Propagating sample variance uncertainties in redshift calibration: simulations, theory, and application to the COSMOS2015 data

Carles Sánchez¹⁰, ^{1*} Marco Raveri, ^{1*} Alex Alarcon² and Gary M. Bernstein¹⁰ ¹Department of Physics & Astronomy, University of Pennsylvania, 209 S. 33rd St, Philadelphia, PA 19104, USA ²HEP Division, Argonne National Laboratory, Lemont, IL 60439, USA

Accepted 2020 August 18. Received 2020 August 13; in original form 2020 May 9

ABSTRACT

Cosmological analyses of galaxy surveys rely on knowledge of the redshift distribution of their galaxy sample. This is usually derived from a spectroscopic and/or many-band photometric calibrator survey of a small patch of sky. The uncertainties in the redshift distribution of the calibrator sample include a contribution from shot noise, or Poisson sampling errors, but, given the small volume they probe, they are dominated by sample variance introduced by large-scale structures. Redshift uncertainties have been shown to constitute one of the leading contributions to systematic uncertainties in cosmological inferences from weak lensing and galaxy clustering, and hence they must be propagated through the analyses. In this work, we study the effects of sample variance on small-area redshift surveys, from theory to simulations to the COSMOS2015 data set. We present a three-step Dirichlet method of resampling a given survey-based redshift calibration distribution to enable the propagation of both shot noise and sample variance uncertainties. The method can accommodate different levels of prior confidence on different redshift sources. This method can be applied to any calibration sample with known redshifts and phenotypes (i.e. cells in a *self-organizing map*, or some other way of discretizing photometric space), and provides a simple way of propagating prior redshift uncertainties into cosmological analyses. As a worked example, we apply the full scheme to the COSMOS2015 data set, for which we also present a new, principled SOM algorithm designed to handle noisy photometric data. We make available a catalogue of the resulting resamplings of the COSMOS2015 galaxies.

Key words: surveys – galaxies: distances and redshifts – galaxies: photometry – cosmology: observations.

1 INTRODUCTION

Imaging (or photometric) galaxy surveys, such as the Sloan Digital Sky Survey (SDSS; York et al. 2000), PanSTARRS (Kaiser, Tonry & Luppino 2000), the Kilo-Degree Survey (KiDS; de Jong et al. 2013), the Dark Energy Survey (DES; Flaugher et al. 2015), the Hyper-Suprime-Cam survey (HSC; Miyazaki et al. 2012), or the Legacy Survey of Space and Time (LSST; LSST Dark Energy Science Collaboration 2012), provide key information about the large-scale structure of the Universe using the weak gravitational lensing and clustering of galaxies, and they constitute one of the most powerful probes for testing cosmological models.

In order to perform unbiased cosmological analyses of imaging surveys it is very important to characterize the redshift distributions n(z) = dN/dz dA of the corresponding galaxy samples, and systematic errors in that characterization may directly lead to biases in the cosmological parameter estimation (Huterer et al. 2006; Cunha et al. 2012; Hildebrandt et al. 2012, 2017; Benjamin et al. 2013; Huterer, Cunha & Fang 2013; Bonnett et al. 2016; Samuroff et al. 2017; Hoyle et al. 2018). Relatedly, recent comparisons between cosmological parameters obtained from imaging surveys (Hildebrandt et al. 2018; Troxel et al. 2018a; Hikage et al. 2019) and the cosmic microwave background (Planck Collaboration VI

2018) have claimed discrepancies of up to 3.2σ in their estimates for the amplitude of density fluctuations in a Lambda cold dark matter (ACDM) universe (Asgari et al. 2019). Even though such discrepancies may be attributed to a failure of the ACDM model (Joudaki et al. 2017), that claim would need significant evidence and thorough testing. Alternatively, some studies suggest it may instead be pointing to systematic biases in the weak lensing analysis methodologies (Troxel et al. 2018b; Asgari et al. 2019; Joudaki et al. 2019; Wright et al. 2019). Moreover, such studies indicate that a major difference in the methodologies of those analyses lies in the redshift calibration, and that this has the potential to produce such discrepancy. Redshift calibration requires substantial improvement for the success of the current and next generations of imaging surveys.

Redshift constraints in photometric surveys usually begin with external data on the redshift distribution of their galaxy sample, which can be considered as a prior on n(z) for subsequent survey analyses. Because spectroscopic or high-quality photometric redshifts are very costly in time and resources, such information typically comes from a small area on the sky and therefore it is subject to both shot noise and sample variance due to the large-scale structure of the Universe. As redshift uncertainties can dominate the error budget in current and future weak lensing analyses, it is very important to propagate such sources of uncertainty into the derived cosmological constraints. However, there is yet no clear way of sampling from that redshift prior while including shot noise and sample variance as sources of

^{*} E-mail: carles.sanchez.alonso@gmail.com (CS); mraveri@sas.upenn.edu (MR)



Figure 1. Graphical representation of the redshift and phenotype description of a galaxy population. The colour scale is the relative density f_{zt} of galaxies as each combination of redshift *z* and phenotype *t*. Apparent is the correlation between phenotypes and redshifts: phenotypes are generally only allowed in a certain redshift range, and some phenotypes will have a tighter redshift distribution than others. This figure shows only a subsample of phenotypes, for easier visualization. The ordering of phenotypes is arbitrary.

uncertainty, and hence these have been frequently overlooked or estimated relying on simplified simulated galaxy catalogues.

There exist ways of estimating redshift distribution uncertainties from the data themselves, using subsampling methods such as bootstrapping. Such methods assume, however, that the subsamples are independent draws of a given random field, which is not true if they are correlated by large-scale structure fluctuations. Recently, some studies have used the Dirichlet distribution to model the information contained in the redshift calibration sample (Leistedt, Mortlock & Peiris 2016; Alarcon et al. 2019; Sánchez & Bernstein 2019), as a way to propagate uncertainties from the prior into a redshift posterior. This is a good choice because the Dirichlet distribution produces samples of a distribution which preserve normalization and have positive elements, among other properties. But again, when the redshift information comes from small patches of the sky, the Dirichlet sampling is only propagating shot noise from the calibrator survey, while if the patches are small enough, sample variance from large-scale structure may be the dominant source of uncertainty (Cunha et al. 2012).

In this paper, we will study the problem of sample variance in redshift estimation in detail, and introduce a number of advancements on several fronts. So far, this problem has mostly been studied in simulations (e.g. Cunha et al. 2012). That has some drawbacks, such as the limitations in redshift range (*N*-body simulations are typically not reliable in a broad redshift range like 0 < z < 5), the fixed cosmology, and the associated statistical uncertainties for the simulation volume. In this work, for the first time, we develop a theoretical estimate of the sample variance contribution to redshift uncertainties, and we validate that using *N*-body simulations. The theoretical estimation has some advantages, such as the unlimited

redshift range and the possibility to explore the effects of supersample covariance, cosmological model dependencies, or redshift-space distortions and lensing magnification. We then introduce a novel sampling method based on the Dirichlet distribution which takes as input the theoretical estimate of sample variance and the redshift-survey catalogue, then produces samples of n(z) distribution that draw from uncertainties due to both the shot noise and the sample variance in the catalogues. This yields the correct sampling of uncertainties in the prior for n(z) and hence propagation of those into cosmology analyses of a photometric survey. The method is based on the phenotype approach described in Sánchez & Bernstein (2019) and Alarcon et al. (2019). Additionally, in our application of the method to the COSMOS2015 data sample (Laigle et al. 2016), we present a new SOM algorithm designed to handle noisy photometric data, to be used in the phenotype characterization of a galaxy population.

This paper is organized as follows. In Section 2, we describe the simulations used in this work and the phenotype approach we will work with throughout the paper. In Section 3, we use those simulations to characterize the shot noise and sample variance uncertainties in the redshift distribution, and write down a parametrization that separates the two contributions. We develop the sample variance contribution from a theoretical perspective in Section 4. In Section 5, we present a sampling scheme that can produce realizations of a redshift distribution including shot noise and sample variance uncertainties. Finally, in Section 6 we validate the results in simulations, and we apply them to real data in Section 7 (using an SOM algorithm described in Appendix A). Conclusions are presented in Section 8.

2 FRAMEWORK AND SIMULATED DATA

We will work in the context of the phenotype redshift approach (Alarcon et al. 2019; Buchs et al. 2019; Sánchez & Bernstein 2019), in which we model the galaxy population as a 2D histogram in redshift z and phenotype t (see Fig. 1 for a graphical description), such that $p(z, t) \equiv f_{zt}$ gives the fraction of the population in each (z, t) bin. N_{zt} will be the counts on that histogram for a finite realized galaxy sample. For the implementation of this scheme, we will use a combination of deep survey observations and selforganizing maps (SOMs). Deep observations are often available for surveys like the DES, Euclid, or LSST via searches for SNe (Abbott et al. 2018; Inserra et al. 2018; Scolnic et al. 2018), and these provide essentially zero-noise photometric measurements and additional filters for galaxies in specific fields (henceforth deep fields, or simply DFs), and provide an empirical sampling of the distribution of galaxies in the observed photometric space. In turn, SOMs provide a data-driven way of mapping and discretizing that observed photometric space (Masters et al. 2015), so that we can use the cells in an SOM trained in the DFs as the definition of our galaxy phenotypes t. In addition, SOMs have been extensively used for redshift studies in the past years (Alarcon et al. 2019; Buchs et al. 2019; Hemmati et al. 2019; Wright et al. 2019), and hence all the results presented in this work can be easily accommodated in all of the current redshift calibration efforts that utilize SOMs for the purpose of redshift calibration.

In this work, as in Alarcon et al. (2019), we will use the public MICE2 simulation,¹ a mock galaxy catalogue created from a light-cone of a dark-matter-only *N*-body simulation that contains

¹The data can be downloaded from CosmoHub (Carretero et al. 2017), https://cosmohub.pic.es/.



Figure 2. Effects of sample variance and shot noise in the redshift distribution of galaxies, as determined from the variation of N_z among the healpixel and random-equivalent sets of simulated spectroscopic surveys, respectively. (*Upper-left panel*): Violin plots with the distribution of each redshift bin in the redshift distributions. (*Lower-left panel*): Distribution of the mean redshifts of each sample redshift distribution. (*Upper-right panel*): Normalized variance in the redshift distributions as a function of redshift. (*Lower-right panel*): Sample variance as a function of redshift, as parametrized in equation (2).

~200 million galaxies over one sky octant (~5000 deg²) and up to z = 1.4. The MICE2 simulation has realistic clustering properties given by a Λ CDM cosmology with parameters $\Omega_{\rm m} = 0.25$, $\Omega_{\rm b} = 0.044$, h = 0.7, $n_{\rm s} = 0.95$, $\Omega_{\Lambda} = 0.75$, $\sigma_8 = 0.8$ and w = -1. The galaxies in the simulation have realistic spectral energy distributions (SEDs) assigned from the COSMOS catalogue (Ilbert et al. 2009) that reproduce the observed colour–magnitude distribution as well as clustering observations as a function of colours and luminosity (see Crocce et al. 2015 for more details). Once the galaxy SED is known, magnitudes are computed based on the luminosity and redshift of the galaxy. The galaxy properties, clustering and lensing in the simulation have been thoroughly validated in Carretero et al. (2015), Fosalba et al. (2015b), Crocce et al. (2015), Fosalba et al. (2015a).

To define phenotypes in the simulation, we create an SOM on a square grid with periodic boundary conditions (the map resembles a torus), similar to the SOM in Masters et al. (2015). The SOM is trained with eight colours, defined as mag -i, where mag = {u, g, r, z, Y, J, H, K}, in a 32 × 32 grid, and presents a median redshift dispersion of 0.030 per cell. This is the same SOM as the deep SOM described in Alarcon et al. (2019); please refer to fig. 2 in that work for a graphical representation. For the redshift part, we discretize the redshift space of the simulation in 42 redshift bins of width 0.03 in the redshift range $0.15 \le z \le 1.41$.

3 SHOT NOISE AND SAMPLE VARIANCE: CHARACTERIZING THE UNCERTAINTIES

The characterization of redshift distributions in large imaging surveys relies on prior information coming from smaller spectroscopic or many-band, deep photometric surveys – we will refer to those as

calibration samples. In order to perform accurate cosmological analyses using such information, those calibration samples should be representative of the entire target population. However, that is very difficult to achieve in practice. Assuming that the calibration samples have the same selection and completeness as the entire survey, there are two statistical reasons why they will still be different than the full sample:

(i) Shot noise: Calibration samples are costly in time and resources, which makes the number of galaxies in them very small compared to the total number of galaxies in a survey. Therefore, *Poisson* fluctuations are significant for them.

(ii) Sample variance: Calibration samples are typically from small sky patches subject to large fluctuations due to the large-scale structure of the Universe, i.e. they may be sitting on some void or cluster of galaxies at a given z and not be a fair representation of the Universe. This effect is known as sample variance, and can introduce fluctuations which are an order of magnitude larger than shot noise (Van Waerbeke et al. 2006; Cunha et al. 2012).

To study the sources of uncertainties that go into redshift priors we construct two sets of simulated spectroscopic redshift samples: each sample in the first set consist of all galaxies from one healpy sky pixel of the simulation (with nside = 2^5), which has an area of $\sim 3.5 \text{deg}^2$. We use 247 of those patches, with a mean number of galaxies of $\approx 56\,000$, and standard deviation of ≈ 3000 . We will refer to these as the *healpixel* samples. Additionally, we construct a set of *random-equivalent* samples (Cunha et al. 2012), by drawing 247 samples, each with 56\,000 galaxies drawn at random from the full simulation. Each of the random-equivalent spectroscopic samples has the same shot noise as the previous case, but without the variance induced by large-scale structure. In the simulation we assume the redshift and

photometry of each galaxy are known exactly, i.e. we use the true values, and then phenotype comes from an SOM cell assignment using that true photometry. In reality, redshifts will typically come from a spectroscopic or high-quality photo-*z* sample, and phenotype will come from an SOM cell placement using deep photometry of those galaxies, so both estimates will have some additional noise in redshift and photometry.

Next, we characterize the imprint that sample variance leaves in redshift constraints coming from calibration samples. This will allow us to develop ways to include such effects into the sampling of the redshift distribution n(z), which will be the subject of Section 5. We will split this section in two parts, separating the effects of sample variance in redshift from those in phenotype.

3.1 Effects in redshift

There is one critical difference between the two sources of uncertainty considered here, especially regarding their importance for redshift inference and calibration. Shot noise depends solely on the number of galaxies in a given redshift bin of the redshift sample, whereas sample variance has additional explicit dependence on redshift due to evolution of the volume elements and the large-scale clustering strength.

We parametrize the redshift distribution of a given patch in the sky by including the contributions from shot noise (Poisson) and sample variance as follows:

$$N_z = \text{Poisson}[\bar{N}f_z(1+\Delta_z)]. \tag{1}$$

Here, N_z is the number of galaxies from the sample in redshift bin z, \overline{N} is the angular average galaxy density, f_z is the shape of the redshift distribution for the whole galaxy population, and Δ_z captures variations in source density at z due to the sample variance effect. If we look at the (normalized) variance of an ensemble of patches due to these effects, we find

$$\frac{\operatorname{Var}(N_z)}{\langle N_z \rangle} = 1 + N f_z \operatorname{Var}(\Delta_z), \tag{2}$$

where $\langle N_z \rangle$ is the average N_z for different patches. The term $Var(\Delta_z)$ corresponds to the contribution from sample variance alone, and is defined to not depend on the galaxy number counts.

Fig. 2 shows the shot noise and sample variance redshift uncertainties for the simulated calibration samples described at the beginning of this section: healpixel samples are shown in red and random-equivalent samples in blue. While the random-equivalent samples only contain shot noise, the healpixel samples also include the contribution from sample variance. In the upper-left panel, we show violin plots with the distribution of N_z 's for each redshift bin. In the upper-right panel we show the normalized variance, equation (2), as a function of redshift for the two sets. As it is clear from the plot, the case with sample variance shows a $>10 \times$ larger normalized variance compared to shot noise. The shot noise contribution to the normalized variance shows a flat behaviour in redshift and a steady value of the normalized variance around 1, as expected for a Poisson distribution. Related to this, the lower-right panel of Fig. 2 shows the contribution from sample variance alone, as parametrized in equation (2) by $Var(\Delta_z)$. For this term, the shot noise case is consistent with zero, as expected, and the sample variance case shows a decreasing trend with redshift, as expected for equally sized redshift bins.

From these two different sources of uncertainty, shot noise and sample variance, and the corresponding sets of redshift distributions derived from them, we can compare the uncertainty they introduce in recovering the mean redshift of the population, as this is a very important quantity for weak lensing analyses. We calculate the mean z for each simulated spectroscopic survey, and the lower-left panel of Fig. 2 plots the histogram of mean z's for the members of the healpixel and random-equivalent samples. The healpixel set, which includes sample variance, presents a scatter on the mean redshift of the distribution that is $\approx 10 \times$ larger than that of the random-equivalent sample, which has only shot noise.

3.2 Effects in phenotype

The left-hand panel of Fig. 3 presents the correlation matrix of N_t , the number of spectroscopic galaxies per phenotype t in a spectroscopic survey, for the healpixel simulation set. We observe a strong pattern of correlations between phenotypes, which must be caused by sample variance (i.e. coming from different patches in the sky), since shot noise will induce no inter-phenotype correlations. Sample variance will produce physical clustering of galaxies in redshift, but due to the intrinsic correlation between phenotype and redshift, it will also result in a correlation among different phenotypes that live at similar redshifts.

Since we know that types being at similar redshifts is the cause of their correlation induced by sample variance, we segment the phenotypes into groups with similar redshifts, which we will call *superphenotypes* (*T*). We do this by assigning each phenotype *t* to one of the 42 redshift bins *z* (as defined in Section 2) according to the mean of the p(z|t) as determined by spectroscopy. In the right half of Fig. 3, we show the correlation matrices of superphenotypes *T* that are constructed by joining all phenotypes *t* that are assigned to groups of one, two, three, or six redshift bins (left to right, top to bottom in the figure). That is, in the first case we have as many superphenotypes as redshift bins, in the second case we have half that amount, and so on (as can be seen by the dimension of the correlation matrices). As expected, the correlation matrices look much more diagonal than that in the left half of the figure.

In order to choose a prescription for joining phenotypes to make superphenotypes that are disjoint in redshift, and hence independent under sample variance, we devise the following metric \hat{O}^z . For a given superphenotype definition *T*, i.e. a number of phenotypes with the same mean redshift that are joined into one superphenotype, we estimate the corresponding redshift overlap matrix O_{ij}^z from their redshift distribution p(z, T) as

$$O_{ij}^{z} = \frac{\sum_{z} p(z|T_{i})p(z|T_{j})}{\sqrt{\left(\sum_{z} p(z|T_{i})^{2}\right)\left(\sum_{z} p(z|T_{j})^{2}\right)}}$$
(3)

and compute its determinant corrected by a dimensional factor:

$$\hat{O}^{z} = \operatorname{Det}\left(O_{ij}^{z}\right)^{N^{T}/N^{z}}.$$
(4)

This metric \hat{O}^z , defined to be between 0 and 1, will become closer to unity if the superphenotypes present small overlap in redshift, and will be smaller otherwise. Fig. 4 shows the redshift distributions of superphenotypes in three different prescriptions (joining one, three and six redshift bins), and the corresponding values of the metric. We can see how the metric becomes close to unity for the case where superphenotypes are well separated in redshift. Later on, we will also use this mechanism for devising a sampling method that includes sample variance uncertainties in Section 5.



Figure 3. (Left half): Correlation matrix of the phenotype distribution in different patches of the simulation, showing that sample variance produces strong correlation among phenotypes. The plot shows only a subset (200 out of 1024) of phenotypes, for easier visualization. (Right half): Correlation matrices of superphenotypes that are constructed by joining phenotypes with mean p(z|t) that lie within one, two, three and six redshift bins (left to right, top to bottom).



Figure 4. Redshift distributions for different superphenotype prescriptions used in Fig. 3. The superphenotypes are constructed by joining phenotypes that share one, three, and six average redshift values (top to bottom), and we show the corresponding values of the overlap metric of equation (4). As expected, superphenotypes that show small redshift overlap between them have a metric values close to unity, while overlaping ones show a smaller value.

4 THEORETICAL ESTIMATE OF SAMPLE VARIANCE

In the last section, we characterized the effects of sample variance in redshift priors from calibration samples. Equation (2) describes that variance in the redshift distribution of sky patches given the galaxy density in them, an estimate of the shape of the distribution, and an estimate of the sample variance contribution, independent of galaxy density. The first two of these ingredients can be obtained directly from the data, while the sample variance term, which can be the leading contribution to the redshift uncertainty of calibration patches, cannot be obtained from the data itself, and has only been estimated with simulations so far.

In this section, we discuss how to compute this sample variance contribution for a given cosmological model.

We want to compute the variance of number density fluctuations over the entire calibrator sample that we assume covers a fraction of sky $f_{sky} = A/4\pi$ given its area of A. To simplify the calculation we assume that the calibrator survey geometry is circular so that the angular scale of the survey is related to its area by $A_f = 2\pi (1 - \cos(\theta_f))$. This approximation works well in practice and the results of this section can be straightforwardly extended to more complicated survey geometries if necessary.

The galaxy fluctuation field, coarse grained over an angular scale $\theta_{f,i}$ is the spherical convolution of the fluctuation field with an angular smoothing filter *W*:

$$\hat{\Delta}_{z}(\boldsymbol{n}) = \int W(\boldsymbol{n} \cdot \boldsymbol{n}') \, \Delta_{z}(\boldsymbol{n}') \, \mathrm{d}\Omega', \qquad (5)$$

that we assumed depends only on angular separation $\cos(\theta) = \mathbf{n} \cdot \mathbf{n}'$ between two points.

In this case, similarly to the flat case, the convolution of two functions is equivalent to the product of their Fourier transforms. If we expand the smoothing filter in a Fourier–Legendre series $W(x) = \sum_{\ell} \tilde{W}_{\ell} P_{\ell}(x)$, in terms of Legendre polynomials P_{ℓ} , it can be shown that the smoothed density field can be decomposed in spherical harmonics as

$$\hat{\Delta}_{z}(\boldsymbol{n}) = \sum_{\ell} \frac{4\pi}{2\ell+1} \tilde{W}_{\ell} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\boldsymbol{n}), \tag{6}$$

where $Y_{\ell m}$ are spherical harmonics with coefficients $a_{\ell m}$. Since the smoothing filter depends only on angular separation we can centre the reference vector at the north pole p and compute the coefficients

of the Fourier–Legendre expansion as

$$\tilde{W}_{\ell} = \sqrt{\frac{2\ell+1}{4\pi}} \int W(\boldsymbol{p} \cdot \boldsymbol{n}) Y_{\ell 0}(\boldsymbol{n}) \,\mathrm{d}\Omega.$$
(7)

Here, we are interested in an angular top-hat smoothing filter, $W(x) \propto \Theta(\theta_f - \arccos(x))$ where θ_f is the angular aperture of the tophat. The Fourier–Legendre coefficients can be computed and result in

$$F_{\ell} \equiv \frac{4\pi}{2\ell+1} \tilde{W}_{\ell} = \frac{2\pi}{A_{\rm f}} \frac{P_{\ell-1}(\cos\theta_{\rm f}) - P_{\ell+1}(\cos\theta_{\rm f})}{2\ell+1}.$$
(8)

With these we can compute the angular correlation function of the smoothed galaxy density field at two different redshifts z and z':

$$\langle \hat{\Delta}_{z}(\boldsymbol{n}) \, \hat{\Delta}_{z'}(\boldsymbol{n}') \rangle = \sum_{\ell} \frac{2\ell+1}{4\pi} \, P_{\ell}(\boldsymbol{n} \cdot \boldsymbol{n}') \, F_{\ell}^{2} \mathcal{C}_{\ell}^{zz'}, \tag{9}$$

which is very similar to the standard result for the correlation function in terms of the harmonic power spectrum, C_{ℓ} , of galaxy number counts fluctuations with the difference that multipoles are here weighted differently because of the smoothing filter.

The covariance of the smoothed field is given by equation (9) at zero angular separation:

$$SV(z, z') \equiv Cov(\hat{\Delta}_z, \hat{\Delta}_{z'}) = \sum_{\ell} \frac{2\ell + 1}{4\pi} F_{\ell}^2 C_{\ell}^{zz'}, \qquad (10)$$

which is the equation that we use to compute the sample variance term once the harmonic power spectra are computed for a given cosmological model. For the calculation of the theory galaxy number counts power spectrum we use CAMB (Lewis, Challinor & Lasenby 2000) and follow the discussion in Challinor & Lewis (2011) for the different effects to include in the calculation. We treat the modelling of bias and its redshift dependence separately, as discussed in the next sections, and assume that it is scale independent.

The simulation from which we extract the sample variance measurement that we seek to match is run with a finite volume. This means that part of the sky is masked and we need to include this effect in the theory calculation. The power spectrum for the masked fluctuations, \tilde{C}_{ℓ} , is computed as $\tilde{C}_{\ell} = \sum_{\ell'} M_{\ell\ell'} C_{\ell'}$ (Hivon et al. 2002) where the mode-coupling matrix, $M_{\ell\ell'}$ is computed with Alonso, Sanchez & Slosar (2019).

4.1 Comparison with simulation

From the MICE2 simulation, discussed in Section 2, we can extract redshifts and positions of galaxies in their rest frame, hence neglecting line-of-sight effects distorting both. In this situation sample variance is only sourced by CDM fluctuations.

We start the comparison by considering the ratios of sample variance for different scales of the calibrator survey. With the approximation that bias between CDM and galaxy number-count fluctuations is scale independent, these ratios do not depend on the modeling of bias and in particular do not depend on its redshift evolution.

We consider four areas for the calibrator survey: 13.4, 3.36, 0.893, and 0.21 deg², which corresponds to healpix nside values of 16, 32, 64, and 128, respectively.

Sample variance estimated from the simulations is expected to be noisy and we can compute the expected covariance of sample



Figure 5. Ratio of sample variance at different angular scales. Theory predictions (continuous line) are computed with equation (10) and their variance is computed with equation (11). Measurements from simulation are shown as dots with bootstrap error bars, as discussed in Section 2. Notice that the results shown in this plot are independent of the modeling of bias. Moreover, since cosmological parameters of the simulation are known, there is no free parameter to optimize to obtain these results.

variance as

 $Cov(SV(z_1, z_2), SV(z_3, z_4))$

$$= \sum_{\ell_1,\ell_2} \frac{(2\ell_1+1)(2\ell_2+1)}{(4\pi)^2} F_{\ell_1}^2 F_{\ell_2}^2 \operatorname{Cov}\left(C_{\ell_1}^{z_1z_2}, C_{\ell_2}^{z_3z_4}\right).$$
(11)

In the following we neglect, for simplicity, any non-Gaussian component to the power spectrum covariance that is then given by

$$\operatorname{Cov}\left(C_{\ell_{1}}^{z_{1}z_{2}}, C_{\ell_{2}}^{z_{3}z_{4}}\right) = \delta_{\ell_{1},\ell_{2}} \frac{C_{\ell_{1}}^{z_{1}z_{3}}C_{\ell_{1}}^{z_{2}z_{4}} + C_{\ell_{1}}^{z_{1}z_{4}}C_{\ell_{1}}^{z_{2}z_{3}}}{f_{\mathrm{sky}}(2\ell_{1}+1)},\tag{12}$$

where f_{sky} is the sky fraction covered by the simulation.

In Fig. 5, we show the ratio of sample variances with respect to sample variance at the largest scale. This is a quantity of particular interest since it does not depend on the treatment of linear bias. As we can see all panels show agreement with the theoretical prediction (10), within the error bars, at the 10–15 per cent level.

2989

Notice that there are no free parameters to obtain these results since the cosmology in the simulation is known and (scale-free) galaxy bias cancels in the ratio.

Some of the discrepancy between the theory prediction and the simulation measurement in Fig. 5, especially at high redshift, is due to the approximation that we have made about survey geometry. To test this we have computed the sample variance term with equation (10) but using the harmonic power spectra measured from the simulation finding better agreement.

To compute the full theory prediction of the sample variance we need to model bias and its redshift dependence (one bias value per redshift bin), between galaxies and CDM. This step is crucial to get a reliable estimate of sample variance and hence we devise two ways of doing it that are complementary and can be used to check the reliability of the theoretical prediction.

The first method consists in computing bias from the halo model (see Cooray & Sheth (2002) for a review). Since we do not have very strict accuracy requirements we model the halo mass function following Sheth & Tormen (1999). To connect this to galaxy counts for our magnitude-limited simulation we need to model the conditional luminosity function that we coarsely approximate by assuming that galaxy luminosity is proportional to halo mass. In this case the halo model gives a prediction for the number of galaxies as a function of redshift that depends on two parameters: the proportionality constant between halo mass and luminosity and the effective mass cut of the simulation M^* . We fit the simulation $\tilde{N}(z)$ to get the best estimate for these two parameters and compute bias as the logarithmic derivative of the $\bar{N}(z)$ as a function of the mass cut-off

The second method consists in measuring bias within the calibration survey. To do so we compute the correlation function of galaxies within the small patch and compare it with the theory prediction, after subtracting shot noise from the measured correlation function. If the theory power spectrum is computed with unit bias and assuming that the measured $a_{\ell m}$ are Gaussian distributed then the estimate of bias at each redshift z_i is given by

$$b^{2}(z_{i}) = \frac{\sum_{\ell} (2\ell+1) C_{\ell}^{z_{i} z_{i} \text{ obs}} / C_{\ell}^{z_{i} z_{i}}}{\sum_{\ell} (2\ell+1)},$$
(13)

where, for simplicity, we have neglected the correlation between the power spectra at different redshifts that is negligible for nonneighbouring redshift bins.

The error on the bias determination can be easily computed from the covariance of the observed power spectra in equation (12).

Comparing the two approaches we find that they agree at the 10 per cent level which is sufficient for our application. Better agreement can be likely achieved by refining the theoretical modeling of bias.

With a model for bias we can compute the complete theory prediction for sample variance which, for the four angular sizes that we consider, is shown in Fig. 6. As we can see the agreement between the measurement and theoretical prediction is good at the 10–20 per cent level. The theory line slightly overestimates the measurement mostly due to the modeling of bias.

4.2 Dependence on finite simulation volume

With the theory prediction for sample variance at hand we can test the impact of fluctuations above the scale of the MICE simulation. To do so we need to compute the theory sample variance neglecting the impact of the simulation mask and the way in which it suppresses power. The result is shown in Fig. 7 along with the previous result for comparison, for one of the calibrator survey scale in particular. As we can see the effect of these large-scale modes can be significant and in particular the estimate of sample variance from the simulation, in this case, underestimates it by about 30 per cent. It is therefore advisable to use the theoretical prediction than the simulation values for sample variance when analysing real data.

4.3 Dependence on cosmological model

In this section, we discuss the dependence of sample variance on cosmological parameters. We test the impact of all relevant Λ CDM parameters: matter density Ω_m , the amplitude and tilt of the primordial scalar power spectrum, A_s and n_s respectively, and the Hubble constant H_0 . We find that Ω_m and A_s influence it the most, as shown in Fig. 8. Note that a variation in A_s , in Fig. 8a), is not simply rescaling the entire theory prediction. This happens because bias is not kept fixed. We fix the halo model parameters that produce the bias prediction that would then take into account the variation of the number of observed objects when varying cosmological parameters. A similar effect happens for a variation in Ω_m .

Given the dependence of sample variance on cosmological parameters we recommend using this theory estimate iteratively. At first using a sensible parameter choice to produce the first sample variance estimate. This can be included in parameter estimation pipelines to obtain the cosmological parameters from the probe of interest. These can be used to recompute sample variance and iterate this process to convergence.

4.4 Other effects

In this section, we discuss the dependence of sample variance on other modeling assumptions that we have made.

First of all we consider the impact of non-linear growth of matter perturbations. We show in Panel (a) of Fig. 9 the difference in the SV prediction when using the non-linear modeling in Takahashi et al. (2012) and only linear perturbation theory. As we can see SV in linear theory would be in general smaller and the difference between the two theory predictions increases as redshift decreases as we would expect. As expected this discrepancy strongly depends on the size of the calibrator patch. For small areas the effect is large and between 50 per cent and 10 per cent for the redshifts that we consider. At the two largest scales instead the effect is very small and in general sub per cent.

Then we consider other effects that were neglected in the comparison with simulations in the previous section. These include all relativistic corrections to galaxy number counts fluctuations, discussed in Challinor & Lewis (2011). In particular, these include lensing magnification and redshift space distortions. Magnification bias in this case can be easily extracted from the MICE simulation and included in the calculation of the theory power spectra. As we can see in Panel (b) of Fig. 9 the combination of these effects is negligible for our purposes. Note that the theory prediction is noisy because magnification bias extracted from the simulation is noisy.

5 SAMPLING OVER A REDSHIFT PRIOR

In the previous sections, we have studied shot noise and sample variance as sources of uncertainty in the characterization of a redshift distribution coming from observations in a limited area in the sky. We have found sample variance to be the dominant source of uncertainty, and we have characterized its effects in redshift priors using simulated



Figure 6. Sample variance at different angular scales for the MICE simulation. Theory predictions (continuous line) are computed with equation (10) and their variance is computed with equation (11). Measurements from simulation are shown as dots with bootstrap error bars, as discussed in Section 2.



Figure 7. Sample variance computed with and without the contribution of modes larger than the MICE simulation box, as shown in legend.

calibration samples and theoretical estimation. In this section we will present a method to sample the redshift information from a calibration survey that manifests the proper level of sample variance. In the next two sections we will validate this method in simulations and apply it to data.

We will make extensive use of the Dirichlet distribution for sampling our priors. The Dirichlet distribution is a family of continuous multivariate probability distributions parametrized by a vector of positive reals, or integers, in our case. They are commonly used as prior distributions in Bayesian statistics, exploiting the fact that the Dirichlet distribution is the conjugate prior of the multinomial distribution. This means that when we have a multinomial likelihood and choose the prior to be Dirichlet distributed, the posterior will also be Dirichlet distributed. This makes the posterior sampling straightforward.

In a general example where N is a vector with M elements n_i , i = 1, ..., M, the corresponding fractions f_i given the n_i values are Dirichlet distributed as

$$p(\{f_i\}|\{n_i\}) = \text{Dir}(\{f_i\};\{n_i\}) \propto \prod_{i=1}^M f_i^{n_i-1}.$$
(14)

N can, for example, be the redshift distribution of a galaxy sample coming from a patch in the sky, and then we can say the prior probabilities on $p(z) = f_z$ from that patch follow a Dirichlet distribution Dir(*N*). Conveniently, the Dirichlet sampling will yield samples of f_z that fulfill two required properties of redshift probabilities, namely $f_z > 0 \forall z$ and $\sum_z f_z = 1$.

It is also important to note that the Dirichlet model does not carry any assumption on the nature of the bins of the input vector, so the bins can be of an arbitrary width in z or have different physical interpretations – they are just different categories in our model. None the less, once such categories are defined, for instance in the prior term, they should be kept consistent for the estimation of the posterior.

Two properties of the Dirichlet distribution will be of particular use: first, $\langle f_i \rangle = n_i / n_{\text{tot}}$, with $n_{\text{tot}} \equiv \sum_j n_j$, such that the expectation



Figure 8. Relative differences in sample variance estimates as a result of variations of cosmological parameters. Different lines show different angular scales, as shown in legend.



Figure 9. Relative differences in sample variance estimates using the linear and non-linear matter power spectrum, Panel (a), and magnification, redshift space distortion and other relativistic effects along the observer line of sight, Panel (b). Different lines show different angular scales, as shown in legend.

value is equal to the distribution of the sample set. Second, the variance of f_i is approximately n_i/n_{tot}^2 . The latter means that if we rescale all of the $n_i \rightarrow n_i/\lambda$, then the mean of the Dirichlet distribution is unchanged, but $\operatorname{Var} f_i \rightarrow \lambda \operatorname{Var} f_i$. We will make use of this property below as a means to inflate the shot-noise variation intrinsic to the Dirichlet distribution so that it approximates sample variance plus shot noise. Next we will describe three ways of sampling a redshift distribution: the first two will not be able to capture the uncertainties from sample variance, while the third one will be designed for that purpose.

5.1 Bootstrap sampling (BT)

We can produce samples of a prior p(z, t) distribution by using the bootstrap resampling scheme on a redshift sample with known z, t pairs. We can produce an arbitrary number of calibration samples which are just random resamplings, with repetition, of the original one. Bootstrap resampling will treat galaxies at all redshifts in the same way, and will not resample the sample variance. We will use the label 'BT' for bootstrap resampling.

5.2 Basic Dirichlet sampling (Dir)

In the past (Alarcon et al. 2019; Sánchez & Bernstein 2019), we have drawn samples from the prior using a Dirichlet distribution on the counts M_{zt} of calibration galaxies in joint bins of redshift and phenotype. The Dirichlet distribution treats the 2d array of counts as a single 1d array of categories, so that

$$\operatorname{Dir}(\{f_{zt}\};\{M_{zt}\}) \propto \prod_{z=1}^{N_z} \prod_{t=1}^{N_t} f_{zt}^{M_{zt}-1}.$$
(15)

When doing this, we are assuming all redshifts and phenotypes have the same uncertainties, uncorrelated beyond the constraint $\sum f_{tz} = 1$, i.e. we are not considering sample variance. This 'Dir' case assumes that the prior information on redshift and on phenotype both arise from the same calibration sample. We might, however, often have stronger prior information on the phenotype distribution compared to the redshift distribution, i.e. we have more galaxies with definitive phenotypes (using high-quality photometry) than have definitive redshifts (using spectroscopic or many-band photometric information).

5.3 Three-step Dirichlet sampling (3sDir)

Now we will construct a sampling method based on the Dirichlet model but using the characterization of sample variance from Section 3. Instead of considering the redshift (z) and phenotype (t) parts together as in Section 5.2, now we will also make use of superphenotypes (T) and we will split the problem by making use of the following relations:

$$p(z, t) = p(t|z, T)p(z|T)p(T),$$
 (16)

$$f_{zt} = f_t^{zT} f_z^T f_T, (17)$$

where the mean of each fraction f in the second row is given by its corresponding term in the first row. All f's must be non-negative, and there are sum constraints $\sum_{zt} f_{zt} = \sum_t f_t^{zT} = \sum_z f_z^T = \sum_T f_T = 1$.

The three-step Dirichlet sampling method consists of drawing, in sequence, values of f_T , then f_z^T , then f_t^{zT} from individual Dirichlet distributions, using appropriate source counts from the calibration

survey to define the n_i in each. When sampling distributions that span multiple redshifts (namely f_z^T and f_T) we will, however, rescale the n_i so as to inflate the variance of the Dirichlet distribution to the level expected for the sum of shot noise and sample variance. The key input is the normalized variance from equation (2) which we will label as $\lambda_z \equiv \text{Var}(N_z)/N_z$. The terms on the right-hand side of the equation can be estimated from theory as in Section 4.

The first sampling step is to draw a set of f_T given the counts M_T of each superphenotype in the calibration sample. For this purpose we define $\bar{\lambda} = \langle \lambda_z \rangle$ as the mean ratio of (shot+sample) variance to shot noise. We then draw f_T from a Dirichlet distribution that scales the observed counts in a manner that generates the desired total noise:

(i)
$$f_T \leftarrow \text{Dir}(M_T/\bar{\lambda})$$

Recall that the superphenotypes were chosen to span large enough redshift ranges that their counts are nearly uncorrelated by sample variance. Thus, this noise-inflated Dirichlet draw will approximate the conjugate to the true distribution for M_T , which is essentially a coarse redshift binning of the sources. The method is not exact, in the sense that the real λ_z do change with z (Fig. 2, upper right), while our Dirichlet partition must assume a fixed $\overline{\lambda}$ over the full range.

One important facet of this stage is that it does not require redshift information. The counts M_T can be made over all calibration fields with deep photometry sufficient to assign (super)-phenotypes, without the need for redshift assignments. This allows reduction of shot noise and sample variance in this step – which was noted by Buchs et al. (2019) as the largest noise source in their implementation of phenotypic redshifts.

The next step is to draw values of f_z^T , i.e. distribute the probabilities f_T into redshift bins. We compute for every superphenotype Tthe noise excess $\lambda_T = \sum_z \lambda_z p(z|T) \approx \sum_z \lambda_z M_{zT}/M_T$. A sample of the $f_{zT} = f_z^T f_T$ are then generated by

(ii)
$$f_{zT} \leftarrow \text{Dir}(M_{zT}/\lambda_T) f_T$$
.

Here, M_{zT} are the counts of calibration survey galaxies in both redshift bin *z* and superphenotype *T*. These counts must be drawn from a calibration field with high-quality redshift assignment. A shortcoming of this step is that the Dirichlet distribution assumes no correlation between fluctuations in distinct *z* bins (except those induced by the $\sum f = 1$ constraint), whereas we know that sample variance does have a finite correlation length in *z*. Fig. 14 shows inter-bin density correlations are <0.1 at $\Delta z \ge 0.05$ for $z \le 1$ in the COSMOS2015 fields. Our method will hence not sample the large-scale structure faithfully on scales below $\Delta z \approx 0.05$.

Finally, we wish to draw a sample of the f_t^{zT} probabilities, i.e. to distribute the probability of a given redshift bin among the phenotypes t at that redshift. Because the superphenotypes T are nearly disjoint in redshift, we opt to simplify this process at some small loss in accuracy by summing over superphenotypes, and using $f_t^z = p(t|z)$ instead of p(t|z, T) by executing the following step for each redshift bin z:

(iii)
$$f_{tz} \leftarrow \text{Dir}(M_{tz}) f_z$$
; with $f_z = \sum_T f_{zT}$.

In this case M_{tz} are the galaxy counts in joint redshift-phenotype space, which again requires a high-quality redshift calibration field.

Note that there is no sample-variance inflation factor λ in the final sampling step, since we are assuming that the sample variance is strictly a redshift phenomenon – within a redshift bin, the phenotype assignments are assumed to have only shot noise. This also means that our sampling method is thus assuming that all phenotypes have the same bias.

In summary, this sampling method for f_{zt} works by splitting the redshift and phenotype parts of the problem and takes advantage of our knowledge of sample variance across redshift from the previous Sections by inflating the variance in the redshift axis. This '3sDir' method reduces to the basic Dir method described above when $\lambda_T = 1$ at all *T*.

The splitting into z and t samplers enables lower noise for the case where one has more calibration data with high-quality photometry than with high-quality redshift information. One could generalize this method to allow, for each phenotype, combining calibration fields with differing levels of redshift uncertainty (e.g. spectroscopic versus many-band photometric redshifts). This is left for future work.

6 VALIDATION IN SIMULATIONS

We apply the three sampling strategies described in Section 5 (BT, Dir, and 3sDir) to the sky patches defined on the MICE simulation and we test their accuracy in reproducing the true uncertainties from sample variance. We use the 247 patches of the simulation (described in Section 2). For any single patch, each of the three methods is able to produce a number of realizations of the $f_z = p(z)$ function, and we can test if those realizations accurately describe the uncertainties of the redshift distribution coming from that patch. For the 3sDir method, we utilize the sample variance theory characterization of Section 4 as an input to the method.

As a first test of the sampling methods, we randomly choose two patches from the simulation, and produce realizations of f_z from them using the three different sampling schemes (BT, Dir, and 3sDir). In Fig. 10, the three panels on the left-hand column show the true redshift distribution of the galaxy population (in grey) and violin plots for the sampling of the redshift distribution from the two random patches by the three different methods, BT, Dir, and 3sDir, in the upper, centre, and lower panels, respectively. For the first two cases, bootstrap and basic Dirichlet, the uncertainties coming from the sampling are underestimated and do not cover the true redshift distribution of the population, and both methods perform similarly. The 3sDir method does a much better job, with the sampling containing the true distribution. The panels in the right-hand column show the distribution for the mean of each of the samples of the redshift distribution produced by the three schemes, minus the true mean redshift of the population. The BT and Dir methods yield very tight distributions for the mean redshift difference, but not containing the truth value of zero. In contrast, the 3sDir method shows wider distributions of the mean redshift, containing the truth for both patches.

A more quantitative test of the fidelity of the sampler of f_z is to compute the *pull* distribution of the mean-*z* errors. For each simulated patch *i*, we generate 100 samples j = 1...100 of f_z , then calculate the mean redshift \bar{z}^{ij} for each distribution. For a given patch we can then calculate a mean error \bar{z}^i over the samples and a standard deviation $\Delta(\bar{z})^i$. The 'pull' $(z_{true} - \bar{z}^i)/\Delta(\bar{z})^i$ for the 247 patches should approach a Normal distribution $\mathcal{N}(0, 1)$ if the sampler is properly estimating the total uncertainty in the redshift distribution, and is unbiased, and if the uncertainties are Gaussian. Similarly, we can compute the standard deviation σ_z of the redshift distribution for each sample from each patch; and we can plot the pull distribution of this summary statistic of p(z).

Fig. 11 shows the pull distribution of both \bar{z} and σ_z for each of the three sampling methods. As expected from the test in Fig. 10, both the bootstrap and the basic Dirichlet method perform poorly in the comparison with a Gaussian $\mathcal{N}(0, 1)$. On the other hand, the 3sDir method performs very well in that comparison (see inset panels



Figure 10. Two random patches chosen from the simulation. We show, for each sampling method (BT, Dir and 3sDir), the sampling of the redshift distribution of the patches (left) and its mean redshift relative to the true mean (right). The true redshift distribution of the population is shown in grey, and the true mean redshift as the vertical dashed line. Clearly only the 3sDir method generates samples with uncertainties large enough to include the truth values.



Figure 11. (*Left-hand panel*): Pull distribution for the mean of the redshift distribution \bar{z} for each of the three sampling methods. If uncertainties are well behaved, pulls should approach a Gaussian N(0, 1) (in grey). Inset shows a zoom in for the 3sDir method. (*Right-hand panel*): Same as in left-hand panel, but for the standard deviation of the redshift distribution.

in Fig. 11), demonstrating that the method properly captures the sample variance uncertainties in the mean and width of the redshift distributions. Both the mean and the width of redshift distribution are key properties for galaxy clustering and cosmic shear studies.

Another quantity that we would like our sampling scheme to reproduce well is the (normalized) redshift variance $Var(N_z)/N_z$ as a function of z. That is, we would like the samples to span the same range of N_z as is expected from the combination of shot and

sample variance. As seen in Figs 2 and 6, this quantity has significant redshift dependence in both simulations and theory. Fig. 12 shows the 3sDir method succeeds in producing samples that follow the expected redshift dependence in normalized variance. On the contrary, the BT and Dir methods yield variances consistent with pure shot noise, far below the sample variance level.

With this we have shown how the proposed three-step Dirichlet sampling scheme can reproduce the uncertainties from sample



Figure 12. Normalized variance in the samples of the redshift distribution as a function of redshift, $Var(N_z)/N_z$, for the three sampling methods (violins). Sample variance and shot noise estimates are shown in grey and black.

variance in the redshift distribution, its mean, its width, and its variance as a function of redshift.

7 APPLICATION TO THE COSMOS2015 DATA SET

We now apply the methods of the previous sections to real data, namely the COSMOS2015 catalogue (Laigle et al. 2016), which provides high-quality redshift estimates at high completeness over the colour–magnitude space down to faint magnitudes. COSMOS2015 has played a key role in the redshift characterization of many past and current cosmological analyses using lensing surveys (Bonnett et al. 2016; Hildebrandt et al. 2018; Hoyle et al. 2018; Hamana et al. 2019; Hikage et al. 2019). We produce resamplings of this catalogue (reweightings) that realize uncertainties from both shot noise and sample variance.

7.1 Redshift estimates

The COSMOS2015 catalogue from Laigle et al. (2016) provides photometry in 30 different UV/visible/IR bands, and probability distribution functions (PDFs) p(z) for the redshift of each galaxy based on this photometry using the LePhare template-fitting code (Arnouts et al. 1999; Ilbert et al. 2006). Due to the extensive photometric coverage of the catalogue, the redshift performance is very good: For bright galaxies, the typical p(z) widths are ~0.01(1 + z), with outlier rates around 0.5 per cent, while for fainter, highredshift galaxies it goes to p(z) widths of ~0.023(1 + z) and outlier rates of about 13 per cent. For the results in this work, we use the redshift estimates of ZMINCHI2, which is the 30-band photometric redshift point prediction corresponding to the minimum χ^2 fit between fluxes and templates, and throughout this section, we use a redshift binning of width 0.05 between redshift 0.06 and 5.01, what makes a total of 99 redshift bins.



Figure 13. (*Upper panel*): Redshift distribution for the COSMOS2015 galaxies used in this section. (*Lower panel*): Redshift distributions for the different superphenotypes made for the COSMOS2015 data. This is similar to Fig. 4 in the simulation. The overlap metric of equation (4) for this case is $\hat{O}^z = 0.93$.

7.2 Phenotype characterization

Similar to our treatment of the MICE2 simulations, phenotypes for COSMOS2015 are defined as cells in an SOM which is trained on photometric data. In this case, we use the following photometric bands: {*u*, *B*, *V*, *r*, *ip*, *zp*, *J*, *H*, *Ks*}. We limit our sample to galaxies with mag(*ip*) < 25.5. Quality cuts on the data² yield a total of 305,835 galaxies placed in the SOM. For classifying the COSMOS2015 catalogue, we make substantial changes to the SOM algorithm with the purpose of improving the classification of galaxies of modest S/N, and to allow magnitude information (not just colours) to be used in redshift estimation. These improvements to the SOM algorithm are detailed in Appendix A. For present purposes it suffices to note that the COSMOS2015 galaxies are each assigned to one of $64 \times 64 = 4096$ phenotypes defined by an SOM cell.

As described for the MICE simulation in Section 3.2 and Fig. 4, we divide the COSMOS2015 phenotypes into six *superphenotypes* by partitioning a list of all phenotypes ordered by their mean redshifts. Due to the large redshift range of the COSMOS2015 galaxy sample, we create four equally spaced superphenotypes between redshift 0 and 1.75, and then two-equally spaced high-redshift superphenotypes between redshift 1.75 and 5. The phenotypes resulting from this prescription are shown in Fig. 13, and they have an overlap metric of 0.93, computed as in equation (4).

7.3 Sample variance estimation

We produce a theoretical estimate for sample variance (SV) using the methods of Section 4. For simplicity we fix all cosmological parameters to their best-fitting Planck 2018 values, as obtained by fitting the Λ CDM model to the combination of the Planck measurements of CMB temperature and polarization, Planck CMB

²We select objects with TYPE = 0, FLAG_PETER = 0, B_IMAFLAGS_ISO = 0, B_FLUXERR_APER3>0, where B runs over all photometric bands we use.



Figure 14. Theory prediction of sample correlation as a function of redshift separation for the COSMOS sample. Different colours represent different redshifts, as indicated in the legend.

lensing reconstruction and Baryon Acoustic Oscillations measurements (Aghanim et al. 2018). Specifically these are physical baryon density $\Omega_b h^2 = 0.02242$, physical CDM density $\Omega_c h^2 = 0.11933$, the reduced Hubble constant h = 0.676, the amplitude of the primordial scalar spectrum $10^9 A_s = 2.105$ and its tilt $n_s = 0.9665$. The crucial part in the prediction of SV is the modeling of galaxy bias. As in Section 4.1 we first extract bias from the COSMOS data by fitting the amplitude of the measured correlation function at different redshifts. Then we use the halo model and we fit its parameters to reproduce the observed COSMOS redshift distribution so that we can predict the value and redshift evolution of bias. We find that these two procedures produce results that match to satisfactory accuracy.

After the recipe for galaxy bias is obtained we proceed with the estimate of SV following equation (10) with a smoothing filter obtained assuming circular survey geometry matching the area of the COSMOS field. In this case, since we are not matching a simulation, we produce an estimate of the full sky SV, taking into account contributions coming from all scales bigger than the COSMOS patch. The resulting SV qualitatively follows the previous results in Fig. 6, decreasing as redshift increases.

The cross-correlations between different redshift bins, $\operatorname{Corr}(\hat{\Delta}_z, \hat{\Delta}_{z'}) = \operatorname{SV}(z, z')/[\sqrt{\operatorname{SV}(z)}\sqrt{\operatorname{SV}(z')}]$, are independent of bias modeling and depend on the long modes along the line of sight. Fig. 14 plots predicted correlation coefficients as a function of redshift separation and at different redshifts. At low redshift ($z \sim 1$), sample correlation decays quickly in redshift and becomes negligible at a redshift separation of $\Delta z \sim 0.1$. On the other hand, the decay is slower at higher redshift. The correlation predictions can be used to choose appropriate smoothing functions for the resampled COSMOS2015 redshift distributions, a topic that we will defer to future work.

7.4 Sampling

We can now apply the 3sDir sampling method from Section 5 and produce realizations of the COSMOS2015 redshift distribution that include uncertainties from sample variance and shot noise. For that, we will use as inputs both the phenotype and superphenotype definitions of Section 7.2 and the theory sample variance estimation of Section 7.3. We will apply the sampling method as described in Section 5.3, and we will assign the minimum sample variance contribution to the two high-redshift superphenotypes defined in Section 7.2. Fig. 15 shows the result of the 3sDir sampling method

applied to the COSMOS2015 data sample. The upper-left panel shows the COSMOS2015 redshift distribution together with the mean and standard deviation of the samples in each redshift bin. The lower-left panel shows the normalized redshift variance that was input to the sampling method, compared to the normalized redshift variance of the samples, showing generally good agreement. The right-hand panels show the distributions of the mean redshift difference between the samples and the input COSMOS2015 data, for the entire redshift range and for a subset at lower redshift. In those panels, we also show the distribution corresponding to shot noise only, without sample variance, for comparison. For the entire redshift range (0 < z < 5) the 3sDir method shows a mean redshift dispersion of 0.019 (0.002 for shot noise only), while for the lower redshift subsample (0 < z < 1.5) it goes down to 0.0097 (0.0008 for shot noise only).

We release with this paper a set of 1000 realizations of the COSMOS2015 n(z) produced by the 3sDir method, which anyone can use to propagate the shot and sample variance of COSMOS2015 into their own analyses. This is presented as a table with 305,835 rows (one for each COSMOS galaxy passing the quality and ip < 25.5 cuts); there are columns for the COSMOS2015 ID number of the galaxy, its ZMINCHI2 redshift, and the bin number z and phenotype t to which it is assigned in the SOM. There are then 1000 columns labelled by index j containing weights to apply to the galaxies to realize the jth 3sDir sample. The weights are equal to $w_{ij} = f_{iz}^j * M_{tot}/M_{tz}$, where f_{iz}^j is the value of f_{iz} 3sDir sample for the tz bin to which galaxy i belongs; $M_{tot} = 305$, 385, and M_{iz} is the number of COSMOS2015 galaxies falling into the tz bin. The file containing the table described above can be downloaded in FITS format from this link.³

To use the 3sDir resamplings, a user makes any particular cuts they desire to the galaxy sample; defines their own redshift bins; and then makes a histogram of the galaxies in these bins, weighting by the w_{ij} . Each column will yield an independent sample of the n(z).

8 SUMMARY AND CONCLUSIONS

Redshift uncertainties have become one of the leading contributions to the systematics budget of imaging galaxy surveys. In particular, they are important enough to bias cosmological constraints in a way that different surveys can be at tension between themselves and with results from other probes such as the CMB, and hence it is crucial to properly include them in cosmological analyses.

Such uncertainties in the redshift distribution of a galaxy population arise mostly because of the limited knowledge of the colourredshift relation, which usually comes from prior knowledge in small patches of the sky known as calibration samples. In those samples, galaxy redshifts are known either through spectroscopy or from highquality photometry, and they have associated uncertainties regarding those estimations. In addition, due to the small size of calibration fields, sample variance from large-scale structure becomes an important source of uncertainty (Cunha et al. 2012). In this paper, we have studied in detail the impact of sample variance and shot noise in the uncertainties associated with the redshift distributions of calibration samples, both from theory and *N*-body simulations, and we have proposed a new scheme to produce realizations of redshift distributions including those uncertainties.

In addition, using the techniques described above and a dedicated SOM algorithm, we have applied the scheme to the COSMOS2015 data sample, producing a theory estimate of the sample variance

³https://cosmos2015resampling.shortcm.li/sbrdMS



Figure 15. (*Upper-left panel*): Redshift distribution for the COSMOS2015 galaxies, and 1σ error bars for the samples produced in this work, which include uncertainties from shot noise and sample variance. (*Lower-left panel*): Input normalized redshift variance to the sampling method, together with resulting normalized redshift variance for the samples produced. Uncertainties on the latter come from bootstrap resampling. (*Right-hand panels*): Distributions of mean redshift for the redshift distribution samples produced, both for the whole redshift range (upper), and for a subset at low redshift (z < 1.5) (lower). We also show the distributions corresponding to shot noise only, without sample variance, for comparison.

contribution for that particular data set, and then generating (and making public) 1000 realizations of its redshift distribution that include the effects of shot noise and the estimated sample variance. From those realizations, we compute an uncertainty in the mean redshift of the COSMOS2015 population to be around 2 per cent (\simeq 1 per cent for z < 1.5). That uncertainty is comparable to the redshift uncertainties estimated in cosmological analyses that have used COSMOS2015 as redshift prior (e.g. Hoyle et al. 2018), which highlights the importance of correctly propagating it into cosmological constraints from galaxy surveys.

In summary, this work introduces three main advances to the problem of propagating redshift uncertainties into the analysis of imaging galaxy surveys:

(i) A theoretical approach to estimating the sample variance contribution to the uncertainty in the redshift distributions of calibration fields. A theoretical estimate has several advantages over a direct estimation from *N*-body simulations, such as the unlimited redshift range and the possibility of handling different modeling effects such as cosmology dependence.

(ii) A sampling scheme, based on the Dirichlet distribution, to produce realizations of the redshift distribution of a calibration field given estimates of shot noise (which can be taken directly from the data) and sample variance (given by (i) or by simulations). The realizations are able to correctly reproduce the sample variance impact in the uncertainty in the mean and standard deviation of redshift distributions, and the redshift trend of the number counts uncertainty.

(iii) A new SOM algorithm designed to improve the phenotype classification of galaxies of modest S/N, and to allow magnitude information (not just colours) to be used in redshift estimation.

These new elements of redshift calibration presented in this work provide an appropriate way of including sample variance uncertainties into the redshift uncertainty budget, which was an important missing piece in past analyses. This work will therefore contribute to an improved N(z) characterization in current and future real survey data, which is a key part of the overall systematic uncertainties in future weak lensing and galaxy clustering analyses.

ACKNOWLEDGEMENTS

The authors thank Scott Dodelson for useful comments on the draft, and Cyrille Doux, Lucas Secco, Daniel Gruen, Justin Myles, Joe DeRose, and Alexandra Amon for helpful conversations about this topic. This work was supported by grants AST-1615555 from the US National Science Foundation, and DE-SC0007901 from the U.S. Department of Energy. MR is supported in part by NASA ATP Grant No. NNH17ZDA001N, and by funds provided by the Center for Particle Cosmology. Work at Argonne National Lab is supported by UChicago Argonne LLC, Operator of Argonne National Laboratory (Argonne). Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated under contract no. DE-AC02-06CH11357.

DATA AVAILABILITY

The data underlying this article are available at

(i) MICE2 simulations: publicly available at https://cosmohub.p ic.es/home under Catalogues labelled as MICECAT, version 2.

(ii) COSMOS2015 data: publicly available at http://cosmos.astro .caltech.edu/page/photom under COSMOS2015 (Laigle et al. 2016).

(iii) The produced set of redshift distributions of COSMOS2015 including the effects of sample variance and shot noise, described in Section 7, are made publicly available at https://cosmos2015resamp ling.shortcm.li/sbrdMS.

REFERENCES

- Abbott T. M. C. et al., 2018, ApJS, 239, 18
- Aghanim N. et al., 2018, preprint (arXiv:1807.06209)
- Alarcon A., Sánchez C., Bernstein G. M., Gaztañaga E., 2019, preprint (arXiv: 1910.07127)
- Alonso D., Sanchez J., Slosar A., 2019, MNRAS, 484, 4127
- Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
- Asgari M. et al., 2019, A&A, 624, A134
- Benjamin J. et al., 2013, MNRAS, 431, 1547
- Bonnett C. et al., 2016, Phys. Rev. D, 94, 042005
- Buchs R. et al., 2019, MNRAS, 489, 820

- Carretero J., Castander F. J., Gaztañaga E., Crocce M., Fosalba P., 2015, MNRAS, 447, 646
- Carretero J. et al., 2017, Proc. Sci., The European Physical Society Conference on High Energy Physics – PoS(EPS-HEP2017). SISSA, Trieste, PoS#488
- Challinor A., Lewis A., 2011, Phys. Rev. D, 84, 043516
- Cooray A., Sheth R. K., 2002, Phys. Rep., 372, 1
- Crocce M., Castander F. J., Gaztañaga E., Fosalba P., Carretero J., 2015, MNRAS, 453, 1513
- Cunha C. E., Huterer D., Busha M. T., Wechsler R. H., 2012, MNRAS, 423, 909
- de Jong J. T. A., Kuijken K., Applegate D., Begeman K., Belikov A., al E., 2013, ESO Messenger, 154, 44
- Flaugher B. et al., 2015, AJ, 150, 150
- Fosalba P., Gaztañaga E., Castander F. J., Crocce M., 2015a, MNRAS, 447, 1319
- Fosalba P., Crocce M., Gaztañaga E., Castander F. J., 2015b, MNRAS, 448, 2987
- Hamana T. et al., 2020, PASJ, 72, 16
- Hemmati S. et al., 2019, ApJ, 877, 117
- Hikage C. et al., 2019, PASJ, 71, 43
- Hildebrandt H. et al., 2012, MNRAS, 421, 2355
- Hildebrandt H. et al., 2017, MNRAS, 465, 1454
- Hildebrandt H. et al., 2020, A&A, 633, A69
- Hivon E., Gorski K. M., Netterfield C. B., Crill B. P., Prunet S., Hansen F., 2002, ApJ, 567, 2
- Hoyle B. et al., 2018, MNRAS, 478, 592
- Huterer D., Takada M., Bernstein G., Jain B., 2006, MNRAS, 366, 101
- Huterer D., Cunha C. E., Fang W., 2013, MNRAS, 432, 2945
- Ilbert O. et al., 2006, A&A, 457, 841
- Ilbert O. et al., 2009, ApJ, 690, 1236
- Inserra C. et al., 2018, A&A, 609, A83
- Joudaki S. et al., 2017, MNRAS, 471, 1259
- Joudaki S. et al., 2020, A&A, 638, L1
- Kaiser N., Tonry J. L., Luppino G. A., 2000, PASP, 112, 768
- Laigle C. et al., 2016, ApJS, 224, 24
- Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258
- Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
- LSST Dark Energy Science Collaboration, 2012, preprint (arXiv:1211.0310)
- Masters D.C. et al., 2015, ApJ, 813, 53
- Masters D. C. et al., 2019, ApJ, 877, 81
- Miyazaki S. et al., 2012, Proc. SPIE, 8446, 84460Z
- Planck Collaboration VI, 2018, preprint (arXiv:1807.06209)
- Samuroff S., Troxel M. A., Bridle S. L., Zuntz J., MacCrann N., Krause E., Eifler T., Kirk D., 2017, MNRAS, 465, L20
- Sánchez C., Bernstein G. M., 2019, MNRAS, 483, 2801
- Scolnic D. M. et al., 2018, preprint (arXiv:1812.00516)
- Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
- Speagle J. S., Eisenstein D. J., 2017, MNRAS, 469, 1186
- Speagle J. S. et al., 2019, MNRAS, 490, 5658
- Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
- Troxel M. A. et al., 2018a, Phys. Rev. D, 98, 043528
- Troxel M. A. et al., 2018b, MNRAS, 479, 4998
- Van Waerbeke L., White M., Hoekstra H., Heymans C., 2006, Astropart. Phys., 26, 91
- Wright A. H., Hildebrandt H., van den Busch J. L., Heymans C., 2020, A&A, 637, A100
- York D. G. et al., 2000, AJ, 120, 1579

APPENDIX A: COSMOS2015 SOM

In this appendix, we describe the SOM created with COSMOS2015 data, which is used in Section 7 to define the phenotypes of that sample. The use of SOMs to discretize the galaxy colour/magnitude space has been discussed extensively in previous works (e.g. Masters et al. 2015; Buchs et al. 2019; Speagle et al. 2019), and we refer

the reader to those works and references therein for more details about the standard SOM algorithm. Here, we will describe the ways in which the algorithm we apply to COSMOS2015 differs from previous implementations.

The gist of the algorithm is as follows: first, we have a population of objects (galaxies) labelled by $i = 1, 2, ..., N_{obj}$, each of which has a measured feature set F_i in a d_F -dimensional space consisting of, in our cases, the fluxes of the object in the u, B, V, r, ip, zp, J, H, and Ks bands. In our case we also have measurement uncertainties σ_i corresponding to each F_i – the standard SOM algorithm does not allow for measurement errors in the training set, so this is one aspect we wish to address.

The SOM is defined by discrete set of points \mathbf{x}_c in an d_{SOM} dimensional space. Typically these are placed in a regular grid in $d_{\text{SOM}} = 2$ dimensions for photo-*z* applications. Each such 'cell' is assigned a location in the feature space, $F_c = F(\mathbf{x}_c)$, known as the 'weights' for the cell in the SOM literature. The final step in specifying the SOM is to define some distance measure $r(F_c, F_i, \sigma_i)$ between a cell and an object. Typically this is taken as the Euclidean metric $|F_c - F_i|$, but this is not required. There is in fact no need for *r* to satisfy the conditions of a metric, nor even to be symmetric in its arguments. Below we will describe a distance function which is much more desirable in the photo-*z* application than the Euclidean distance.

An arbitrary object *i* is assigned to an SOM cell by simply selecting the cell *c* which minimizes $r(F_c, F_i, \sigma_i)$. The SOM literature calls *c* the 'best matching unit' (BMU) for the object.

The algorithm for assigning the weights F_c to the cells using a sequence of training objects F_j , $j = 1, 2, ..., N_{\text{train}}$ from the population is

(i) initialize the weights to a set $F_c^{(0)}$ in the regime spanned by the data;

(ii) for each training point *j*, find its BMU, then alter the full set of weights according to

$$\boldsymbol{F}_{c}^{(j)} = \boldsymbol{F}_{c}^{(j-1)} + A(j)H(\boldsymbol{x}_{c} - \boldsymbol{x}_{\text{BMU}}, j) \boldsymbol{\Delta} \left(\boldsymbol{F}_{c}^{(j-1)}, \boldsymbol{F}_{i}, \boldsymbol{\sigma}_{j} \right).$$
(A1)

In the training phase, we have a shift in the feature space, Δ , to draw the BMU weights towards the training object. In the standard algorithm Δ is simply $F_i - F_c^{(j-1)}$, but we will alter this. The function *H* describes a neighbourhood around the BMU in the SOM space which is dragged towards the training point, which contracts with advancing iterations *j*. Finally, there is an overall learning rate *A* which also typically decreases as training proceeds.

We adopt the functional forms for *A* and *H* given by Speagle et al. (2019). We adopt a regular grid of **x** in two dimensions for our cells, as have other photo-*z* applications, in our case a 64×64 array. We do *not* adopt period boundary conditions in SOM space, because there is no natural periodicity in the feature space of galaxy colours and magnitudes.

The main alterations we make to previous methods are to the distance function r and the training shift Δ . Previous works have struggled with the choice between using fluxes, magnitudes, and/or colours as the elements of the feature space. Fluxes are the natural space in the sense that measurement errors are nearly Gaussian in flux space, making a Euclidean distance the best match to a (log) probability in flux space. It is also true, however, that galaxy *colours* – i.e. ratios of fluxes, or differences in log(flux) – are more sensitive indicators of redshift than are fluxes, making a Euclidean metric in colour a better way to group galaxies into cells of common redshift. For this reason, most SOM-based photo-*z* techniques use colours as features. But this method fails catastrophically when one



Figure A1. The COSMOS2015 SOM, coloured by the mean redshift in each cell. White cells have no members.

or more bands have modest or low S/N (or when observed fluxes are negative!), as colour becomes wildly uncertain, and the measurement noise can come to dominate the choice of BMU. We want our functions *r* and Δ to respect the meaning of the error bars, and not use unreliable information to classify objects or train the SOM. A further shortcoming of pure colour-based SOMs is that redshift is known to very with flux (or magnitude) at fixed colour, as demonstrated by Speagle et al. (2019) and Masters et al. (2019). We therefore want a distance function with sensitivity to overall flux or magnitude level, but much more weakly than to colour.

The solution we find is as follows. First, we use fluxes as our features, so that negative measured fluxes can be treated properly. But we force the cells to have positive weights (fluxes), since we consider the cell weights to be noiseless galaxy 'templates'. We then define a nominal distance function to be a sum over the elements of the flux vector (indexed by band b) as follows. First, we convert the object and cell fluxes into units of signal-to-noise ratio (SNR), specifying a maximum for the object SNR as a means of softening the specificity of the cells:

$$s_{ib} \equiv \max\left(\sigma_{ib}, F_{ib}/\text{SNR}_{\text{max}}\right),$$
 (A2)

$$\nu_{ib} \equiv F_{ib}/s_{ib},\tag{A3}$$

$$\nu_{cb} \equiv F_{cb}/s_{ib}.\tag{A4}$$

Next we define a weighting function that will be used to transition from the high- to low-SNR regimes:

$$w_{ib} = e^{2(v_{ib} - 4)}. (A5)$$

Now we define a distance function

$$\tilde{r}(\boldsymbol{F}_{c}, \boldsymbol{F}_{i}, \boldsymbol{\sigma}_{i}) = \sum_{b} \left[\frac{\operatorname{asinh} v_{cb} + w_{ib} \log 2v_{cb}}{1 + w_{ib}} - \operatorname{asinh} v_{ib} \right]^{2} \times \left(1 + v_{ib}^{2} \right),$$
(A6)

which has the desirable properties of approaching Euclidean in logflux at high SNR, and Euclidean in linear flux at low SNR, weighting each band by its SNR (up to a maximum), and monotonically increasing away from $F_i = F_c$.

The final step is to make the cells 'fuzzy' in overall flux level, in the sense of Speagle & Eisenstein (2017), by allowing the cell fluxes to be scaled by an overall constant e^s :

$$r(\boldsymbol{F}_{c}, \boldsymbol{F}_{i}, \boldsymbol{\sigma}_{i}) = \inf_{s} \left[\tilde{r}(e^{s}\boldsymbol{F}_{c}, \boldsymbol{F}_{i}, \boldsymbol{\sigma}_{i}) + \frac{s^{2}}{\sigma_{s}^{2}} \right].$$
(A7)

The cell's fluxes are thus rescaled to find the minimal distance to the object's fluxes, subject to a penalty that is quadratic in the log of the scaling factor. The parameter σ_s determines, essentially, the width in magnitudes of the smeared cells. We thus realize a distance that can be quite sharp in any colour while being broad in overall flux.

We will not detail the shift function Δ here, just noting that it operates as a simple shift in log-flux space when SNR is high in a given band, but produces no shift during training in bands where the difference between the flux of the training object and the flux of the cell have less than 2σ significance.

The COSMOS2015 SOM is created using SNR_{max} = 50, which essentially sets the sensitivity to colour at ≈ 0.02 mag; and with $\sigma_s =$ 0.4, which sets the sensitive to overall flux at ≈ 0.4 mag. Fig. A1 shows the mean redshift of COSMOS2015 galaxies assigned to each SOM cell, plotted across the SOM space **x**. The redshift standard deviation of galaxies assigned to each cell has a median of 0.21. Keep in mind that the SOM was constructed without regard to redshift. The smooth behaviour of redshift across the SOM shows that the variation of redshift in the 9-band flux space is reasonably well traced by the embedded 2D manifold defined by the SOM.

This paper has been typeset from a TEX/LATEX file prepared by the author.