin 2 (lower panel) for the toy galaxy sample of Section 5.1. As introduced in Section 3, $q_{\alpha\beta}$ is the probability of the observed Type-α galaxy to be a true Type-β galaxy. We see that given the photo-z 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 (7).

calculate the estimator variance as $\langle \{\widehat{w}_i(\theta_k) - w_i^{\rm true}(\theta_k)\}^2 \rangle$ 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 five times the size of the data catalog, and restrict CF calculation to 0.01–3 deg scales. Figure 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 autocorrelations. Decontamination lowers the amplitude of the cross-correlations, and we find that both estimators correct for the contamination and reduce the bias, leading to estimates closer to the truth. This is more apparent in Figure 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 5.

5.2. Realistic Example: Optimistic Case

Now we consider a more realistic scenario: a true galaxy sample with $0.7 \leqslant z \leqslant 1.0$, with three redshift bins ($0.7 \leqslant z < 0.8$, $0.8 \leqslant z < 0.9$, $0.9 \leqslant z < 1.0$) for the tomographic

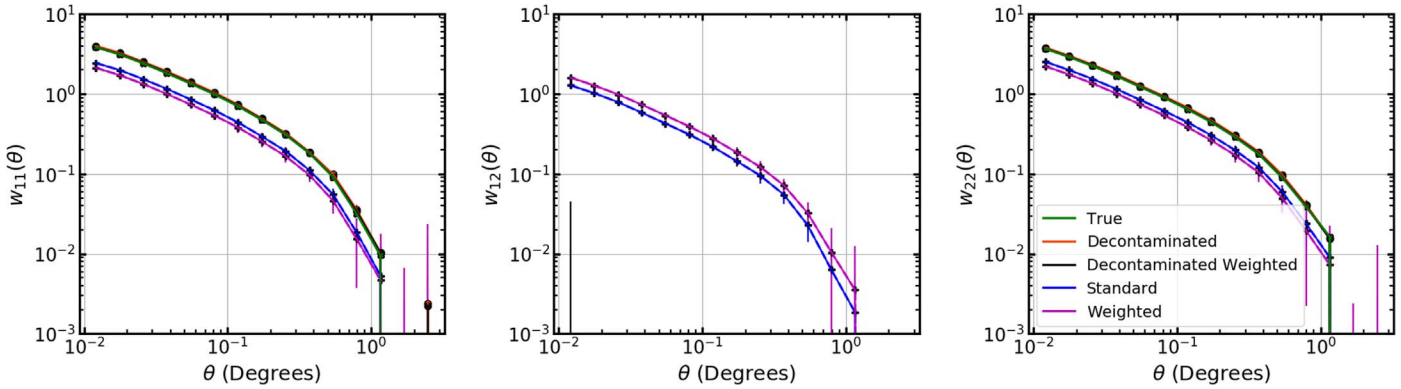

Figure 5. Correlation functions estimates and the estimator variance in the toy galaxy sample with only two redshift bins (presented in Section 5.1). We see that just as `Decontamination` (red) recovers the truth (green) using the correlations on the contaminated subsamples (blue), the `Decontaminated 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.

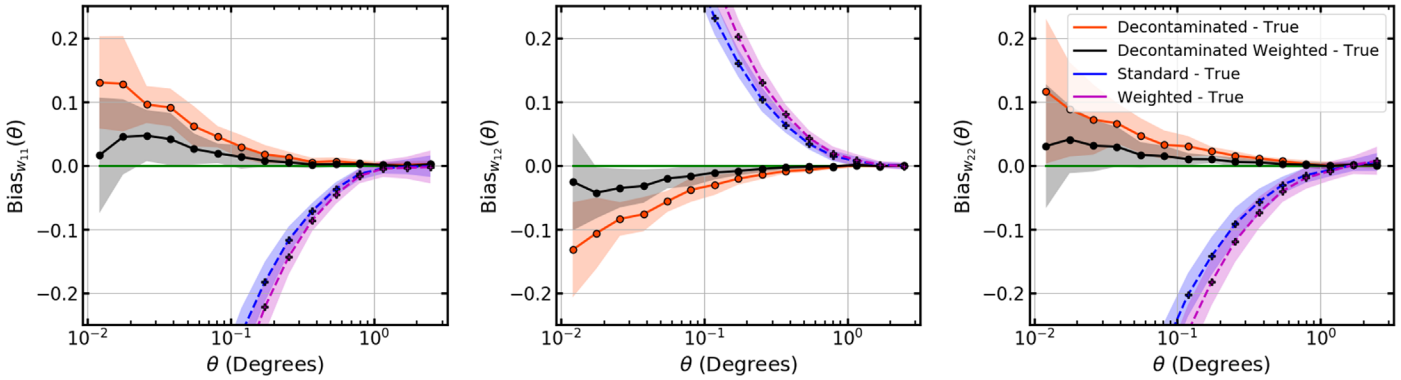

Figure 6. Bias in correlation functions for the toy galaxy sample of Section 5.1, with 1σ uncertainties in each estimator indicated with the shaded regions. We see that the `Decontaminated Weighted` estimator (black) leads to a bias smaller than that from the `Decontaminated` estimator (red); the green line indicates zero bias.

clustering analysis. As before, we query the galaxies in nine 10×10 deg^2 patches along decl. $= 0$, 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 5; all patches have a similar number of galaxies (1080–1147 K) and face similar contamination (23–26%, 44–46%, and 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 7 shows the distributions of the true redshifts and the photo-zs using one of the patches (with 1095,404 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 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 9 illustrates our results, showing only the

Figure 7. True and observed redshift histograms for the mock galaxy sample of Section 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.

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

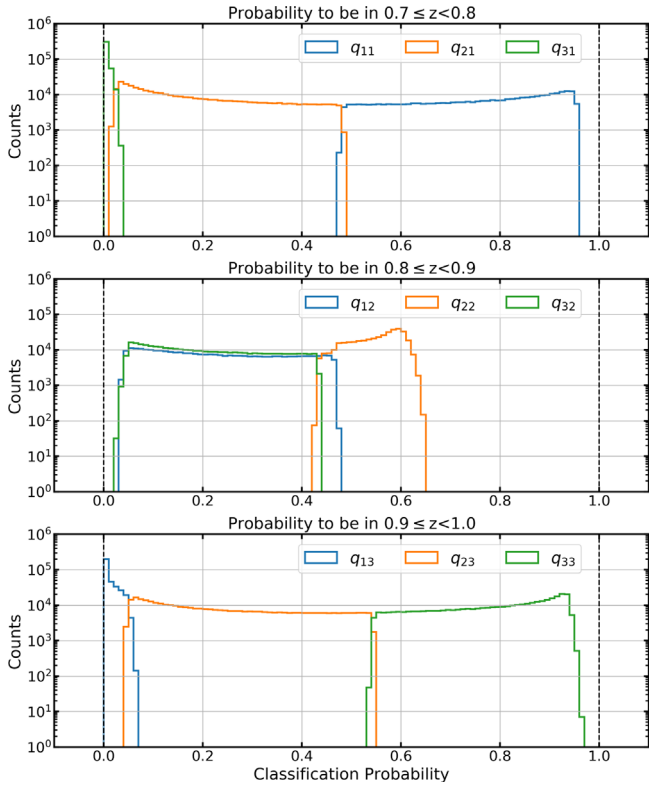

Figure 8. Distribution of the classification probabilities to be in the three target redshift bins for the mock galaxy sample of Section 5.2. The middle bin sees the largest contamination and therefore has no objects that have a very high probability to be in any target bin.

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 $\langle \{\widehat{w}_i(\theta_k) - w_i^{\rm true}(\theta_k)\}\{\widehat{w}_j(\theta_k) - w_j^{\rm true}(\theta_k)\}\rangle$ where i, j run over all the correlations (both auto and cross) and the expectation value is over all the realizations;[11] note here that the diagonal of this covariance matrix is the estimator variance used to generate uncertainties shown in Figures 5, 6, and 9. We show the estimator covariances for the mock galaxy sample considered here in Figure 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`.

Next, 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 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 matrix, 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 10, we note that LSS sample variance largely dominates over the estimator variance for the 10×10 patches considered here—a reassuring result. A comparison between the two sources of variance for larger effective survey area is left for future work.

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

5.3. Realistic Example: Pessimistic Case

Now we consider a more pessimistic scenario for the true galaxy sample of Section 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 ± 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 versus 4000 Å 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 difference between the two cases more explicitly, Figure 12 shows an

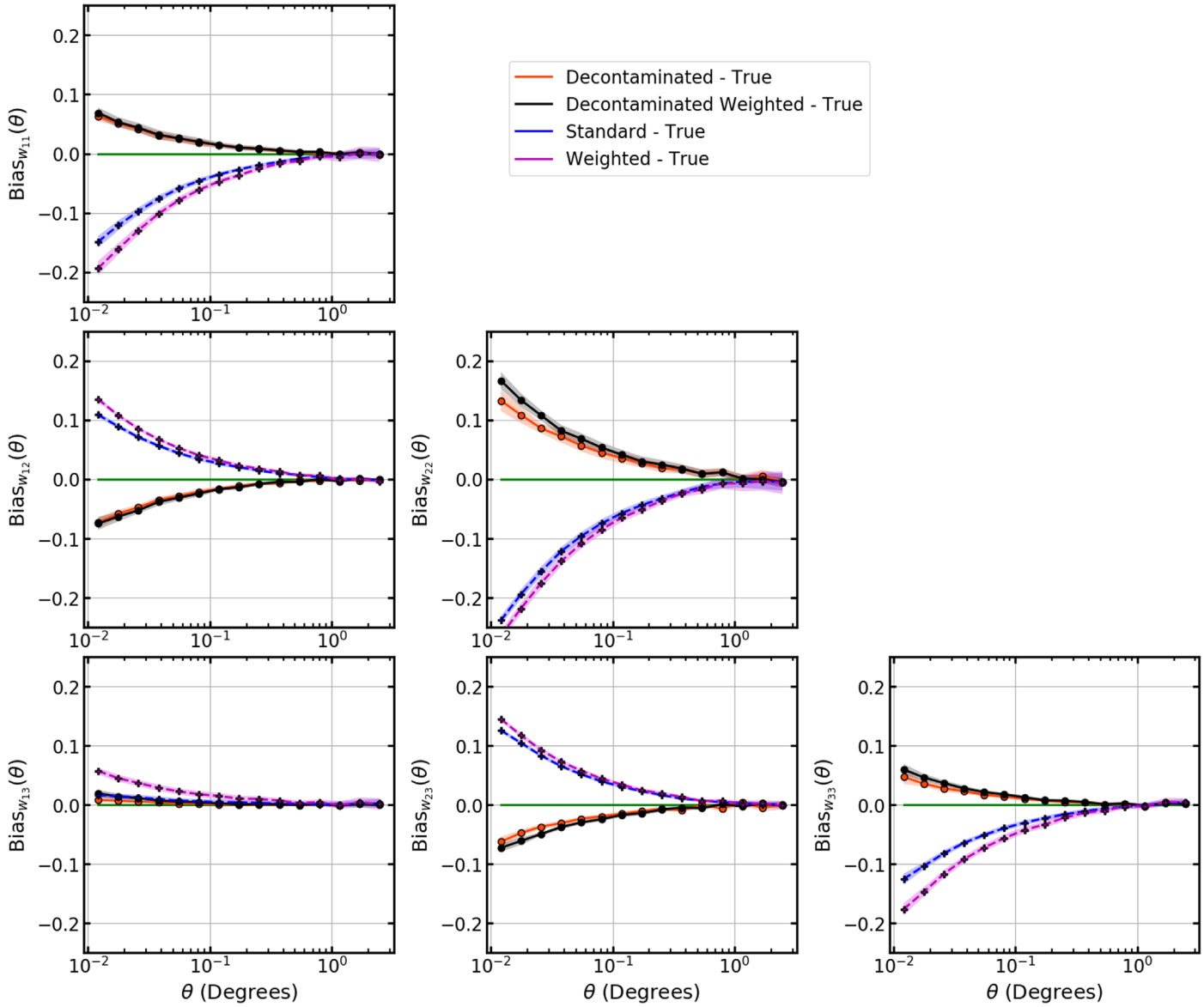

Figure 9. Bias in the correlation functions in the three-sample case of Section 5.2, with 1σ uncertainties in each estimator indicated with the shaded regions. We see that as in the toy example in Section 5.1, just as `Decontamination` (red) reduces the bias using the correlations on the contaminated 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.

example set of PDFs for the case of all-Gaussian PDFs versus half-bimodal ones.

Figure 13 shows the distributions of the true redshifts and the photo-zs using one of the patches (with 1095,404 galaxies as before, but now with 40%, 55%, and 35% contamination in the three redshift bins, respectively). Comparing it to Figure 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 14 shows the classification probabilities for all the galaxies here; comparing it to Figure 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 patches and estimate the mean CFs and the covariances. Figure 15 shows the residuals in the CF estimates, and we see that the decontaminated estimators are able to reduce the bias significantly. Figure 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 10. Finally, Figure 17 shows the full covariance matrices. Here too, we see that, as in Figure 11 for the all-Gaussian case, our decontaminated estimators approximate the true covariances more effectively than 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.

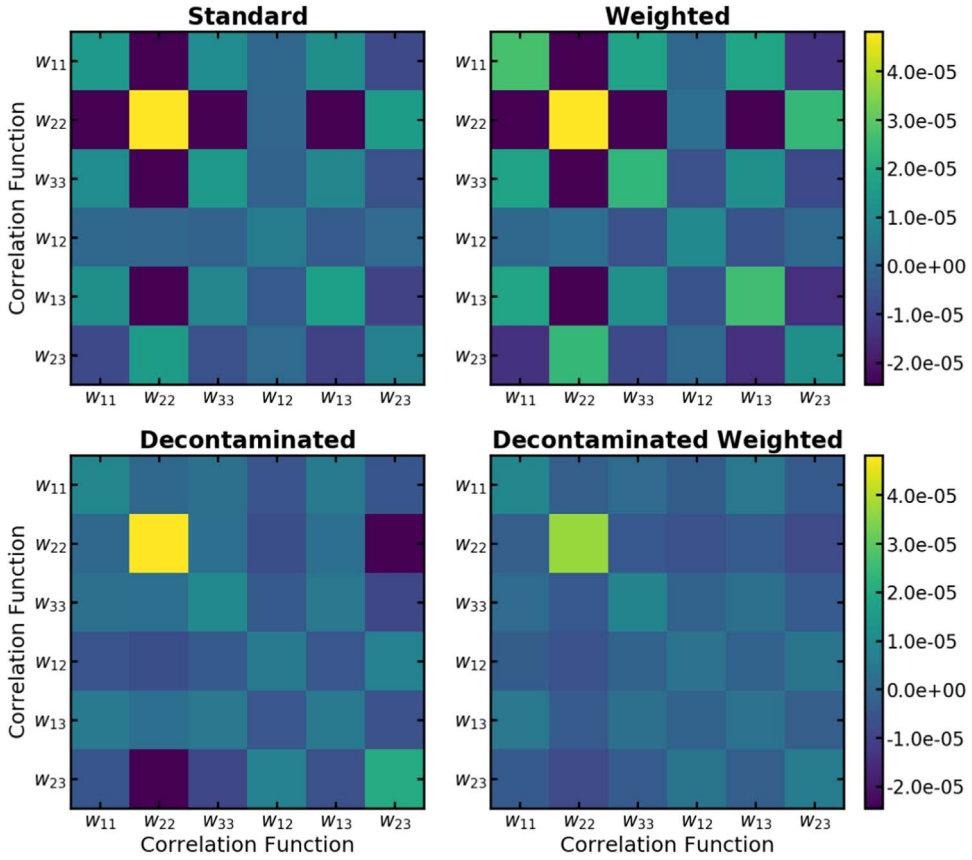

Figure 10. Estimator covariances across redshift bins for the case with three target redshift bins of Section 5.2 for an example theta bin (with $\theta = 0.^{\circ}79$ as the nominal center of the bin in $\log(\theta)$); these probe the covariances in the estimators without accounting for LSS sample 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 $\langle\{\widehat{w}_i(\theta_k) - w_i^{\rm true}(\theta_k)\}\{\widehat{w}_j(\theta_k) - w_j^{\rm true}(\theta_k)\}\rangle$ 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`.

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 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 versus 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.

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 the measured signal. In somewhat of a contrast, Zhu et al. (2015) use both weighted data and random pairs, along with 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

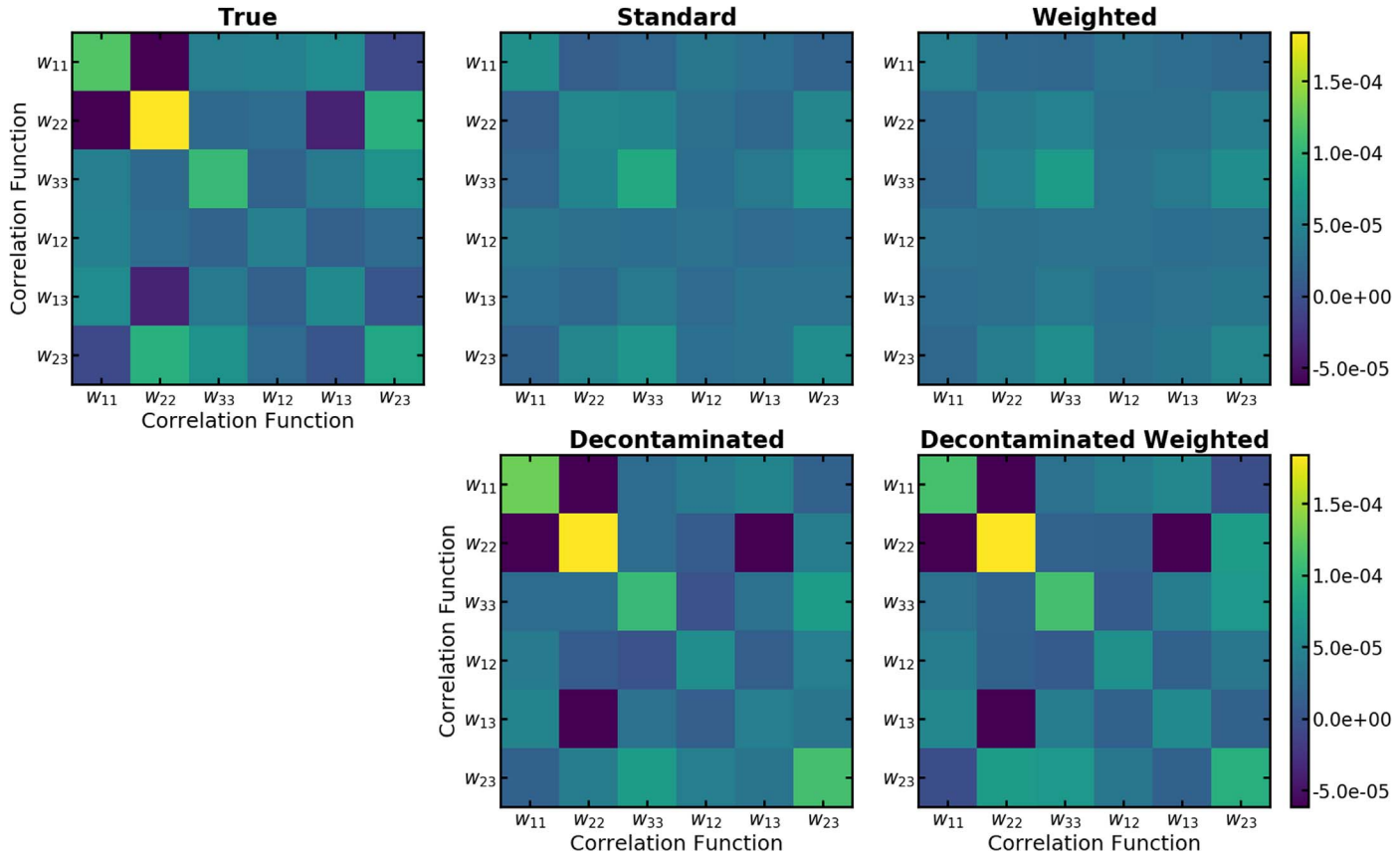

Figure 11. Full covariances across redshift bins for the case with three target redshift bins of Section 5.2 for an example theta bin (with $\theta = 0.^{\circ}79$ as the nominal center of the bin in $\log(\theta)$); these probe the covariances in the estimators while accounting for LSS sample variance. 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 because LSS at the boundary of the two bins makes w_{12} nonzero 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.

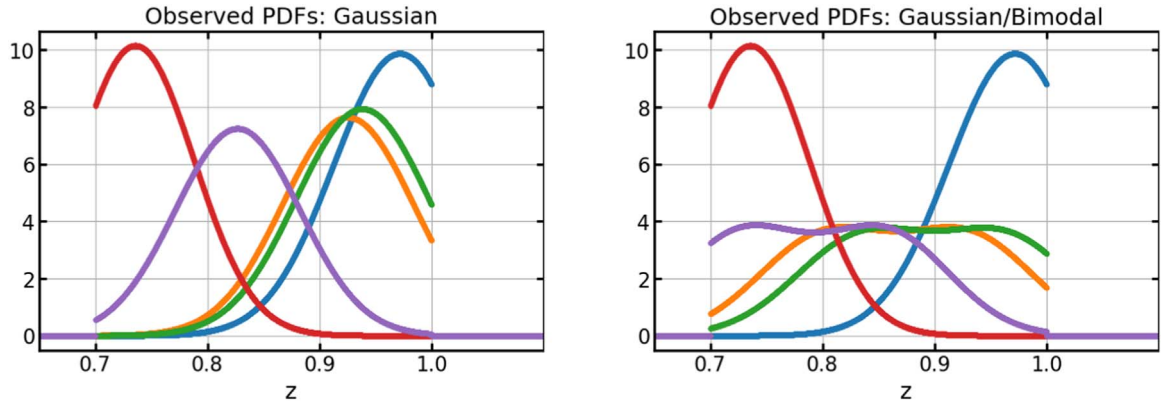

Figure 12. An example set of PDFs to compare the case of all-Gaussian PDFs of Section 5.2 vs. the case presented in Section 5.3 where half of the galaxies have bimodal PDFs. The left panel shows the observed photo-z PDFs for the case of all-Gaussian PDFs, while the right panel shows them for the case where half of the galaxies have bimodal PDFs. The colors correspond to the same objects across the panels.

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 minimize variance by introducing an additional parameter for each pair of galaxies, i.e., $\mathrm{w}_{ij,\mathrm{opt}}^{\alpha\beta} = \Upsilon_{ij}(q, k)\mathrm{w}_{ij}^{\alpha\beta}$, where $\Upsilon_{ij}(q, k)$ are the optimization parameters that minimize the variance of the estimator for each bin k. 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

Figure 13. True and observed redshift histograms for the mock galaxy sample of Section 5.3. As in Figure 7, the bin edges are shown using the vertical dashed lines. We see that, as in Figure 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.

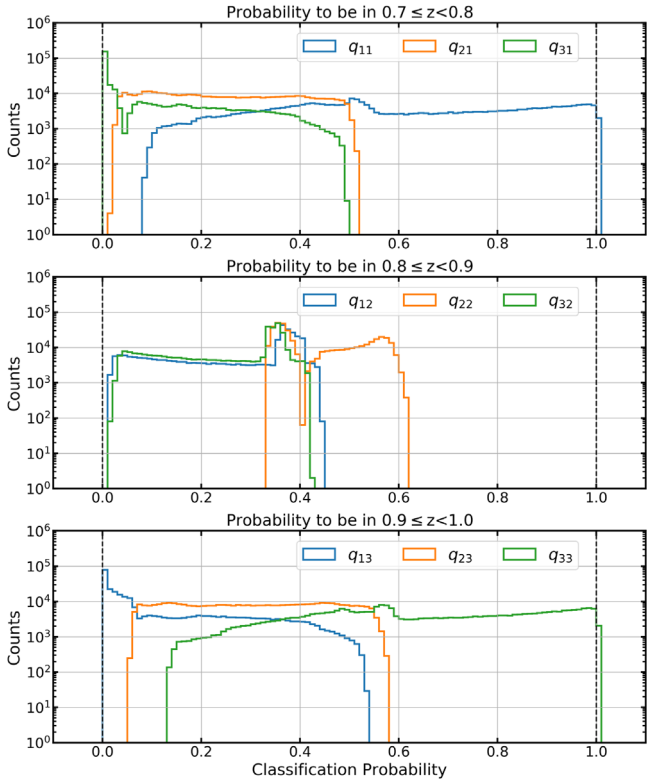

Figure 14. Distribution of the classification probabilities to be in the three target redshift bins for the mock galaxy sample of Section 5.3. As in Figure 8, the middle bin sees the largest contamination and therefore has no objects that have a very high probability to be in any target bin.

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 we can then 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 (H. Awan et al. 2020, in preparation).

We further 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 because neither decontamination nor the pair weights have an explicit redshift dependence, hence allowing for decontaminating and weighting any types. Finally, we can also extend the 2D formulation to 3D, where it will be relevant for HETDEX[12] (Hill et al. 2008), *Euclid*, and *WFIRST*, as they face emission line contaminants, as well as LSST, where the projected correlation

[12] http://hetdex.org/

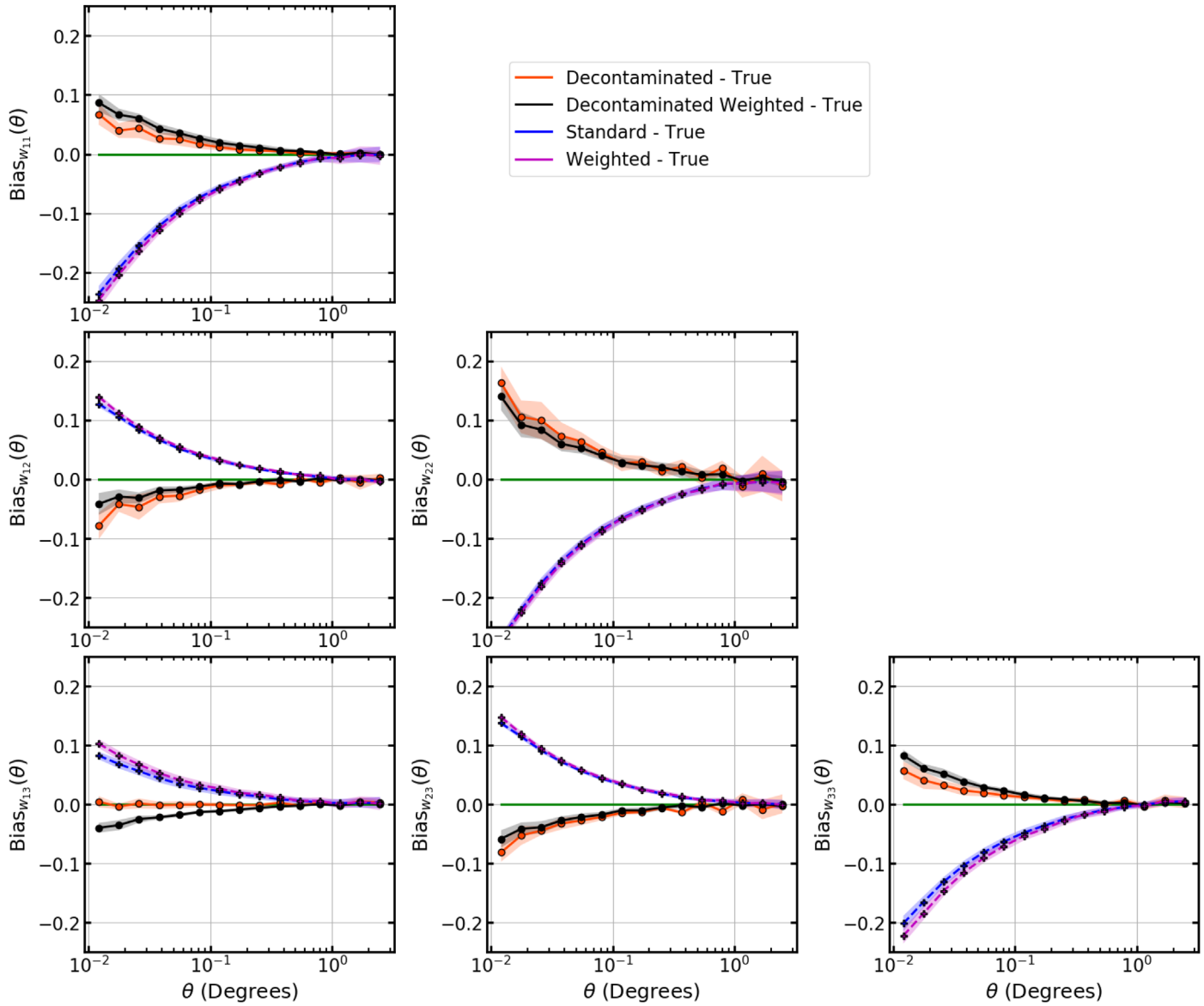

Figure 15. Bias in the correlation functions in the three sample case of Section 5.3. As in Figure 9, the 1σ uncertainties in each estimator are indicated with the shaded regions. We see that, as for the all-Gaussian photo-z PDFs case, both decontaminated estimators significantly reduce the bias and lead to estimates closer to the truth.

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.

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 data sets, 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 scalable 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 data sets. 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

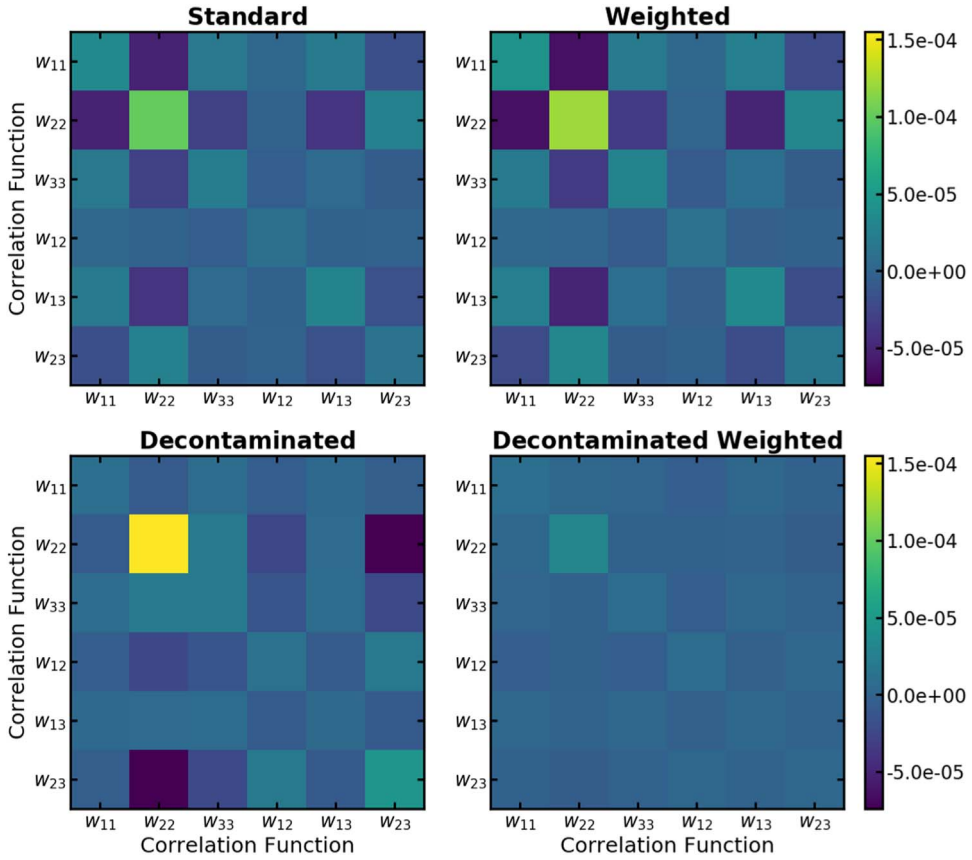

Figure 16. Estimator covariances across redshift bins for the case of Section 5.3 for the same example theta bin as in Figure 10. As in Figure 10, the left column shows estimator covariances in contaminated samples constructed using photo-z point estimates before (top) and after (bottom) decontamination, 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`.

direction of reducing biases and uncertainties in the measurement of cosmological parameters from upcoming surveys.

We thank David Alonso, Nelson Padilla, and Javier Sánchez for their helpful feedback. H.A. also thanks Kartheik Iyer and Willow Kion-Crosby for insightful discussions through the various stages of this work. H.A. 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)[13] 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.A. 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).

[13] http://oarc.rutgers.edu

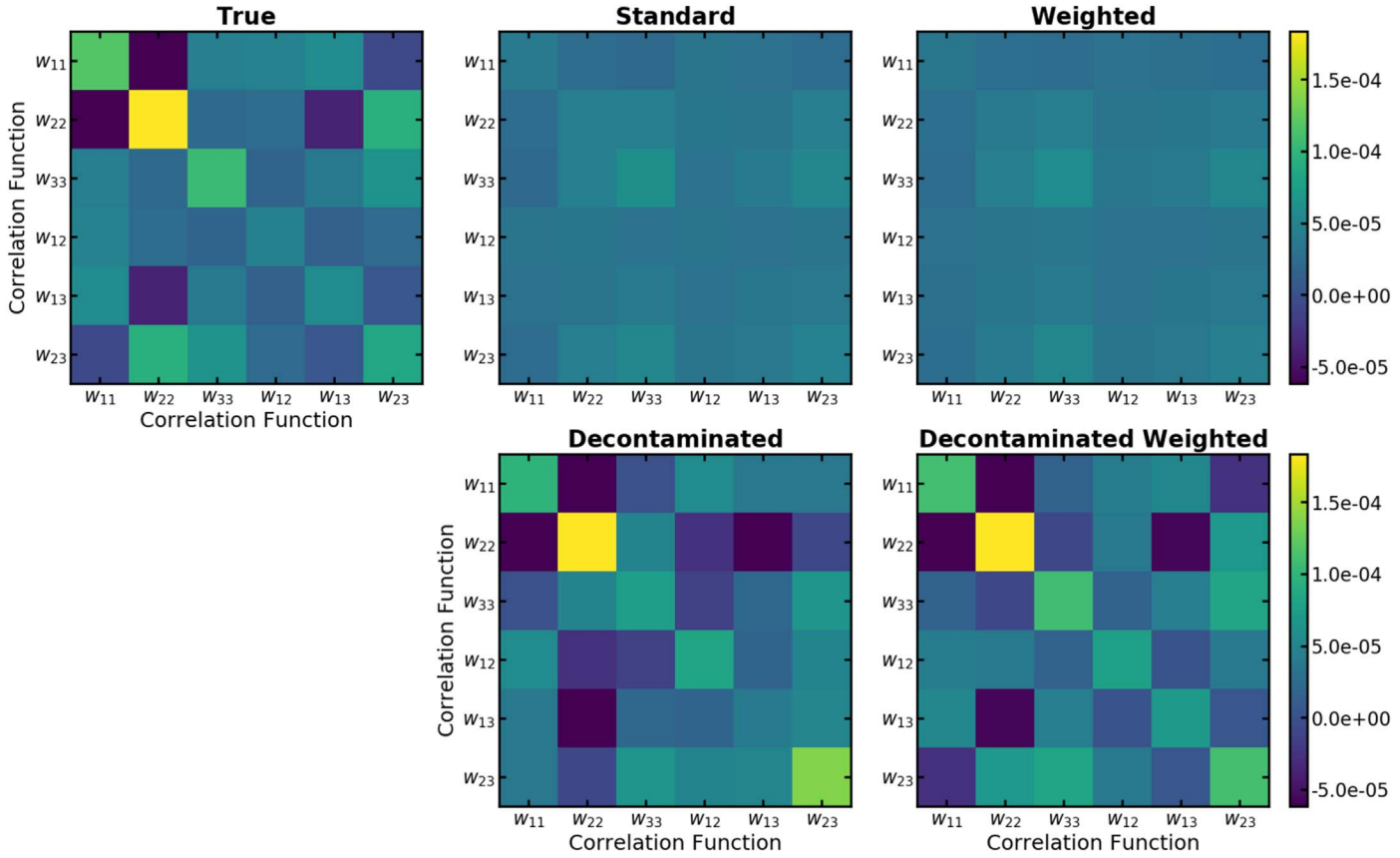

Figure 17. Full covariances across redshift bins for the case of Section 5.3 for the same example theta bin as in Figure 11. As in Figure 11, the top left panel shows the true covariances across multiple realizations of the LSS, the middle column shows covariances in contaminated samples constructed 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.

Appendix A
`Decontaminated` Estimator: Decontamination, Bias, and Variance

A.1. Decontamination Derivation

Here, we rederive the decontamination equation (Equation (11)) using the definition of angular correlation function. We start with Equation (1), rewriting it as

$$dP_{\alpha\beta}(\theta_k) = \eta^{\rm pair}_{\alpha\beta}[1 + w_{\alpha\beta}(\theta_k)]d\Omega_\alpha d\Omega_\beta \ = \mathcal{N}_{\alpha\beta}[1 + w_{\alpha\beta}(\theta_k)]\frac{d\Omega_\alpha}{V_\alpha}\frac{d\Omega_\beta}{V_\beta}, \tag{19}$$

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 $\mathcal{N}_{\alpha\beta} \rightarrow \langle\mathcal{N}_{\alpha\beta}\rangle$, 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

$$dP_{\alpha\beta}(\theta_k) = \mathcal{N}_{\alpha\beta,{\rm obs}}[1 + w^{\rm obs}_{\alpha\beta}(\theta_k)]\frac{d\Omega}{V}\frac{d\Omega}{V} \ = \sum_{\gamma,\delta}\mathcal{N}^{\gamma\delta,{\rm true}}_{\alpha\beta,{\rm obs}}[1 + w^{\rm true}_{\gamma,\delta}(\theta_k)]\frac{d\Omega}{V}\frac{d\Omega}{V}, \tag{20}$$

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

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

Since we have

$$\frac{\mathcal{N}^{\gamma\delta,\mathrm{true}}_{\alpha\beta,\mathrm{obs}}}{\mathcal{N}^{\mathrm{tot}}_{\alpha\beta,\mathrm{obs}}} = f_{\alpha\gamma} f_{\beta\delta} \tag{22}$$

$$\Rightarrow [1 + w^{\mathrm{obs}}_{\alpha\beta}(\theta_k)] = f_{\alpha1} f_{\beta1}[1 + w^{\mathrm{true}}_{11}(\theta_k)] + \{f_{\alpha1} f_{\beta2} + f_{\alpha2} f_{\beta1}\}[1 + w^{\mathrm{true}}_{12}(\theta_k)] + f_{\alpha2} f_{\beta2}[1 + w^{\mathrm{true}}_{22}(\theta_k)]. \tag{23}$$

Therefore, for α, $\beta = 1, 2$, Equation (23) becomes

$$[1 + w^{\mathrm{obs}}_{12}(\theta_k)] = f_{11} f_{21}[1 + w^{\mathrm{true}}_{11}(\theta_k)] + \{f_{11} f_{22} + f_{12} f_{21}\}[1 + w^{\mathrm{true}}_{12}(\theta_k)] + f_{12} f_{22}[1 + w^{\mathrm{true}}_{22}(\theta_k)]. \tag{24}$$

Now, since

$$f_{11} f_{21} + \{f_{11} f_{22} + f_{12} f_{21}\} + f_{12} f_{22} = f_{11}[f_{21} + f_{22}] + f_{12}[f_{21} + f_{22}] = 1, \tag{25}$$

we have

$$w^{\mathrm{obs}}_{12}(\theta_k) = f_{11} f_{21} w^{\mathrm{true}}_{11}(\theta_k) + \{f_{11} f_{22} + f_{12} f_{21}\} w^{\mathrm{true}}_{12}(\theta_k) + f_{12} f_{22} w^{\mathrm{true}}_{22}(\theta_k), \tag{26}$$

which agrees with Equation (11). Similar results follow for $(\alpha, \beta) = (1, 1), (2, 2)$.

A.2. Estimator Bias

We expect that the `Decontaminated` estimators are unbiased given their construction (i.e., Equation (10)). However, for brevity, we formally show that they are indeed unbiased. By definition, an unbiased estimator is such that

$$\langle \widehat{w} \rangle = w_{\mathrm{true}}, \tag{27}$$

where the expectation value is over many realizations of the survey. Now, using Equations (11) and (12), we have

$$\begin{aligned}\left\langle \left[\widehat{w}_{AA}(\theta_k) \;\; \widehat{w}_{AB}(\theta_k) \;\; \widehat{w}_{BB}(\theta_k)\right]^T \right\rangle &= \left\langle [D_{\mathrm{S}}]^{-1} \left[w^{\mathrm{obs}}_{AA}(\theta_k) \;\; w^{\mathrm{obs}}_{AB}(\theta_k) \;\; w^{\mathrm{obs}}_{BB}(\theta_k)\right]^T \right\rangle \\ &= [D_{\mathrm{S}}]^{-1}[D_{\mathrm{S}}] \left[w^{\mathrm{true}}_{AA}(\theta_k) \;\; w^{\mathrm{true}}_{AB}(\theta_k) \;\; w^{\mathrm{true}}_{BB}(\theta_k)\right]^T = \left[w^{\mathrm{true}}_{AA}(\theta_k) \;\; w^{\mathrm{true}}_{AB}(\theta_k) \;\; w^{\mathrm{true}}_{BB}(\theta_k)\right]^T,\end{aligned} \tag{28}$$

where the second equality follows by substituting Equation (11). Hence, the `Decontaminated` estimators are unbiased. We note here that $[D_{\mathrm{S}}]$ in Equation (12) is effectively a decontamination matrix: it removes the contamination from the biased estimates, $w^{\mathrm{obs}}_{\alpha\beta}$, in the presence of sample contamination. A similar argument follows for the case where we have M target samples, using Equation (108). We also note that Equation (28) is valid only when $f_{\alpha\beta}$ are accurate averages of the classification probabilities.

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 (12), we have

$$\left[\sigma^2_{\widehat{w}_{AA}}(\theta_k) \;\; \sigma^2_{\widehat{w}_{AB}}(\theta_k) \;\; \sigma^2_{\widehat{w}_{BB}}(\theta_k)\right]^T = \{[D_{\mathrm{S}}]^{-1}\}^2_{ij} \left[\sigma^2_{w^{\mathrm{obs}}_{AA}}(\theta_k) \;\; \sigma^2_{w^{\mathrm{obs}}_{AB}}(\theta_k) \;\; \sigma^2_{w^{\mathrm{obs}}_{BB}}(\theta_k)\right]^T, \tag{29}$$

where $\{[D_{\mathrm{S}}]^{-1}\}^2_{ij}$ denotes that matrix resulting from squaring each individual coefficient in the matrix $[D_{\mathrm{S}}]^{-1}$. We also note that the above derivation assumes no covariance between the observed correlations (i.e., $w^{\mathrm{obs}}_{\alpha\beta}$), which is incorrect for the case of neighboring redshift bins, given the shared LSS between them; this is discussed in more detail when we discuss the covariance matrices in Section 5.2. To consider the covariance matrix for the `Decontaminated` estimators, we start with Equation (12), which is reproduced here:

$$[\widehat{w}_{AA}(\theta_k) \;\; \widehat{w}_{AB}(\theta_k) \;\; \widehat{w}_{BB}(\theta_k)]^T = [D_{\mathrm{S}}]^{-1}[w^{\mathrm{obs}}_{AA}(\theta_k) \;\; w^{\mathrm{obs}}_{AB}(\theta_k) \;\; w^{\mathrm{obs}}_{BB}(\theta_k)]^T. \tag{30}$$

Given Equation (28), we therefore have

$$\langle [\widehat{w}_{AA}(\theta_k) \;\; \widehat{w}_{AB}(\theta_k) \;\; \widehat{w}_{BB}(\theta_k)]^T \rangle = [D_{\mathrm{S}}]^{-1} \langle [w^{\mathrm{obs}}_{AA}(\theta_k) \;\; w^{\mathrm{obs}}_{AB}(\theta_k) \;\; w^{\mathrm{obs}}_{BB}(\theta_k)]^T \rangle, \tag{31}$$

where we assume that $[D_S]$ 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

$$[\Delta\widehat{w}_{AA}(\theta_k) \;\; \Delta\widehat{w}_{AB}(\theta_k) \;\; \Delta\widehat{w}_{BB}(\theta_k)]^T = [D_S]^{-1}[\Delta w^{\rm obs}_{AA}(\theta_k) \;\; \Delta w^{\rm obs}_{AB}(\theta_k) \;\; \Delta w^{\rm obs}_{BB}(\theta_k)]^T. \tag{32}$$

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

$$C_{\widehat{w}}(\theta_k) = \langle [\Delta\widehat{w}_{AA}(\theta_k) \;\; \Delta\widehat{w}_{AB}(\theta_k) \;\; \Delta\widehat{w}_{BB}(\theta_k)]^T [\Delta\widehat{w}_{AA}(\theta_k) \;\; \Delta\widehat{w}_{AB}(\theta_k) \;\; \Delta\widehat{w}_{BB}(\theta_k)] \rangle. \tag{33}$$

Using Equation (32) and its transpose, we then have

$$\begin{aligned} C_{\widehat{w}}(\theta_k) &= \langle [D_S]^{-1}[\Delta w^{\rm obs}_{AA}(\theta_k) \;\; \Delta w^{\rm obs}_{AB}(\theta_k) \;\; \Delta w^{\rm obs}_{BB}(\theta_k)]^T [\Delta w^{\rm obs}_{AA}(\theta_k) \;\; \Delta w^{\rm obs}_{AB}(\theta_k) \;\; \Delta w^{\rm obs}_{BB}(\theta_k)] [[D_S]^{-1}]^T \rangle \\ &= [D_S]^{-1} \langle [\Delta w^{\rm obs}_{AA}(\theta_k) \;\; \Delta w^{\rm obs}_{AB}(\theta_k) \;\; \Delta w^{\rm obs}_{BB}(\theta_k)]^T [\Delta w^{\rm obs}_{AA}(\theta_k) \;\; \Delta w^{\rm obs}_{AB}(\theta_k) \;\; \Delta w^{\rm obs}_{BB}(\theta_k)] \rangle [[D_S]^{-1}]^T \\ &= [D_S]^{-1} C_{w^{\rm obs}}(\theta_k) [[D_S]^{-1}]^T, \end{aligned} \tag{34}$$

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 $[D_S]$ is constant.

Both $C_{w^{\rm obs}}(\theta_k)$ and $C_{\widehat{w}}(\theta_k)$ can be determined via bootstrap, as done for the example considered in Section 5.2, with the estimated covariance matrices presented in Figures 11 and 17. We note that $C_{\widehat{w}}(\theta_k)$ may be calculated using $C_{w^{\rm obs}}(\theta_k)$ given Equation (34), assuming that $[D_S]$ 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 (34), we note that if $C_{w^{\rm obs}}(\theta_k)$ is diagonal, i.e., there are no covariances in the observed correlations, Equation (34) leads to the variance in the `Decontaminated` estimators as given by Equation (29).

Appendix B
Decontamination: From `Decontaminated` with Full Sample to Weighted

Here, we present the methodology to decontaminate the `Weighted` correlation function introduced in Equation (13), using the formalism introduced in Appendix A.1. To develop intuition, we first extend the methodology in Appendix A.1 to consider an unweighted full observed sample, followed by considering the weighted full sample.

B.1. `Decontaminated`: Full Sample

We extend the treatment in Appendix A.1 to consider an unweighted full sample. The analog of Equation (20) is then

$$dP(\theta_k) = \mathcal{N}_{\rm tot_{obs}}[1 + w^{\rm full}(\theta_k)] \frac{d\Omega}{V}\frac{d\Omega}{V} = \sum_{\gamma,\delta} \mathcal{N}^{\gamma\delta,\rm true}_{\rm tot_{obs}}[1 + w^{\rm true}_{\gamma,\delta}(\theta_k)] \frac{d\Omega}{V}\frac{d\Omega}{V}. \tag{35}$$

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

$$\begin{aligned} \mathcal{N}_{\rm tot_{obs}}[1 + w^{\rm full}(\theta_k)] &= \mathcal{N}^{11,\rm true}_{\rm tot_{obs}}[1 + w^{\rm true}_{11}(\theta_k)] + \mathcal{N}^{12,\rm true}_{\rm tot_{obs}}[1 + w^{\rm true}_{12}(\theta_k)] \\ &\quad + \mathcal{N}^{21,\rm true}_{\rm tot_{obs}}[1 + w^{\rm true}_{21}(\theta_k)] + \mathcal{N}^{22,\rm true}_{\rm tot_{obs}}[1 + w^{\rm true}_{22}(\theta_k)]. \end{aligned} \tag{36}$$

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

$$\sum_{i}^{N^{\gamma}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{\delta}_{\rm tot_{obs}}} \mathfrak{q}^{\gamma}_i \mathfrak{q}^{\delta}_j = \widehat{\mathcal{N}}^{\gamma\delta,\rm true}_{\rm tot_{obs}}. \tag{37}$$

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

$$\begin{aligned} \mathcal{N}_{\rm tot_{obs}}[1 + w^{\rm full}(\theta_k)] &= \sum_{i}^{N^{1}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{1}_{\rm tot_{obs}}} \mathfrak{q}^{1}_i \mathfrak{q}^{1}_j [1 + w^{\rm true}_{11}(\theta_k)] + \sum_{i}^{N^{1}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{2}_{\rm tot_{obs}}} \mathfrak{q}^{1}_i \mathfrak{q}^{2}_j [1 + w^{\rm true}_{12}(\theta_k)] \\ &\quad + \sum_{i}^{N^{2}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{1}_{\rm tot_{obs}}} \mathfrak{q}^{2}_i \mathfrak{q}^{1}_j [1 + w^{\rm true}_{21}(\theta_k)] + \sum_{i}^{N^{2}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{2}_{\rm tot_{obs}}} \mathfrak{q}^{2}_i \mathfrak{q}^{2}_j [1 + w^{\rm true}_{22}(\theta_k)]. \end{aligned} \tag{38}$$

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. Thus, for $\alpha, \beta = 1, 2$, Equation (38) becomes

$$
\begin{aligned}
N_{1,\rm obs}N_{2,\rm obs}[1+w_{11}^{\rm obs}(\theta_k)] &= \left(\sum_i^{N_{1,\rm obs}}\sum_{j\neq i}^{N_{2,\rm obs}} \mathrm{q}_i^1\mathrm{q}_j^1\right)[1+w_{11}^{\rm true}(\theta_k)] + \left(\sum_i^{N_{1,\rm obs}}\sum_{j\neq i}^{N_{2,\rm obs}} \mathrm{q}_i^1\mathrm{q}_j^2\right)[1+w_{12}^{\rm true}(\theta_k)] \\
&\quad + \left(\sum_i^{N_{1,\rm obs}}\sum_{j\neq i}^{N_{2,\rm obs}} \mathrm{q}_i^2\mathrm{q}_j^1\right)[1+w_{21}^{\rm true}(\theta_k)] + \left(\sum_i^{N_{1,\rm obs}}\sum_{j\neq i}^{N_{2,\rm obs}} \mathrm{q}_i^2\mathrm{q}_j^2\right)[1+w_{22}^{\rm true}(\theta_k)] \\
&= \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} \mathrm{q}_i^1\mathrm{q}_j^1\right)[1+w_{11}^{\rm true}(\theta_k)] + \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} \mathrm{q}_i^1\mathrm{q}_j^2\right)[1+w_{12}^{\rm true}(\theta_k)] \\
&\quad + \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} \mathrm{q}_i^2\mathrm{q}_j^1\right)[1+w_{21}^{\rm true}(\theta_k)] + \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} \mathrm{q}_i^2\mathrm{q}_j^2\right)[1+w_{22}^{\rm true}(\theta_k)] \\
&\xrightarrow{\text{simplify } q\text{s}} \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} q_{i,11}q_{j,12}\right)[1+w_{11}^{\rm true}(\theta_k)] + \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} q_{i,11}q_{j,22}\right)[1+w_{12}^{\rm true}(\theta_k)] \\
&\quad + \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} q_{i,12}q_{j,21}\right)[1+w_{21}^{\rm true}(\theta_k)] + \left(\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}} q_{i,12}q_{j,22}\right)[1+w_{22}^{\rm true}(\theta_k)] \\
&\xrightarrow{q\text{s}=f\text{s}} \left(f_{11}f_{21}\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}}\right)[1+w_{11}^{\rm true}(\theta_k)] + \left(f_{11}f_{22}\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}}\right)[1+w_{12}^{\rm true}(\theta_k)] \\
&\quad + \left(f_{12}f_{21}\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}}\right)[1+w_{21}^{\rm true}(\theta_k)] + \left(f_{12}f_{22}\sum_i^{N_{1,\rm obs}}\sum_{j}^{N_{2,\rm obs}}\right)[1+w_{22}^{\rm true}(\theta_k)] \\
&= f_{11}f_{21}N_{1,\rm obs}N_{2,\rm obs}[1+w_{11}^{\rm true}(\theta_k)] + f_{11}f_{22}N_{1,\rm obs}N_{2,\rm obs}[1+w_{12}^{\rm true}(\theta_k)] \\
&\quad + f_{12}f_{21}N_{1,\rm obs}N_{2,\rm obs}[1+w_{21}^{\rm true}(\theta_k)] + f_{12}f_{22}N_{1,\rm obs}N_{2,\rm obs}[1+w_{22}^{\rm true}(\theta_k)] \qquad (39)
\end{aligned}
$$

$$
\Rightarrow [1+w_{12}^{\rm obs}(\theta_k)] = f_{11}f_{21}[1+w_{11}^{\rm true}(\theta_k)] + \{f_{11}f_{22}+f_{12}f_{21}\}[1+w_{12}^{\rm true}(\theta_k)] + f_{12}f_{22}[1+w_{22}^{\rm true}(\theta_k)], \qquad (40)
$$

which agrees with Equation (26). Similar results follow for $(\alpha, \beta) = (1, 1) = (2, 2)$.

B.2. `Weighted`: Full Sample

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

$$
d\widetilde{P}_{\alpha\beta}(\theta_k) = \widetilde{\mathcal{N}}_{\rm tot_{obs}}^{\alpha\beta,\rm obs}[1+\widetilde{w}_{\alpha\beta}(\theta_k)]\frac{d\Omega}{V}\frac{d\Omega}{V}, \qquad (41)
$$

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

$$
\widetilde{\mathcal{N}}_{\rm tot_{obs}}^{\alpha\beta,\rm obs} = \sum_i^{N_{\rm tot_{obs}}^{\alpha}}\sum_{j\neq i}^{N_{\rm tot_{obs}}^{\beta}} \mathrm{w}_{ij}^{\alpha\beta}. \qquad (42)
$$

Now, when writing the analog of Equations (20)–(35), we need to account for pair weights, leading us to

$$
d\widetilde{P}_{\alpha\beta}(\theta_k) = \widetilde{\mathcal{N}}_{\rm tot_{obs}}^{\alpha\beta,\rm obs}[1+\widetilde{w}_{\alpha\beta}(\theta_k)]\frac{d\Omega}{V}\frac{d\Omega}{V} = \sum_{\gamma,\delta}\widetilde{\mathcal{N}}_{\rm tot_{obs}}^{\gamma\delta,\rm true}[1+w_{\gamma,\delta}^{\rm true}(\theta_k)]\frac{d\Omega}{V}\frac{d\Omega}{V}, \qquad (43)
$$

where we have the analog of Equation (37):

$$\sum_{i}^{N^{\alpha}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{\beta}_{\rm tot_{obs}}} \mathrm{w}_{ij}^{\alpha\beta} \mathrm{q}_i^{\alpha} \mathrm{q}_j^{\beta} = \widetilde{\widetilde{\mathcal{N}}}^{\alpha\beta,\rm true}_{\rm tot_{obs}}. \tag{44}$$

Now, expanding the sum in Equation (43), we have

$$\begin{aligned}\widetilde{\mathcal{N}}^{\alpha\beta,\rm obs}_{\rm tot_{obs}}[1+\widetilde{w}_{\alpha\beta}(\theta_k)] = \widetilde{\mathcal{N}}^{11,\rm true}_{\rm tot_{obs}}[1+w^{\rm true}_{11}(\theta_k)] + \widetilde{\mathcal{N}}^{12,\rm true}_{\rm tot_{obs}}[1+w^{\rm true}_{12}(\theta_k)] \\ + \widetilde{\mathcal{N}}^{21,\rm true}_{\rm tot_{obs}}[1+w^{\rm true}_{21}(\theta_k)] + \widetilde{\mathcal{N}}^{22,\rm true}_{\rm tot_{obs}}[1+w^{\rm true}_{22}(\theta_k)].\end{aligned} \tag{45}$$

Substituting Equation (37) to estimate the true counts, we have

$$\begin{aligned}\left(\sum_{i}^{N^{\alpha}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{\beta}_{\rm tot_{obs}}} \mathrm{w}_{ij}^{\alpha\beta}\right)[1+\widetilde{w}^{\rm full}_{\alpha\beta}(\theta_k)] = \left(\sum_{i}^{N^{\alpha}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{\beta}_{\rm tot_{obs}}} \mathrm{w}_{ij}^{\alpha\beta}\mathrm{q}_i^1\mathrm{q}_j^1\right)[1+w^{\rm true}_{11}(\theta_k)] + \left(\sum_{i}^{N^{\alpha}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{\beta}_{\rm tot_{obs}}} \mathrm{w}_{ij}^{\alpha\beta}\mathrm{q}_i^1\mathrm{q}_j^2\right)[1+w^{\rm true}_{12}(\theta_k)] \\ + \left(\sum_{i}^{N^{\alpha}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{\beta}_{\rm tot_{obs}}} \mathrm{w}_{ij}^{\alpha\beta}\mathrm{q}_i^2\mathrm{q}_j^1\right)[1+w^{\rm true}_{21}(\theta_k)] + \left(\sum_{i}^{N^{\alpha}_{\rm tot_{obs}}} \sum_{j\neq i}^{N^{\beta}_{\rm tot_{obs}}} \mathrm{w}_{ij}^{\alpha\beta}\mathrm{q}_i^2\mathrm{q}_j^2\right)[1+w^{\rm true}_{22}(\theta_k)].\end{aligned} \tag{46}$$

Note that this equation reduces to `Decontaminated` as in Equation (39) when weights are set to 1 for target subsample and 0 for the rest, and that we basically have theta-independent decontamination.

Appendix C
`Weighted` Estimator: Variance and Practical Notes

C.1. `Weighted` *Estimator: Variance*

Here, we follow the procedure in LS93 to estimate the variance of the `Weighted` estimator introduced in Equation (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 Appendix 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

$$\begin{aligned}(\overline{DD})_{\alpha\beta}(\theta_k) &= \langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle(1+\eta(\theta_k)) \\ (\overline{RR})(\theta_k) &= \langle(\overline{RR})(\theta_k)\rangle(1+\gamma(\theta_k)),\end{aligned} \tag{47}$$

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

$$\langle\eta(\theta_k)\rangle = \langle\gamma(\theta_k)\rangle = 0, \tag{48}$$

and hence, we have

$$\begin{aligned}\sigma^2_{\eta}(\theta_k) &= \langle\eta^2(\theta_k)\rangle - \cancelto{0}{\langle\eta(\theta_k)\rangle^2} = \langle\eta^2(\theta_k)\rangle \\ \sigma^2_{\gamma}(\theta_k) &= \langle\gamma^2(\theta_k)\rangle - \cancelto{0}{\langle\gamma(\theta_k)\rangle^2} = \langle\gamma^2(\theta_k)\rangle \\ \mathrm{cov}(\eta,\gamma)(\theta_k) &= \langle\eta(\theta_k)\gamma(\theta_k)\rangle - \cancelto{0}{\langle\eta(\theta_k)\rangle}\cancelto{0}{\langle\gamma(\theta_k)\rangle} = 0\end{aligned} \tag{49}$$

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. Thus, given Equations (13) and (47), we have

$$1 + \widetilde{w}_{\alpha\beta}(\theta_k) = \frac{\widetilde{(DD)}_{\alpha\beta}(\theta_k)}{RR(\theta_k)} = \frac{(\overline{DD})_{\alpha\beta}(\theta_k)}{\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{N_r(N_r-1)/2}{(\overline{RR})(\theta_k)} = \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle(1+\eta(\theta_k))}{\langle(\overline{RR})(\theta_k)\rangle(1+\gamma(\theta_k))}, \tag{50}$$

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

$$RR(\theta_k) = \frac{\sum_i^{N_r}\sum_{j>i}^{N_r}\bar{\Theta}_{ij,k}}{\sum_i^{N_r}\sum_{j>i}^{N_r}} = \frac{\sum_i^{N_r}\sum_{j>i}^{N_r}\bar{\Theta}_{ij,k}}{N_r(N_r-1)/2} \tag{51}$$

$$\begin{aligned}
\Rightarrow 1 + \langle\widetilde{w}_{\alpha\beta}(\theta_k)\rangle &= \left\langle \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle(1+\eta(\theta_k))}{\langle(\overline{RR})\rangle(\theta_k)(1+\gamma(\theta_k))} \right\rangle \\
&= \frac{N_r(N_r-1)}{2} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \left\langle \frac{1}{\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \right\rangle \left\langle \frac{(1+\eta(\theta_k))}{(1+\gamma(\theta_k))} \right\rangle \\
&\approx \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \langle(1+\eta(\theta_k))(1-\gamma(\theta_k)+\gamma^2(\theta_k))\rangle \\
&= \frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \langle 1-\gamma(\theta_k)+\gamma^2(\theta_k)+\eta(\theta_k)-\eta(\theta_k)\gamma(\theta_k)+\eta(\theta_k)\gamma^2(\theta_k)\rangle,
\end{aligned} \tag{52}$$

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

$$\begin{aligned}
\mathrm{var}[\widetilde{w}_{\alpha\beta}](\theta_k) = \sigma^2_{\widetilde{w}_{\alpha\beta}}(\theta_k) &= \mathrm{var}\left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \quad [1-\gamma(\theta_k)+\gamma^2(\theta_k)+\eta(\theta_k)-\eta(\theta_k)\gamma(\theta_k)+\eta(\theta_k)\gamma^2(\theta_k)] \right] \\
&\approx \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \right]^2 \mathrm{var}[1-\gamma(\theta_k)+\eta(\theta_k)] \\
&= \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \right]^2 [\sigma^2_\gamma(\theta_k) + \sigma^2_\eta(\theta_k) - 2\cancelto{0}{\mathrm{cov}(\eta(\theta_k),\gamma(\theta_k))}] \\
&= \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}} \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle}{\langle(\overline{RR})(\theta_k)\rangle} \right]^2 [\langle\gamma^2(\theta_k)\rangle + \langle\eta^2(\theta_k)\rangle]
\end{aligned} \tag{53}$$

where, again, we only keep the terms up to the second order in fluctuations. Here, as derived from Equation (47), we have the second moments of the fluctuations defined as

$$\langle\eta^2(\theta_k)\rangle = \frac{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle - \langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle^2}{\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle^2}, \tag{54}$$

$$\langle\gamma^2(\theta_k)\rangle = \frac{\langle(\overline{RR})(\theta_k)\cdot(\overline{RR})(\theta_k)\rangle - \langle(\overline{RR})(\theta_k)\rangle^2}{\langle(\overline{RR})(\theta_k)\rangle^2}. \tag{55}$$

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 $\langle(\overline{RR})(\theta_k)\rangle$, $\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle$, and $\langle(\overline{DD})_{\alpha\beta}\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle$; we do not need the second moment of the random pair counts, since $\langle\gamma^2\rangle$ is simply the variance of the random data and hence the variance of the Poisson distribution.

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 the first moment of random pairs in Appendix C.1.2; random pairs are uncorrelated in the limit of large N_r, and hence present a simpler case. We then calculate the first moment of correlated data pairs in Appendix C.1.3, followed by the second moment for the correlated data pairs in Appendix C.1.4.

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

$$\langle (\overline{RR})(\theta_k) \rangle = \left\langle \sum_{j<i}^{K} \rho_i \rho_j \bar{\Theta}_{ij,k} \right\rangle = \frac{1}{2} \sum_{i \neq j}^{K} \langle \rho_i \rho_j \rangle \bar{\Theta}_{ij,k}, \tag{56}$$

where we have borrowed the notation introduced in Equation (5) to express the Heavisides. Now, the probability of finding two random points in two cells, chosen without replacement, is

$$\langle \rho_i \rho_j \rangle = \frac{N_r(N_r - 1)}{K(K-1)}, \tag{57}$$

and similar to LS93 Equation (10), we have

$$\sum_{i \neq j}^{K} \bar{\Theta}_{ij,k} = K(K-1) G_p(\theta_k), \tag{58}$$

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 (57) and (58) into Equation (56), we have

$$\langle (\overline{RR})(\theta_k) \rangle = \frac{1}{2} \frac{N_r(N_r-1)}{K(K-1)} [K(K-1) G_p(\theta_k)] = \frac{N_r(N_r-1)}{2} G_p(\theta_k). \tag{59}$$

C.1.3. Data Pairs: First Moment

Here, we have $N_{\rm tot}$ points distributed randomly over the survey area. As in Appendix 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. It follows, given Equation (14), that

$$\langle (\overline{DD})_{\alpha\beta}(\theta_k) \rangle = C_\Omega \left\langle \sum_{i \neq j}^{K} \mathtt{w}_{ij}^{\alpha\beta} \nu_i \nu_j \bar{\Theta}_{ij,k} \right\rangle = C_\Omega \sum_{i \neq j}^{K} \langle \mathtt{w}_{ij}^{\alpha\beta} \rangle \langle \nu_i \nu_j \rangle \bar{\Theta}_{ij,k}, \tag{60}$$

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

$$\langle \nu_i \nu_j \rangle = \frac{N_{\rm tot}(N_{\rm tot} - 1)}{K(K-1)} [1 + w_{\alpha\beta}(\theta_k)]. \tag{61}$$

Therefore, using Equations (58) and (61), Equation (60) becomes

$$\begin{aligned}\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle &= C_\Omega\langle \mathrm{w}_{ij}^{\alpha\beta}\rangle|_{i\neq j}\frac{N_{\rm tot}(N_{\rm tot}-1)}{K(K-1)}\,[1+w_{\alpha\beta}(\theta_k)][K(K-1)G_p(\theta_k)]\\ &= C_\Omega\left[\frac{\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}}{N_{\rm tot}(N_{\rm tot}-1)}\right][1+w_{\alpha\beta}(\theta_k)]G_p(\theta_k)N_{\rm tot}(N_{\rm tot}-1)\\ &= C_\Omega[1+w_{\alpha\beta}(\theta_k)]G_p(\theta_k)\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}.\end{aligned} \tag{62}$$

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.,

$$w_\Omega \equiv \int_\Omega G_p(\theta_k)w_{\alpha\beta}(\theta_k)d\Omega, \tag{63}$$

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

$$\int_\Omega G_p(\theta_k)d\Omega = 1. \tag{64}$$

Therefore, we have

$$\begin{aligned}\int_\Omega\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle\,d\Omega &= \sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\\ \Rightarrow\int_\Omega C_\Omega G_p(\theta_k)[1+w_{\alpha\beta}(\theta_k)]\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta} &= \sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\\ \Rightarrow C_\Omega &= \frac{1}{1+w_\Omega},\end{aligned} \tag{65}$$

where we make use of Equation (64). Therefore, Equation (62) becomes

$$\langle(\overline{DD})_{\alpha\beta}(\theta_k)\rangle = G_p(\theta_k)\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}. \tag{66}$$

C.1.4. Data–Data Pairs

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

$$\begin{aligned}\langle(\overline{DD})_{\alpha\beta}\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle &= \left\langle\sum_{j\neq i}^{K}\mathrm{w}_{ij}^{\alpha\beta}\nu_i\nu_j\bar{\Theta}_{ij,k}\sum_{l\neq m}^{K}\mathrm{w}_{ml}^{\alpha\beta}\nu_m\nu_l\bar{\Theta}_{ml,k}\right\rangle\\ &= \sum_{j\neq i}^{K}\sum_{l\neq m}^{K}\langle\nu_i\nu_j\nu_m\nu_l\rangle\langle\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{ml}^{\alpha\beta}\rangle\bar{\Theta}_{ij,k}\bar{\Theta}_{ml,k}.\end{aligned} \tag{67}$$

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 Appendix C.1.3):

1. No indices overlap: there are $K(K-1)(K-2)(K-3)$ cases of the sort as 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

$$\langle\nu_i\nu_j\nu_m\nu_l\rangle = \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)}{K(K-1)(K-2)(K-3)}[1+w_{ij}(\theta_k)+w_{im}(\theta_k)+w_{il}(\theta_k)+w_{jm}(\theta_k)+w_{jl}(\theta_k)+w_{ml}(\theta_k)]. \tag{68}$$

Here, given that 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,

$$\langle \nu_i \nu_j \nu_m \nu_l \rangle = \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)}{K(K-1)(K-2)(K-3)}[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega]. \tag{69}$$

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 $\theta_k \pm d\theta_k/2$. Thus, the total number of quadrilaterals is

$$\sum_{\text{unique}\{i,j,l,m\}}^{K} \bar{\Theta}_{ij,k}\bar{\Theta}_{ml,k} = K(K-1)(K-2)(K-3)G_q(\theta_k), \qquad i \neq j,\, m \neq l. \tag{70}$$

Note that as in Equation (64), $G_q(\theta_k)$ is also normalized to unity, i.e.,

$$\int_\Omega G_q(\theta_k) d\Omega = 1. \tag{71}$$

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

$$\begin{aligned}
\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \rangle_{\rm quad} &= C_{\rm quad} \sum_{j\neq i\neq l\neq m}^{K} \langle \nu_i \nu_j \nu_m \nu_l \rangle \langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \rangle \bar{\Theta}_{ij,k}\bar{\Theta}_{ml,k} \\
&= C_{\rm quad} N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega] G_q(\theta_k) \langle \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \rangle_{i\neq j\neq m\neq l} \\
&= C_{\rm quad} N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega]\; G_q(\theta_k) \left[\frac{\sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta}}{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)(N_{\rm tot}-3)} \right] \\
&= C_{\rm quad}[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega] G_q(\theta_k) \sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta},
\end{aligned} \tag{72}$$

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.,

$$\begin{aligned}
\int \langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \rangle_{\rm quad} d\Omega &= \sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta} \\
\Rightarrow \int \{C_{\rm quad}[1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega] G_q(\theta_k)\} d\Omega &= 1 \\
\Rightarrow C_{\rm quad} = \frac{1}{1 + 2\int w_{\alpha\beta}(\theta_k) G_q(\theta_k) d\Omega + 4w_\Omega} &= \frac{1}{1 \;+\; 2w_{\Omega,q} + 4w_\Omega},
\end{aligned} \tag{73}$$

where we have used Equation (71) and have defined a new mean:

$$w_{\Omega,q} \equiv \int w_{\alpha\beta}(\theta_k) G_q(\theta_k) d\Omega. \tag{74}$$

Therefore,

$$\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \rangle_{\rm quad} = \left[\frac{1 + 2w_{\alpha\beta}(\theta_k) + 4w_\Omega}{1 \;+\; 2w_{\Omega,q} + 4w_\Omega} \right] G_q(\theta_k) \sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathtt{w}_{ij}^{\alpha\beta} \mathtt{w}_{ml}^{\alpha\beta}. \tag{75}$$

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

$$\begin{aligned}
\langle \nu_i \nu_j \nu_m \nu_l \rangle|_{i=m} &= \langle \nu_i^2 \nu_j \nu_l \rangle = \langle \nu_i \nu_j \nu_l \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + w_{ij}(\theta_k) + w_{il}(\theta_k) + w_{jl}(\theta_k)] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + 3w_{\alpha\beta}(\theta_k)] \\
\langle \nu_i \nu_j \nu_m \nu_l \rangle|_{i=l} &= \langle \nu_i^2 \nu_l \nu_m \rangle = \langle \nu_i \nu_l \nu_m \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + w_{il}(\theta_k) + w_{im}(\theta_k) + w_{lm}(\theta_k)] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + 3w_{\alpha\beta}(\theta_k)] \\
\langle \nu_i \nu_j \nu_m \nu_l \rangle|_{j=m} &= \langle \nu_i \nu_j^2 \nu_l \rangle = \langle \nu_i \nu_j \nu_l \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + w_{ij}(\theta_k) + w_{il}(\theta_k) + w_{jl}(\theta_k)] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + 3w_{\alpha\beta}(\theta_k)] \\
\langle \nu_i \nu_j \nu_m \nu_l \rangle|_{j=l} &= \langle \nu_i \nu_j^2 \nu_m \rangle = \langle \nu_i \nu_j \nu_m \rangle \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + w_{ij}(\theta_k) + w_{im}(\theta_k) + w_{jm}(\theta_k)] \\
&= \frac{N_{\rm tot}(N_{\rm tot}-1)(N_{\rm tot}-2)}{K(K-1)(K-2)}[1 + 3w_{\alpha\beta}(\theta_k)], \qquad (76)
\end{aligned}$$

where we note that $\langle \nu \rangle = \langle \nu^2 \rangle = N_{\rm tot}/K$ 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. Thus, the total number of triangles is

$$\sum_{\text{unique}\{i,j,m\};l=i}^{K} \bar{\Theta}_{ij,k}\bar{\Theta}_{ml,k} = K(K-1)(K-2)G_t(\theta_k), \quad i \neq j,\, m \neq i \qquad (77)$$

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

$$\int_\Omega G_t(\theta_k)d\Omega = 1. \qquad (78)$$

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

$$\begin{aligned}
\langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \rangle_{\rm tri} &= C_{\rm tri} N_{\rm tot}(N_{\rm tot}-1)\,(N_{\rm tot}-2) G_t(\theta_k)[1 + 3w_{\alpha\beta}(\theta_k)] \\
&\quad \times \{ \langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta} \rangle_{i=m\neq j\neq l} + \langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta} \rangle_{i=l\neq j\neq m} + \langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta} \rangle_{i\neq j=m\neq l} + \langle \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{ml}^{\alpha\beta} \rangle_{i\neq j=l\neq m} \} \\
&= C_{\rm tri} G_t(\theta_k)[1 + 3w_{\alpha\beta}(\theta_k)] \sum_{i\neq j\neq l}^{N_{\rm tot}} \{ \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{il}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{li}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{jl}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{lj}^{\alpha\beta} \}, \qquad (79)
\end{aligned}$$

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.,

$$\begin{aligned}
&\int \langle (\overline{DD})_{\alpha\beta} \cdot (\overline{DD})_{\alpha\beta}(\theta_k) \rangle_{\rm tri} d\Omega = \sum_{i\neq j\neq l}^{N_{\rm tot}} \{ \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{il}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{li}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{jl}^{\alpha\beta} + \mathtt{w}_{ij}^{\alpha\beta}\mathtt{w}_{lj}^{\alpha\beta} \} \\
&\Rightarrow \int \{ C_{\rm tri}[1 + 3w_{\alpha\beta}(\theta_k)] G_t(\theta_k) \} d\Omega = 1 \\
&\Rightarrow C_{\rm tri} = \frac{1}{1 + 3\int w_{\alpha\beta}(\theta_k) G_t(\theta_k) d\Omega + 3w_\Omega} = \frac{1}{1 + 3w_{\Omega,t}}, \qquad (80)
\end{aligned}$$

where we have used Equation (78) and have defined a new mean:

$$w_{\Omega,t} \equiv \int w_{\alpha\beta}(\theta_k) G_t(\theta_k) d\Omega. \qquad (81)$$

Therefore,

$$\langle(\overline{DD})_{\alpha\beta}\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle_{\rm tri}=\left[\frac{1+3w_{\alpha\beta}(\theta_k)}{1+3w_{\Omega,t}}\right]G_t(\theta_k)\sum_{i\neq j\neq l}^{N_{\rm tot}}\{{\rm w}_{ij}^{\alpha\beta}{\rm w}_{il}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{li}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{jl}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{lj}^{\alpha\beta}\}. \tag{82}$$

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

$$\begin{aligned}\langle\nu_i\nu_j\nu_m\nu_l\rangle_{i=m,j=l}&=\langle\nu_i\nu_j\nu_i\nu_j\rangle=\langle\nu_i^2\nu_j^2\rangle=\langle\nu_i\nu_j\rangle=\frac{N_{\rm tot}(N_{\rm tot}-1)}{K(K-1)}[1+w_{\alpha\beta}(\theta_k)]\\ \langle\nu_i\nu_j\nu_m\nu_l\rangle_{i=l,j=m}&=\langle\nu_i\nu_j\nu_j\nu_i\rangle=\langle\nu_i^2\nu_j^2\rangle=\langle\nu_i\nu_j\rangle=\frac{N_{\rm tot}(N_{\rm tot}-1)}{K(K-1)}[1+w_{\alpha\beta}(\theta_k)].\end{aligned} \tag{83}$$

Here, Equation (58) applies, giving us the contribution to the second moment of the pair counts by the pairs as

$$\begin{aligned}\langle(\overline{DD})_{\alpha\beta}\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle_{\rm pairs}&=C_{\rm pairs}N_{\rm tot}(N_{\rm tot}-1)G_p(\theta_k)[1+w_{\alpha\beta}(\theta_k)]\{\langle{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ml}^{\alpha\beta}\rangle_{i=m\neq j=l}+\langle{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ml}^{\alpha\beta}\rangle_{i=l\neq j=m}\}\\ &=C_{\rm pairs}N_{\rm tot}(N_{\rm tot}-1)G_p(\theta_k)[1+w_{\alpha\beta}(\theta_k)]\{\langle{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ij}^{\alpha\beta}\rangle_{i\neq j}+\langle{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ji}^{\alpha\beta}\rangle_{i\neq j}\}\\ &=C_{\rm pairs}G_p(\theta_k)[1+w_{\alpha\beta}(\theta_k)]\sum_{i\neq j}^{N_{\rm tot}}\{{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ij}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ji}^{\alpha\beta}\},\end{aligned} \tag{84}$$

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.,

$$\begin{aligned}\int\langle(\overline{DD})_{\alpha\beta}\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle_{\rm pairs}d\Omega&=\sum_{i\neq j}^{N_{\rm tot}}\{{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ij}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ji}^{\alpha\beta}\}\\ \Rightarrow\int\{C_{\rm pairs}[1+w_{\alpha\beta}(\theta_k)]G_p(\theta_k)\}d\Omega&=1\\ \Rightarrow C_{\rm pairs}&=\frac{1}{1+w_\Omega},\end{aligned} \tag{85}$$

where we have used Equation (64); this results matches with Equation (65) as it should. Therefore,

$$\langle(\overline{DD})_{\alpha\beta}\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle_{\rm pairs}=G_p(\theta_k)\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\rm tot}}\{{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ij}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ji}^{\alpha\beta}\}. \tag{86}$$

Combining the three cases, i.e., Equations (75), (82), and (86), Equation (67) becomes

$$\begin{aligned}\langle(\overline{DD})_{\alpha\beta}\cdot(\overline{DD})_{\alpha\beta}(\theta_k)\rangle&=\sum_{j\neq i}^{K}\sum_{l\neq m}^{K}\langle\nu_i\nu_j\nu_m\nu_l\rangle\langle{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ml}^{\alpha\beta}\rangle\bar{\Theta}_{ij,k}\bar{\Theta}_{ml,k}\\ &=\left[\frac{1+2w_{\alpha\beta}(\theta_k)+4w_\Omega}{1+2w_{\Omega,q}+4w_\Omega}\right]G_p(\theta_k)^2\sum_{i\neq j\neq m\neq l}^{N_{\rm tot}}{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ml}^{\alpha\beta}\\ &+\left[\frac{1+3w_{\alpha\beta}(\theta_k)}{1+3w_{\Omega,t}}\right]G_t(\theta_k)\sum_{i\neq j\neq l}^{N_{\rm tot}}\{{\rm w}_{ij}^{\alpha\beta}{\rm w}_{il}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{li}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{jl}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{lj}^{\alpha\beta}\}\\ &+G_p(\theta_k)\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\rm tot}}\{{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ij}^{\alpha\beta}+{\rm w}_{ij}^{\alpha\beta}{\rm w}_{ji}^{\alpha\beta}\},\end{aligned} \tag{87}$$

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

C.1.5. Fluctuations

Now, substituting Equations (66) and (87) in Equation (54), we have

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

$$=\frac{\begin{array}{l}\left[\frac{1+2w_{\alpha\beta}(\theta_k)+4w_\Omega}{1\ +\ 2w_{\Omega,q}+4w_\Omega}\right]\sum_{i\neq j\neq m\neq l}^{N_{\rm tot}} \mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{ml}^{\alpha\beta}\\ +\left[\frac{1+3w_{\alpha\beta}(\theta_k)}{1+3w_{\Omega,t}}\right]\frac{G_t(\theta_k)}{G_p^2(\theta_k)}\sum_{i\neq j\neq l}^{N_{\rm tot}}\{\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{il}^{\alpha\beta}+\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{li}^{\alpha\beta}+\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{jl}^{\alpha\beta}+\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{lj}^{\alpha\beta}\}\\ +\frac{1}{G_p(\theta_k)}\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\rm tot}}\{\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{ij}^{\alpha\beta}+\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{ji}^{\alpha\beta}\}\end{array}}{\left(\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\right)^2}-1. \tag{88}$$

As for $\langle\gamma^2(\theta_k)\rangle$, given Equation (59), it takes the form

$$\langle\gamma^2(\theta_k)\rangle = \frac{2}{N_r(N_r-1)G_p(\theta_k)}. \tag{89}$$

C.1.6. Variance

We now go back to Equation (53) and attempt to evaluate it. First, substituting Equations (66) and (59), we have

$$\begin{aligned}\sigma^2_{\widehat{w}_{\alpha\beta}}(\theta_k) &= \left[\frac{N_r(N_r-1)}{2\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}}\frac{G_p(\theta_k)\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}}{\frac{N_r(N_r-1)}{2}G_p(\theta_k)}\right]^2[\langle\gamma^2(\theta_k)\rangle+\langle\eta^2(\theta_k)\rangle]\\ &= \left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]^2[\langle\gamma^2(\theta_k)\rangle+\langle\eta^2(\theta_k)\rangle].\end{aligned} \tag{90}$$

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

$$\sigma^2_{\widehat{w}_{\alpha\beta}}(\theta_k)\xrightarrow[\text{large } N_r]{}\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]^2\langle\eta^2(\theta_k)\rangle, \tag{91}$$

where $\langle\eta^2(\theta_k)\rangle$ is given by Equation (88). The expression can be simplified: we first look at the leading-order term, i.e., the quadrilateral contribution:

$$\sigma^2_{\widehat{w}_{\alpha\beta}}(\theta_k)\xrightarrow[\text{order}]{\text{leading}}\frac{\left[\frac{1+2w_{\alpha\beta}(\theta_k)+4w_\Omega}{1\ +\ 2w_{\Omega,q}+4w_\Omega}\right]\sum_{i\neq j\neq m\neq l}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{ml}^{\alpha\beta}}{\left(\left[\frac{1+w_{\alpha\beta}(\theta_k)}{1+w_\Omega}\right]\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\right)^2}-1. \tag{92}$$

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

$$\sigma^2_{\widehat{w}_{\alpha\beta}}(\theta_k)\xrightarrow[\text{correlations}]{\text{weak}}\frac{\sum_{i\neq j\neq m\neq l}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\mathrm{w}_{ml}^{\alpha\beta}}{\left(\sum_{i\neq j}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\right)^2}-1, \tag{93}$$

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

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 (18), which is valid for when there are two galaxy types in our observed sample, we essentially have the unbiased estimator for the AA autocorrelation function as

$$\widehat{w}_{AA}(\theta_k) = C_{AA}(\theta_k)\widetilde{w}^{\rm obs}_{AA}(\theta_k) + C_{AB}(\theta_k)\widetilde{w}^{\rm obs}_{AB}(\theta_k) + C_{BB}(\theta_k)\widetilde{w}^{\rm obs}_{BB}(\theta_k), \tag{94}$$

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 (18). Given the dependency of all terms and factors on the pair weights, we have the variance of the unbiased estimator as

$$\begin{aligned}\sigma^2_{\widehat{w}_{AA}}(\theta_k) = {} & C^2_{AA}(\theta_k)\sigma^2_{\widetilde{w}_{AA}}(\theta_k) + C^2_{AB}(\theta_k)\sigma^2_{\widetilde{w}_{AB}}(\theta_k) + C^2_{BB}(\theta_k)\sigma^2_{\widetilde{w}_{BB}}(\theta_k) \\ & - 2{\rm cov}[C_{AA}(\theta_k),\, \widetilde{w}^{\rm obs}_{AA}(\theta_k)] - 2{\rm cov}[C_{AB}(\theta_k),\, \widetilde{w}^{\rm obs}_{AB}(\theta_k)] - 2{\rm cov}[C_{BB}(\theta_k),\, \widetilde{w}^{\rm obs}_{BB}(\theta_k)] \\ & - 2\widetilde{w}^{\rm obs}_{AA}(\theta_k)\widetilde{w}^{\rm obs}_{AB}(\theta_k){\rm cov}[C_{AA}(\theta_k),\, C_{AB}(\theta_k)] - 2\widetilde{w}^{\rm obs}_{AA}(\theta_k)\widetilde{w}^{\rm obs}_{BB}(\theta_k){\rm cov}[C_{AA}(\theta_k),\, C_{BB}(\theta_k)] \\ & - 2\widetilde{w}^{\rm obs}_{AB}(\theta_k)\widetilde{w}^{\rm obs}_{BB}(\theta_k){\rm cov}[C_{AB}(\theta_k),\, C_{BB}(\theta_k)].\end{aligned} \tag{95}$$

This expression is unwieldy to evaluate for the general case, even if when we use the leading-order, weak-correlation approximation as in Equation (93). Therefore, we resort to numerical estimation of the variance.

C.2. `Weighted` Estimator: Practical Notes

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 (13). Specifically considering Equation (14) for the autocorrelation, we have

$$(\widetilde{DD})_{AA}(\theta_k) = \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ij}^{AA}\bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ij}^{AA}}, \tag{96}$$

while for the cross, we have

$$(\widetilde{DD})_{AB}(\theta_k) = \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ij}^{AB}\bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ij}^{AB}}. \tag{97}$$

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

$${\rm w}_{ij}^{AB} = {\rm w}_{ji}^{BA}, \tag{98}$$

and since the sums are reindexable, we have

$$(\widetilde{DD})_{BA}(\theta_k) = \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ij}^{BA}\bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ij}^{BA}} = \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ji}^{AB}\bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}} {\rm w}_{ji}^{AB}} = (\widetilde{DD})_{AB}(\theta_k). \tag{99}$$

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

C.2.2. Pair Weights

While we have used simple pair weights in this work, i.e., ${\rm w}_{ij}^{\alpha\beta} = {\rm q}_i^{\alpha}{\rm q}_j^{\beta}$, the `Weighted` estimator presented in Equation (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 (14): it requires summing over all the pair weights—a task that is computationally prohibitive when working with large data sets 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 (13) with its full

details, i.e.,

$$\widehat{w}^{\rm obs}_{\alpha\beta}(\theta_k)+1=\frac{(\widetilde{DD})_{\alpha\beta}(\theta_k)}{RR(\theta_k)}=\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\bar{\Theta}_{ij,k}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}}\frac{\sum_i^{N_r}\sum_{j\neq i}^{N_r}}{\sum_i^{N_r}\sum_{j\neq i}^{N_r}\bar{\Theta}_{ij,k}}$$
$$=\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\bar{\Theta}_{ij,k}}{\sum_i^{N_r}\sum_{j\neq i}^{N_r}\bar{\Theta}_{ij,k}}\frac{\sum_i^{N_r}\sum_{j\neq i}^{N_r}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}},\tag{100}$$

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, given that 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.,

$$\widehat{w}^{\rm obs}_{\alpha\beta}(\theta_k)+1=\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\bar{\Theta}_{ij,k}}{\sum_i^{N_r}\sum_{j\neq i}^{N_r}\bar{\Theta}_{ij,k}}\frac{\sum_m^{N_{\rm bin}}\sum_i^{N_r}\sum_{j\neq i}^{N_r}\bar{\Theta}_{ij,m}}{\sum_m^{N_{\rm bin}}\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{\alpha\beta}\bar{\Theta}_{ij,m}},\tag{101}$$

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

C.3. Direct Decontamination

Here, we attempt to find weights that allow us to decontaminate *while* estimating the correlations—a step toward optimal weights. To achieve this, we consider Equation (17) which is reproduced here for convenience:

$$\begin{bmatrix}\langle\widehat{w}^{\rm obs}_{AA}(\theta_k)\rangle\\ \langle\widehat{w}^{\rm obs}_{AB}(\theta_k)\rangle\\ \langle\widehat{w}^{\rm obs}_{BB}(\theta_k)\rangle\end{bmatrix}=\begin{bmatrix}\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}\mathrm{q}_i^A\mathrm{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}\{\mathrm{q}_i^A\mathrm{q}_j^B+\mathrm{q}_i^B\mathrm{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}\mathrm{q}_i^B\mathrm{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}}\\ \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}\mathrm{q}_i^A\mathrm{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}\{\mathrm{q}_i^A\mathrm{q}_j^B+\mathrm{q}_i^B\mathrm{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}\mathrm{q}_i^B\mathrm{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}}\\ \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{BB}\mathrm{q}_i^A\mathrm{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{BB}\{\mathrm{q}_i^A\mathrm{q}_j^B+\mathrm{q}_i^B\mathrm{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{BB}} & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{BB}\mathrm{q}_i^B\mathrm{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{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{102}$$

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

$$\begin{bmatrix}\langle\widehat{w}^{\rm obs}_{AA}(\theta_k)\rangle\\ \langle\widehat{w}^{\rm obs}_{AB}(\theta_k)\rangle\\ \langle\widehat{w}^{\rm obs}_{BB}(\theta_k)\rangle\end{bmatrix}=\begin{bmatrix}\frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}\mathrm{q}_i^A\mathrm{q}_j^A}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AA}} & 0 & 0\\ 0 & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}\{\mathrm{q}_i^A\mathrm{q}_j^B+\mathrm{q}_i^B\mathrm{q}_j^A\}}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{AB}} & 0\\ 0 & 0 & \frac{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{BB}\mathrm{q}_i^B\mathrm{q}_j^B}{\sum_i^{N_{\rm tot}}\sum_{j\neq i}^{N_{\rm tot}}\mathrm{w}_{ij}^{BB}}\end{bmatrix}\begin{bmatrix}w^{\rm true}_{AA}(\theta_k)\\ w^{\rm true}_{AB}(\theta_k)\\ w^{\rm true}_{BB}(\theta_k)\end{bmatrix}.\tag{103}$$

To consider a simple scenario, we first assume that the pair weights are a linear product of the weights of individual weights, i.e., $\mathrm{w}_{ij,\rm opt}^{\alpha\beta}=\mathrm{w}_{i,\rm opt}^{\alpha}\mathrm{w}_{j,\rm opt}^{\beta}$, which follows that we only need to find $\mathrm{w}_{i,\rm opt}^{\alpha}$ and $\mathrm{w}_{i,\rm opt}^{\beta}$ (where we note α, β can be either A or B). Then, we must have the nondiagonal terms in Equation (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:

$$\mathrm{w}_{i,\rm opt}^{\alpha}=\mu^{\alpha}+\nu^{\alpha}q_i^{\alpha},\tag{104}$$

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 two-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 nonlinear functions of probability, meriting further investigation as part of a larger investigation of optimizing the weights.

Appendix D
Generalized Estimators

D.1. `Decontaminated` Estimator

As an extension of our derivation for two samples in Section 3.1, we now consider three samples, with galaxies of Types A, B, C present in our sample. For instance, we have

$$\begin{aligned} w_{AB}^{\rm obs}(\theta_k) = {} & f_{AA} f_{BA} w_{AA}^{\rm true}(\theta_k) + \{f_{AA} f_{BB} + f_{AB} f_{BA}\} w_{AB}^{\rm true}(\theta_k) + f_{AB} f_{BB} w_{BB}^{\rm true}(\theta_k) \\ & + \{f_{AB} f_{BC} + f_{AC} f_{BB}\} w_{BC}^{\rm true}(\theta_k) + f_{AC} f_{BC} w_{CC}^{\rm true}(\theta_k) + \{f_{AA} f_{BC} + f_{AC} f_{BA}\} w_{CA}^{\rm true}(\theta_k). \end{aligned} \tag{105}$$

Therefore, similar to the construction of Equation (12), we have

$$\begin{bmatrix} \widehat{w}_{AA}(\theta_k) \\ \widehat{w}_{AB}(\theta_k) \\ \widehat{w}_{BB}(\theta_k) \\ \widehat{w}_{BC}(\theta_k) \\ \widehat{w}_{CC}(\theta_k) \\ \widehat{w}_{CA}(\theta_k) \end{bmatrix} = \begin{bmatrix} \varsigma_{AA}^{AA} & 2\varsigma_{AB}^{AA} & \varsigma_{AB}^{AB} & 2\varsigma_{AC}^{AB} & \varsigma_{AC}^{AC} & 2\varsigma_{AC}^{AA} \\ \varsigma_{BA}^{AA} & \varsigma_{BB}^{AA} + \varsigma_{AB}^{BA} & \varsigma_{AB}^{BB} & \varsigma_{BC}^{AB} + \varsigma_{AC}^{BB} & \varsigma_{BC}^{AC} & \varsigma_{BC}^{AA} + \varsigma_{AC}^{BA} \\ \varsigma_{BA}^{BA} & 2\varsigma_{BA}^{BB} & \varsigma_{BB}^{BB} & 2\varsigma_{BC}^{BB} & \varsigma_{BC}^{BC} & 2\varsigma_{BC}^{BA} \\ \varsigma_{CB}^{BA} & \varsigma_{CB}^{BA} + \varsigma_{BB}^{CA} & \varsigma_{CB}^{BB} & \varsigma_{CC}^{BB} + \varsigma_{BC}^{CB} & \varsigma_{CC}^{BC} & \varsigma_{BC}^{BA} + \varsigma_{BC}^{BA} \\ \varsigma_{CA}^{CA} & 2\varsigma_{CB}^{CA} & \varsigma_{CB}^{CB} & 2\varsigma_{CC}^{CB} & \varsigma_{CC}^{CC} & 2\varsigma_{CC}^{CA} \\ \varsigma_{CA}^{AA} & \varsigma_{CB}^{AA} + \varsigma_{AB}^{CA} & \varsigma_{CB}^{AB} & \varsigma_{CC}^{AB} + \varsigma_{AC}^{CB} & \varsigma_{CC}^{AC} & \varsigma_{CC}^{AA} + \varsigma_{AC}^{CA} \end{bmatrix}^{-1} \begin{bmatrix} w_{AA}^{\rm obs}(\theta_k) \\ w_{AB}^{\rm obs}(\theta_k) \\ w_{BB}^{\rm obs}(\theta_k) \\ w_{BC}^{\rm obs}(\theta_k) \\ w_{CC}^{\rm obs}(\theta_k) \\ w_{CA}^{\rm obs}(\theta_k) \end{bmatrix}, \tag{106}$$

where we have defined the following for brevity:

$$\varsigma_{mn}^{ij} = f_{A_i A_j} f_{A_m A_n} = \varsigma_{ij}^{mn}. \tag{107}$$

Extending the idea to M samples, we can write the analog of the unbiased estimator for `Decontamination`, given by Equation (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}^{\rm obs}(\theta_k) \\ w_{12}^{\rm obs}(\theta_k) \\ \vdots \\ w_{\gamma\gamma}^{\rm obs}(\theta_k) \\ w_{\gamma(\gamma+1)}^{\rm obs}(\theta_k) \\ \vdots \\ w_{MM}^{\rm obs}(\theta_k) \\ w_{M1}^{\rm obs}(\theta_k) \end{bmatrix}. \tag{108}$$

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

$$\left[\sigma^2_{\widehat{w}_{A_1A_1}} \sigma^2_{\widehat{w}_{A_1A_2}} \cdots \sigma^2_{\widehat{w}_{A_\gamma A_\gamma}} \sigma^2_{\widehat{w}_{A_\gamma A_{\gamma+1}}} \cdots \sigma^2_{\widehat{w}_{A_M A_M}} \sigma^2_{\widehat{w}_{A_M A_1}}\right]^T = \{[D_S^{\rm gen}]^{-1}\}^2_{ij} \left[\sigma^2_{w^{\rm obs}_{A_1A_1}} \sigma^2_{w^{\rm obs}_{A_1A_2}} \cdots \sigma^2_{w^{\rm obs}_{A_\gamma A_\gamma}} \sigma^2_{w^{\rm obs}_{A_\gamma A_{\gamma+1}}} \cdots \sigma^2_{w^{\rm obs}_{A_M A_M}} \sigma^2_{w^{\rm obs}_{A_M A_1}}\right]^T, \tag{109}$$

where $[D_S^{\rm gen}]$ is the square matrix in Equation (108), and as in Appendix A.3, $\{[D_S^{\rm gen}]^{-1}\}^2_{ij}$ denotes that matrix resulting from squaring each individual coefficient in the matrix $[D_S^{\rm gen}]^{-1}$. The covariance matrix for the M-samples case follows the derivation in Equation (34), with all of its assumptions.

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

$$\begin{bmatrix} \widehat{w}_{AA}(\theta_k) \\ \widehat{w}_{AB}(\theta_k) \\ \widehat{w}_{BB}(\theta_k) \\ \widehat{w}_{BC}(\theta_k) \\ \widehat{w}_{CC}(\theta_k) \\ \widehat{w}_{CA}(\theta_k) \end{bmatrix} = \begin{bmatrix} \varkappa_{AA}^{AA} & 2\varkappa_{AB}^{AA} & \varkappa_{AB}^{AB} & 2\varkappa_{AC}^{AB} & \varkappa_{AC}^{AC} & 2\varkappa_{AC}^{AA} \\ \varkappa_{BA}^{AA} & \varkappa_{BB}^{AA} + \varkappa_{AB}^{BA} & \varkappa_{AB}^{BB} & \varkappa_{BC}^{AB} + \varkappa_{AC}^{BB} & \varkappa_{BC}^{AC} & \varkappa_{BC}^{AA} + \varkappa_{AC}^{BA} \\ \varkappa_{BA}^{BA} & 2\varkappa_{BA}^{BB} & \varkappa_{BB}^{BB} & 2\varkappa_{BC}^{BB} & \varkappa_{BC}^{BC} & 2\varkappa_{BC}^{BA} \\ \varkappa_{CB}^{BA} & \varkappa_{CB}^{BA} + \varkappa_{BB}^{CA} & \varkappa_{CB}^{BB} & \varkappa_{CC}^{BB} + \varkappa_{BC}^{CB} & \varkappa_{CC}^{BC} & \varkappa_{BC}^{BA} + \varkappa_{BC}^{BA} \\ \varkappa_{CA}^{CA} & 2\varkappa_{CB}^{CA} & \varkappa_{CB}^{CB} & 2\varkappa_{CC}^{CB} & \varkappa_{CC}^{CC} & 2\varkappa_{CC}^{CA} \\ \varkappa_{CA}^{AA} & \varkappa_{CB}^{AA} + \varkappa_{AB}^{CA} & \varkappa_{CB}^{AB} & \varkappa_{CC}^{AB} + \varkappa_{AC}^{CB} & \varkappa_{CC}^{AC} & \varkappa_{CC}^{AA} + \varkappa_{AC}^{CA} \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{w}_{AA}^{\rm obs}(\theta_k) \\ \widetilde{w}_{AB}^{\rm obs}(\theta_k) \\ \widetilde{w}_{BB}^{\rm obs}(\theta_k) \\ \widetilde{w}_{BC}^{\rm obs}(\theta_k) \\ \widetilde{w}_{CC}^{\rm obs}(\theta_k) \\ \widetilde{w}_{CA}^{\rm obs}(\theta_k) \end{bmatrix}, \tag{110}$$

where we have defined the following for brevity:

$$\varkappa^{uv}_{mn} = \frac{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{A_u A_v} \mathrm{q}_i^{A_m} \mathrm{q}_j^{A_n}}{\sum_i^{N_{\rm tot}} \sum_{j\neq i}^{N_{\rm tot}} \mathrm{w}_{ij}^{A_u A_v}}. \tag{111}$$

Extending the idea to M samples, we can write the analog of our unbiased estimator for `Decontaminated Weighted`, given by Equation (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^{11}_{12} & \dots & \varkappa^{1\gamma}_{1\gamma} & 2\varkappa^{1\gamma}_{1(\gamma+1)} & \dots & \varkappa^{1M}_{1M} & 2\varkappa^{11}_{1M} \\ \varkappa^{11}_{21} & \varkappa^{11}_{22} + \varkappa^{21}_{12} & \dots & \varkappa^{1\gamma}_{2\gamma} & \varkappa^{1\gamma}_{2(\gamma+1)} + \varkappa^{2\gamma}_{1(\gamma+1)} & \dots & \varkappa^{1M}_{2M} & \varkappa^{1M}_{21} + \varkappa^{2M}_{11} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots \\ \varkappa^{\gamma 1}_{\gamma 1} & 2\varkappa^{\gamma 1}_{\gamma 2} & \dots & \varkappa^{\gamma\gamma}_{\gamma\gamma} & 2\varkappa^{\gamma\gamma}_{\gamma(\gamma+1)} & \dots & \varkappa^{\gamma M}_{\gamma M} & 2\varkappa^{\gamma M}_{\gamma 1} \\ \varkappa^{\gamma 1}_{(\gamma+1)1} & \varkappa^{\gamma 1}_{(\gamma+1)2} + \varkappa^{(\gamma+1)1}_{\gamma 2} & \dots & \varkappa^{\gamma\gamma}_{(\gamma+1)\gamma} & \varkappa^{\gamma\gamma}_{(\gamma+1)(\gamma+1)} + \varkappa^{(\gamma+1)\gamma}_{\gamma(\gamma+1)} & \dots & \varkappa^{M}_{(\gamma+1)M} & \varkappa^{M}_{(\gamma+1)1} + \varkappa^{(\gamma+1)M}_{\gamma 1} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots \\ \varkappa^{M1}_{M1} & 2\varkappa^{M1}_{M2} & \dots & \varkappa^{M\gamma}_{M\gamma} & 2\varkappa^{M\gamma}_{M(\gamma+1)} & \dots & \varkappa^{MM}_{MM} & 2\varkappa^{MM}_{M1} \\ \varkappa^{M1}_{11} & \varkappa^{M1}_{12} + \varkappa^{11}_{M2} & \dots & \varkappa^{M\gamma}_{1\gamma} & \varkappa^{M\gamma}_{1(\gamma+1)} + \varkappa^{1\gamma}_{M(\gamma+1)} & \dots & \varkappa^{MM}_{1M} & \varkappa^{MM}_{11} \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{w}^{\rm obs}_{11}(\theta_k) \\ \widetilde{w}^{\rm obs}_{12}(\theta_k) \\ \vdots \\ \widetilde{w}^{\rm obs}_{\gamma\gamma}(\theta_k) \\ \widetilde{w}^{\rm obs}_{\gamma(\gamma+1)}(\theta_k) \\ \vdots \\ \widetilde{w}^{\rm obs}_{MM}(\theta_k) \\ \widetilde{w}^{\rm obs}_{M1}(\theta_k) \end{bmatrix}. \tag{112}$$

ORCID iDs

Humna Awan https://orcid.org/0000-0003-2296-7717
Eric Gawiser https://orcid.org/0000-0003-1530-8713

References

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., et al. 2018, PhRvD, 98, 043526
Addison, G. E., Bennett, C. L., Jeong, D., Komatsu, E., & Weiland, J. L. 2018, arXiv:1811.10668
Armijo, J., Cai, Y.-C., Padilla, N., Li, B., & Peacock, J. A. 2018, MNRAS, 478, 3627
Asorey, J., Carrasco Kind, M., Sevilla-Noarbe, I., Brunner, R. J., & Thaler, J. 2016, MNRAS, 459, 1293
Bailoni, A., Spurio Mancini, A., & Amendola, L. 2017, MNRAS, 470, 688
Balaguera-Antolínez, A., Bilicki, M., Branchini, E., & Postiglione, A. 2018, MNRAS, 476, 1050
Beisbart, C., & Kerscher, M. 2000, ApJ, 545, 6
Benjamin, J., van Waerbeke, L., Ménard, B., & Kilbinger, M. 2010, MNRAS, 408, 1168
Bernstein, G. M. 1994, ApJ, 424, 569
Bianchi, D., Burden, A., Percival, W. J., et al. 2018, MNRAS, 481, 2338
Bianchi, D., & Percival, W. J. 2017, MNRAS, 472, 1106
Blake, C., Achitouv, I., Burden, A., & Rasera, Y. 2019, MNRAS, 482, 578
Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., & Fosalba, P. 2015, MNRAS, 447, 646
Carretero, J., Tallada, P., Casals, J., et al. 2017, in Proc. Sci. 314, The European Physical Society Conference on High Energy Physics, ed. P. Collins & K. Collins (Trieste: SISSA), 488
Chaves-Montero, J., Angulo, R. E., & Hernández-Monteagudo, C. 2018, MNRAS, 477, 3892
Connolly, A. J., Scranton, R., Johnston, D., et al. 2002, ApJ, 579, 42
Cooray, A., & Sheth, R. 2002, PhR, 372, 1
Crocce, M., Carretero, J., Bauer, A. H., et al. 2016, MNRAS, 455, 4301
Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., & Carretero, J. 2015, MNRAS, 453, 1513
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560
Elsner, F., Leistedt, B., & Peiris, H. V. 2016, MNRAS, 456, 2095
Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23
Fosalba, P., Crocce, M., Gaztañaga, E., & Castander, F. J. 2015a, MNRAS, 448, 2987
Fosalba, P., Gaztañaga, E., Castander, F. J., & Crocce, M. 2015b, MNRAS, 447, 1319
Grasshorn Gebhardt, H. S., Jeong, D., Awan, H., et al. 2018, arXiv:1811.06982
Harker, G., Cole, S., Helly, J., Frenk, C., & Jenkins, A. 2006, MNRAS, 367, 1039
Hernández-Aguayo, C., Baugh, C. M., & Li, B. 2018, MNRAS, 479, 4824
Hill, G. J., Gebhardt, K., Komatsu, E., et al. 2008, in ASP Conf. Ser. 399, Panoramic Views of Galaxy Formation and Evolution, ed. T. Kodama, T. Yamada, & K. Aoki (San Francisco, CA: ASP), 115
Hoffmann, K., Bel, J., Gaztañaga, E., et al. 2015, MNRAS, 447, 1724
Jarvis, M., Bernstein, G., & Jain, B. 2004, MNRAS, 352, 338
Kerscher, M., Szapudi, I., & Szalay, A. S. 2000, ApJL, 535, L13
Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Leistedt, B., Peiris, H. V., Elsner, F., et al. 2016, ApJS, 226, 24
Leung, A. S., Acquaviva, V., Gawiser, E., et al. 2017, ApJ, 843, 130
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
Morrison, C. B., & Hildebrandt, H. 2015, MNRAS, 454, 3121
Newman, J. A. 2008, ApJ, 684, 88
Peacock, J. A., Percival, W. J., & Verde, L. 2004, MNRAS, 347, 645
Pearson, D. W., Samushia, L., & Gagrani, P. 2016, MNRAS, 463, 2708
Peebles, P. 1993, Principles of Physical Cosmology (Princeton, NJ: Princeton Univ. Press)
Percival, W. J., & Bianchi, D. 2017, MNRAS, 472, L40
Robaina, A. R., & Bell, E. F. 2012, MNRAS, 427, 901

Ross, A. J., Banik, N., Avila, S., et al. 2017, MNRAS, 472, 4456
Ross, A. J., Percival, W. J., Sánchez, A. G., et al. 2012, MNRAS, 424, 564
Scranton, R., Johnston, D., Dodelson, S., et al. 2002, ApJ, 579, 48
Shafer, D. L., & Huterer, D. 2015, MNRAS, 447, 2961
Sheth, R. K., Connolly, A. J., & Skibba, R. 2005, arXiv:astro-ph/0511773
Sheth, R. K., & Tormen, G. 2004, MNRAS, 350, 1385
Skibba, R., Sheth, R. K., Connolly, A. J., & Scranton, R. 2006, MNRAS, 369, 68
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv:1503.03757
The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., et al. 2018, arXiv:1809.01669
Vargas-Magaña, M., Bautista, J. E., Hamilton, J. C., et al. 2013, A&A, 554, A131
Villalobos, J. J., Parashar, M., Rodero, I., & Brennan-Tonetta, M. 2018, High Performance Computing at the Rutgers Discovery Informatics Institute, Rutgers University, Technical Report, doi:10.13140/RG.2.2.11579.87846
White, M. 2016, JCAP, 2016, 057
White, M., & Padmanabhan, N. 2009, MNRAS, 395, 2381
Zehavi, I., Blanton, M. R., Frieman, J. A., et al. 2002, ApJ, 571, 172
Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, ApJ, 736, 59
Zhu, F., Padmanabhan, N., & White, M. 2015, MNRAS, 451, 236