Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **498,** 2614–2631 (2020) doi:10.1093/mnras/staa2478
Advance Access publication 2020 August 17

Redshift inference from the combination of galaxy colours and clustering in a hierarchical Bayesian model – Application to realistic *N*-body simulations

Alex Alarcon [1,2]★ Carles Sánchez,[3]★ Gary M. Bernstein [3] and Enrique Gaztañaga [1,2]

[1]*Institut d'Estudis Espacials de Catalunya (IEEC), E-08193 Barcelona, Spain*
[2]*Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, E-08193 Barcelona, Spain*
[3]*Department of Physics and Astronomy, University of Pennsylvania, 209 S. 33rd St., Philadelphia, PA 19104, USA*

Accepted 2020 August 10. Received 2020 June 16; in original form 2019 October 22

ABSTRACT

Photometric galaxy surveys constitute a powerful cosmological probe but rely on the accurate characterization of their redshift distributions using only broad-band imaging, and can be very sensitive to incomplete or biased priors used for redshift calibration. A hierarchical Bayesian model has recently been developed to estimate those from the robust combination of prior information, photometry of single galaxies, and the information contained in the galaxy clustering against a well-characterized tracer population. In this work, we extend the method so that it can be applied to real data, developing some necessary new extensions to it, especially in the treatment of galaxy clustering information, and we test it on realistic simulations. After marginalizing over the mapping between the clustering estimator and the actual density distribution of the sample galaxies, and using prior information from a small patch of the survey, we find the incorporation of clustering information with photo-z's tightens the redshift posteriors and overcomes biases in the prior that mimic those happening in spectroscopic samples. The method presented here uses all the information at hand to reduce prior biases and incompleteness. Even in cases where we artificially bias the spectroscopic sample to induce a shift in mean redshift of $\Delta\bar{z} \approx 0.05$, the final biases in the posterior are $\Delta\bar{z} \lesssim 0.003$. This robustness to flaws in the redshift prior or training samples would constitute a milestone for the control of redshift systematic uncertainties in future weak lensing analyses.

Key words: cosmology: observations – dark energy – large-scale structure of Universe.

1 INTRODUCTION

Galaxy surveys provide key information about the large-scale structure of the Universe, constituting one of the most powerful probes for testing cosmological models. There exist two main categories of surveys. On one hand, spectroscopic surveys such as 2dF (Colless et al. 2001), the VIMOS-VLT Deep Survey (Le Fèvre et al. 2005), WiggleZ (Drinkwater et al. 2010), Baryon Oscillation Spectroscopic Survey (Dawson et al. 2013), and Dark Energy Spectroscopic Instrument (DESI Collaboration 2016) provide 3D information about the galaxies they measure, but they are expensive in time and resources. On the other hand, imaging or photometric surveys like the Sloan Digital Sky Survey (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 Large Synoptic Survey Telescope (LSST; LSST Dark Energy Science Collaboration 2012; Ivezić et al. 2019) use less time per galaxy, and enable weak gravitational lensing measurements via galaxy shapes – but provide only a crude view of the line-of-sight dimension of the Universe, since galaxy redshifts are estimated from only their observed broad-band fluxes.

In order to perform unbiased cosmological analyses of imaging surveys it is critical to characterize the redshift distributions $n(z) = \mathrm{d}N/\mathrm{d}z\,\mathrm{d}A$ of the corresponding galaxy samples, and unaccounted errors in such characterization will directly lead to biases in the cosmological parameter estimation (Huterer et al. 2006; Cunha et al. 2012; Hildebrandt et al. 2012; Benjamin et al. 2013; Huterer, Cunha & Fang 2013; Bonnett et al. 2016; Hildebrandt et al. 2017; Joudaki et al. 2017a; Hoyle et al. 2018). Recently, there has been a number of comparisons between cosmological parameters obtained from imaging suveys (Hildebrandt et al. 2018; Troxel et al. 2018a; Hikage et al. 2019) and the cosmic microwave background (Planck Collaboration VI 2018) which have claimed discrepancies of up to 3.2σ in their results (Asgari et al. 2019). Even though such discrepancies could be attributed to a failure of the ΛCDM model (Joudaki et al. 2017b), such a claim would need significant evidence and thorough testing. 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. 2020). Moreover, such studies indicate that a major difference in analysis methodologies lies in the redshift calibration, and that this can produce such discrepancy. Redshift calibration clearly needs substantial improvement for the current- and next-generation photometric surveys.

Several techniques for estimating the redshift distributions of imaging surveys have been developed in the last decades, which can be broadly separated in three categories.

★ E-mail: aalarcongonzalez@anl.gov (AA); carless@sas.upenn.edu (CS)

(i) Direct spectroscopic measurement of redshifts is an obvious tactic. Since spectroscopic redshifts are expensive, this method can presently be applied to only a subset of the photometric galaxy population. The resulting shot noise implies that $O(10^5)$ unbiased spectra must be obtained (Ma & Bernstein 2008) to reach the needed precision in $n(z)$ for current (Stage III), potentially lowered to $O(10^4)$ in future (Stage IV) surveys by careful targeting to span the galaxy population (Masters et al. 2015). Such 'direct' calibration is also subject to large-scale-structure (LSS) variance if (as is most practical) it is conducted over a small sky area. Even if numerical requirements are met, the direct method is subject to redshift biases because of differential rates of success in obtaining a reliable redshift across the redshift and magnitude range of the target population (Speagle et al. 2019). In this paper we will refer to information obtained through direct spectroscopy (or many-band photometry) as the 'spectroscopic prior.' It is essential that any $n(z)$ estimation be robust to the noise and biases that exist in real-life spectroscopic surveys.

(ii) Photometric redshifts compare a set of observed fluxes (or colours or potentially other measurable features) F_i of source i to those expected for galaxies at various redshifts to infer the redshift of the individual target galaxy. The map $z(F)$ is based on some mix of theoretical models of galaxy spectra with empirical knowledge from direct spectroscopy. The inference $p(z|F)$ can be made using explicit template-fitting methods (e.g. `Hyperz`, Bolzonella, Miralles & Pell 2000; `BPZ`, Benitez 2000; Coe et al. 2006; `LePhare`, Arnouts et al. 2002; Ilbert et al. 2006; `EAZY`, Brammer, van Dokkum & Coppi 2008), or machine-learning 'training' methods (e.g. `ANNz`, Collister & Lahav 2004; `ArborZ`, Gerdes et al. 2010; `TPZ`, Carrasco Kind & Brunner 2013; `SkyNet`, Bonnett 2015). The most basic, completely empirical, form of photometric redshift determination is to assign to each target the redshift of its nearest neighbour (by some metric in colour/mag space) among a subset with spectroscopic redshifts. This reweighting of the spectroscopy by imaging data was proposed by Lima et al. (2008), and multiple current implementations of it attain various levels of rigor in the treatment of observational errors. Comparisons of different photometric methods have been performed in simulated and real data (Hildebrandt et al. 2010; Dahlen et al. 2013; Sánchez et al. 2014). The limitations of photometric methods are that the map $z(F)$ can be ambiguous even with noiseless data, therefore requiring that the correct $p(z|F)$ incorporate accurate priors on the relative abundance $n(z, F)$. And of course the photometric method inherits any biases or deficiencies of its theoretical/empirical training basis.

(iii) Clustering redshifts use data coming from large-area surveys to constrain $n(z)$, i.e. using the observed sky positions θ_i of the sources in comparison to the positions of a tracer population with secure redshifts. The simple principle is that the targets' θ_i will show no correlation with the tracers unless they are physically colocated, i.e. at a common redshift. The tracers need not be a representative sampling of the targets. The weaknesses of this method are that it will, of course, only provide information at redshifts where abundant tracers are known; that the information per target is weak; that the inference of $n(z)$ is degenerate with a redshift dependence of the 'bias', i.e. the relation between tracer space density and target space density; and that lensing magnification can introduce additional signal unrelated to the physical overlap between targets and tracers. As a consequence, the accuracy of the clustering method is enormously improved if photometric information can be used to select subpopulations known to have narrow redshift range. The application of this method is typically based on 2-point statistics between the source population and tracer population. (Newman 2008; Ménard et al. 2013; Schmidt et al. 2013).

Some recent analyses have attempted the combination of photometric and clustering constraints on the same survey data in the presence of prior spectroscopic information (Davis et al. 2017; DES Collaboration 2017; Hildebrandt et al. 2017; Gatti et al. 2018; Hoyle et al. 2018), but the comparisons have been performed just by means of basic visual cross-checks on the two independently derived $n(z)$'s, or using some single summary statistic of $n(z)$, such as its mean.

Sánchez & Bernstein (2019), hereafter SB19, present a framework to combine these three pieces of information (prior, photometry, and clustering) in a principled way to assign a posterior probability to $n(z)$ using a hierarchical Bayesian model (see also Leistedt, Mortlock & Peiris 2016). The framework provides posterior samples of the redshift distribution of a population constrained by all sources of information, and SB19 demonstrated its performance on simple, idealized simulations. In this work, we extend the method so that it can be applied to real data from galaxy surveys, with the main addition being a practical, realistic treatment of the galaxy clustering information. Using the public MICE2 N-body simulations, we define a galaxy subsample as tracers with known redshifts to develop a clustering probability based on a kernel density estimator (hereafter KDE). We incorporate a redshift-dependent biasing function that maps the local tracer KDE output to the actual density distribution of the target galaxies. The method includes marginalization over the biasing functions' parameters, since in data there will be no sufficiently accurate prior on the biasing relations.

The methods developed in this work provide the necessary tools for the application of the framework to real data. The simulation that is used, even though it is not intended to completely mimic the real data, has all of the parameters relevant to the $n(z)$ accuracy in a realistic range, i.e. galaxy and tracer density, clustering amplitudes and power spectra, noise levels, and sizes of spectroscopic, wide, and deep samples used in the application of the scheme to the DES. Therefore, the methods and the results presented in this paper demonstrate the capabilities of the framework in a realistic setting.

This paper is organized as follows. In Section 2 we present the details of the methodology and the phenotype approach. Section 3 describes the simulated galaxy catalogue used to test the methodology. We follow with a description of the density estimation used to incorporate the clustering information in Section 4. We describe the Gibbs sampling technique used to sample the posterior on all the model parameters in Section 5. Section 6 shows the main results in this work, with priors coming from precise redshifts over a small patch of sky. We examine three cases in which these spectroscopic priors are biased, inspired by shortcomings in real data. Section 7 presents a discussion of the methodology and its application to real data, and we summarize and conclude in Section 8.

2 FRAMEWORK

We work under the framework presented in SB19, in which galaxy 'types' are defined by observed properties rather than rest-frame properties, and we call them phenotypes. The individual galaxies i are seen as being drawn from a pool of possible phenotypes t_i, redshifts z_i, and angular positions θ_i, with intrinsic mean density $n(t, z)$ on the sky. The t_i and z_i and $n(t, z)$ are noiseless latent variables, with the observations yielding a catalogue with the θ_i and a noisy set of observable features which will be denoted as F_i – namely apparent magnitudes/colours. The clustering information is included by considering that the sky density of galaxies of type t at redshift z is modulated by some factor $1 + \delta_z^t(\theta)$. In this paper, we will simplify the galaxy density field to be type-independent, $\delta_z^t \to \delta_z$. The latent densities $\delta_z^t(\theta)$ will be constrained using a 'tracer' galaxy

Table 1. Summary of the notation used throughout this paper.

F	Galaxy set of observed features
t	Galaxy phenotype (or simply type)
z	Galaxy redshift
θ	Galaxy angular position on the sky
s	Indicator of successful detection/selection
$\mathcal{L}_{it}$	Probability of measuring galaxy i with F_i given t
$\boldsymbol{F}, \boldsymbol{t}, \boldsymbol{z}, \boldsymbol{\theta}$	Set of properties for all galaxies in the sample
N	Number of galaxies in the sample
N_t	Number of types
N_z	Number of redshifts
A	Effective area of the survey for source detection
n	Mean galaxy density per unit solid angle
$n(z)$	Mean galaxy density per unit solid angle per z
$\delta_z(\theta)$	Target density fluctuation at a given z and θ
$\tilde{\delta}_z(\theta)$	Tracer density fluctuation at a given z and θ
$\hat{\delta}_z(\theta)$	Tracer density estimator at a given z and θ
π_δ	Density fluctuation field hyperparameters
$\boldsymbol{\delta}$	Set of $\delta_z(\theta)$ for all redshifts and positions
$\mathcal{B}$	Mapping relation between estimated density field and true clustering probability
b_z^t	Parameters of the $\mathcal{B}$ function for type t at redshift z
$\boldsymbol{b}$	Set of b_z^t for all types and redshifts
f_{zt}	Joint type and redshift probability $p(z, t)$
$\boldsymbol{f}$	Set of f_{zt} for all types and redshifts
N_{zt}	Number of sources assigned to redshift z and type t
$\boldsymbol{N}$	Set of N_{zt} for all redshifts and types
M_{zt}	Number of sources in the prior at redshift z and type t
$\boldsymbol{M}$	Set of M_{zt} for all redshifts and types
Δz	Difference between the means of estimated and true $n(z)$'s
D_{KL}	Kullback–Leibler divergence between estimated and true $n(z)$'s

population known to be at redshift z. Our notation will be that the vector quantities $\boldsymbol{F}$, $\boldsymbol{t}$, $\boldsymbol{z}$, and $\boldsymbol{\theta}$ denote the full set of properties of all selected galaxies, i.e. $\boldsymbol{F} = \{F_1, F_2, \ldots, F_N\}$ (a summary of all the notation can be found in Table 1).

2.1 Generative model

As in SB19, the fundamental assumption of the method is that galaxies are drawn from a Cox process (Cox 1955) or doubly stochastic Poisson process, i.e. we assume that each galaxy is Poisson sampled from a latent, stochastic density field. The problem simplifies when considering the redshift z as an integer indexing a set of finite-width redshift bins, where each bin has an independent density fluctuation field $\delta_z(\theta)$, i.e. $\langle \delta_{z_i}(\theta)\delta_{z_j}(\theta)\rangle = 0$ for $z_i \neq z_j$. We will also assume that we have a finite set of phenotypes indexed by integer t. Each phenotype has a mean sky density of $n^t = nf_t$, where we place n as the total density of all detectable galaxy phenotypes, and $f_t = p(t)$ being the fraction of the population in each type, with $\sum_t f_t = 1$ as a constraint. Then the redshift distribution of type t will be $p(z|t) = f_z^t$, and we will also denote

$$f_{tz} \equiv p(z, t) = p(z|t)p(t) = f_z^t f_t. \tag{1}$$

We are considering the sky to be populated with galaxies with a finite variety of redshifts and phenotypes, where phenotypes specify a galaxy's noiseless, observer-frame appearance. We assume there is some selection function s with the probability of a galaxy being selected, possibly depending on sky position, specified as a selection function $p(s|t, \theta)$. We will always assume that we know nothing about the non-selected galaxies, not even that they exist; the observed data are the positions $\boldsymbol{\theta}$ and features $\boldsymbol{F}$ of the selected galaxies. All galaxies of phenotype t observed under the same conditions are assumed to have the same selection function $p(s|t, \theta)$ and the same probability $p(F, s|t, \theta)$ of being selected and measured to have image features F. Finally, we will allow for some local biasing function, $\mathcal{B}_z^t$, with parameters $\boldsymbol{b}_z^t$, depending on both redshift and phenotype, to relate the galaxies' spatial distribution to the underlying tracer density fluctuation $\tilde{\delta}_z$. Now the selected galaxies can be considered as being a Poisson sampling of the following density field:

$$\rho(z, \theta, t|n, \boldsymbol{f}, \boldsymbol{b}, \tilde{\boldsymbol{\delta}}) = nf_{tz}\mathcal{B}_z^t\left(\tilde{\delta}_z(\theta), \boldsymbol{b}_z^t\right) p(s|t, \theta). \tag{2}$$

The $\mathcal{B}$ term describes the spatial variation of the expected detection rate due to density fluctuations. The last term describes density fluctuations due to variable observing conditions. In this work, we will consider the bias function to be independent of type, so $\mathcal{B}_z^t \to \mathcal{B}_z$, and the biasing parameters likewise are independent of t.

With knowledge of the survey noise properties and the noiseless appearance of phenotype t, we can determine the likelihood $p(F, s|t, \theta, z)$ of a galaxy of phenotype t at location θ, z being selected and measured with features F. Note that this likelihood will not depend on z since the phenotype's observables are independent of z, by construction. Therefore, for each observed galaxy i and phenotype t, we can assign a feature/selection likelihood

$$\mathcal{L}_{it} \equiv p(F_i, s|t_i, \theta_i). \tag{3}$$

This function will depend on the quality of the observations at sky position θ_i and the measurement and selection algorithms. We will assume that this likelihood is known a priori, e.g. by the result of analysing the injection of artificial copies of the phenotype into the real survey images (Suchyta et al. 2016).

Then the probability of selecting a set of galaxies $i \in \{1 \ldots N\}$ at positions $\boldsymbol{\theta}$ with features $\boldsymbol{F}$, types $\boldsymbol{t}$, and redshifts $\boldsymbol{z}$ takes the standard Poisson form:

$$p\left(\boldsymbol{F}, \boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{z}|n, \boldsymbol{f}, \boldsymbol{b}, \tilde{\boldsymbol{\delta}}\right) = \exp\left[-n\sum_t f_t A^t(\boldsymbol{f}^t, \boldsymbol{b}^t, \tilde{\boldsymbol{\delta}})\right] \times \prod_i \mathcal{L}_{it} n f_{t_i z_i} \mathcal{B}_{z_i}^{t_i}\left(\tilde{\delta}_{z_i}(\theta_i), \boldsymbol{b}_{z_i}\right). \tag{4}$$

The exponentiated quantity is, as required for Poisson distributions, the expected number of detections $\langle N\rangle$ for the entire sample. This can be determined from knowledge of the survey properties:

$$\begin{aligned} A^t(\boldsymbol{f}, \boldsymbol{b}, \tilde{\boldsymbol{\delta}}) &\equiv \sum_z \int \mathrm{d}^2\theta\; p(s|t, \theta) f_z^t \mathcal{B}_z^t\left(\tilde{\delta}_z(\theta), \boldsymbol{b}_z\right) \\ &= \int \mathrm{d}^2\theta\; p(s|t, \theta) \sum_z f_z^t \mathcal{B}_z^t\left(\tilde{\delta}_z(\theta), \boldsymbol{b}_z\right) \\ &\approx \int \mathrm{d}^2\theta\; p(s|t, \theta), \end{aligned} \tag{5}$$

where we have assumed that the clustering information integrated over the mask of the survey approximately keeps its average value of unity, $\int \mathrm{d}^2\theta\; p(s|t, \theta)\mathcal{B}_z^t\left(\tilde{\delta}_z(\theta), \boldsymbol{b}_z\right) \approx 1$.

In order to provide the full generative model for the data, we must also specify the process $p(\tilde{\boldsymbol{\delta}}|\pi_\delta)$ generating the stochastic density fluctuation fields given some hyperparameters π_δ. For instance, that could be a lognormal distribution where π_δ specifies the power spectrum. We also require priors $p(\boldsymbol{b})$ and $p(n)$, plus any prior information on $p(\boldsymbol{f})$ aside from the constraint that $\sum_{tz} f_{tz} = 1$.

2.2 Redshift inference

The principal quantity of interest is the underlying redshift distribution

$$n(z) = n \sum_t f_{tz}. \tag{6}$$

In most applications of redshift inference, we are only concerned with the shape, not the normalization, of $n(z)$, and therefore we will focus here on the fractions $\boldsymbol{f}$, rather than n. In addition, in many applications it is also useful to know the individual redshifts of galaxies $\boldsymbol{z}$, and in order to enable a Gibbs sampling scheme, which is the simplest way of sampling our posterior (see Section 5), we will need to keep $\boldsymbol{b}$ and $\boldsymbol{t}$ as conditional variables. We can use Bayes' theorem to write down the posterior joint probability of these variables of interest:

$$\begin{aligned} p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{t}, \boldsymbol{b}|\boldsymbol{F}, \boldsymbol{\theta}, \pi_\delta) \propto \int \mathrm{d}n\, \mathrm{d}\tilde{\boldsymbol{\delta}} \\ p\left(\boldsymbol{F}, \boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{z}|n, \boldsymbol{f}, \boldsymbol{b}, \tilde{\boldsymbol{\delta}}\right) \\ p(\tilde{\boldsymbol{\delta}}|\pi_\delta)\, p(n)\, p(\boldsymbol{f})\, p(\boldsymbol{b}). \end{aligned} \tag{7}$$

We have already derived the first term under the integral in equation (4). In this paper, as in SB19, we will work with the approximation that we can replace the stochastic tracer density fluctuation $\tilde{\delta}_z(\theta)$ with some deterministic estimator $\hat{\delta}_z(\theta)$ of the realization of the density fields in the generative probability of equation (4). Under that approximation we can ignore the hyperparameters generating the density field π_δ but we lose the ability to use the information available from the clustering of the target galaxies. Performing the marginalization over n assuming the effective area of the survey is independent of phenotype (see SB19 for more details), the posterior distribution for redshift and phenotype information in equation (7) becomes

$$p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{t}, \boldsymbol{b}|\boldsymbol{F}, \boldsymbol{\theta}) \propto p(\boldsymbol{f}) p(\boldsymbol{b}) \prod_i \mathcal{L}_{it_i}\, f_{t_i z_i}\, \mathcal{B}_{z_i}\left(\hat{\delta}_{iz_i}, \boldsymbol{b}_{z_i}\right), \tag{8}$$

$$\hat{\delta}_{iz} \equiv \hat{\delta}_z(\theta_i). \tag{9}$$

The roles of the main three sources of information in redshift estimation are clearly present and differentiated in the posterior of equation (8). First, there is a term for the prior probability that any galaxy is of phenotype t and redshift z, f_{tz}. Secondly, the photometric information for a galaxy is in $\mathcal{L}_{it}$, which is the likelihood of galaxy i resembling phenotype t and passing selection. Thirdly, clustering information enters as the last term, describing the modification of the probability by our estimator for the density fluctuation field.

In more detail: the prior term can be estimated using a subset of galaxies with well-characterized phenotypes and redshifts, which we will call the spectroscopic sample. It requires deep (low-noise) photometric data, plus either spectroscopic or high-quality photometric redshifts, of a fair subsample of the sources. The clustering information will require another galaxy subpopulation, the tracers, having well-characterized redshift information and spanning a large area and redshift range of the survey (but no need to span them completely). This can be a population of galaxies with accurate photometric redshift estimates, like e.g. luminous red galaxies (LRGs). We will refer to all galaxies in the sample of interest as target galaxies, for which we will only have the measurements of $\boldsymbol{F}$and $\boldsymbol{\theta}$.

2.3 Realistic set up: SOM implementation

To discretize the phenotypes for a general imaging survey, we propose to use a combination of wide and deep survey observations and self-organizing maps (SOMs; Masters et al. 2015). Deep observations are often available for surveys like the DES by summing observations of fields being monitored for high-z SNe. These provide essentially noiseless photometric measurements and observations in additional filter bands for galaxies in specific fields (henceforth deep fields, or simply DFs). The DFs provide an empirical sampling of the distribution of galaxies in feature ($\boldsymbol{F}$) space. In turn, SOMs provide a data-driven way of mapping and discretizing that feature space, so that each cell c of the so-called deep SOM cell constitutes a phenotype t.

Another term that we will need in the data application is the noise or measurement likelihood, $\mathcal{L}_{it} \equiv p(F_i, s|t_i, \theta_i)$. We follow the approach of Buchs et al. (2019) and construct the measurement likelihood by training another SOM on wide-field data of the galaxy survey of interest; we will refer to this one as the wide SOM and its cells, $\hat{c}$, span the space of features $\boldsymbol{F}$ observed in the wide-field survey (i.e. every detected galaxy will be assigned to one wide cell, $\hat{c}$). Crucially, it is possible to inject artificial copies of galaxies with deep photometry, and hence well specified phenotypes, into the real images of the survey, and measure their (noisy) wide-field properties (Suchyta et al. 2016). Then, for a set of injected galaxies, we will know both the cells in the deep and wide SOMs ($\hat{c}$ and c), so that we can construct the mapping between deep and wide SOMs which corresponds to our measurement likelihood:

$$\mathcal{L}_{it} \equiv p(F_i, s|t_i, \theta_i) \equiv p(\hat{c}_i, s|c_i, \theta_i). \tag{10}$$

One other major part in the application of the method to data is the addition of clustering information, that is, the construction of the density field estimator using a tracer population and the creation of biasing functions $\mathcal{B}_z^t$ relating that estimate to the true underlying density fluctuation field of the selected galaxies. This will be treated in Section 4.

3 SIMULATIONS

SB19 demonstrated the performance of the hierarchical Bayesian model (HBM) for redshift estimation described in the previous section in a simplified simulation with idealized galaxy properties and noise distributions, and perfect knowledge of the density fluctuation field. Now, instead, we test our methodology on the public MICE2 simulation,[1] a mock galaxy catalogue created from a light-cone of a dark-matter-only N-body simulation that contains ∼200 million galaxies over one sky octant (∼5000 deg^2) and up to $z = 1.4$. Several important differences with respect to the SB19 simulation make this analysis more realistic and allow the method described herein to be applicable to analysis of real data.

First, the MICE2 simulation has realistic clustering properties given by a ΛCDM cosmology with parameters $\Omega_m = 0.25$, $\Omega_b = 0.044$, $h = 0.7$, $n_s = 0.95$, $\Omega_\Lambda = 0.75$, $\sigma_8 = 0.8$, and $w = -1$. In addition, we do not assume true knowledge of the density field but rather infer the clustering information from a set of galaxy tracers, described below. Secondly, galaxies 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. (2015a,b), Crocce et al. (2015).

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

3.1 Target and tracer sample selection

We select a galaxy sample within a square footprint defined by the cuts $30 \leq \mathrm{RA[deg]} \leq 60$ and $0 \leq \mathrm{Dec[deg]} \leq 35$, representing an area of around $1000\,\mathrm{deg}^2$, with the redshift range $0.2 \leq z \leq 1.2$, and we place a magnitude cut at $i_{\mathrm{DES}} < 24$, where i_{DES} represents the i photometric filter in DES (Flaugher et al. 2015). We use both positions and fluxes without magnification, and we leave a thorough study of magnification effects for future work. To reduce runtimes, we cull the galaxy catalogue by a factor $\sim$2 by selecting only those galaxies with a subset of SEDs. This downsampling retains a representative sampling of populations (Elliptical, Spiral, Starburst) and dust attenuation laws and values present in the simulation.[2]

The tracer sample is a subsample of the full population, randomly drawn to maintain a constant comoving density similar to that of the REDMAGIC DES Y1 galaxy sample in its first three lens bins (Elvin-Poole et al. 2018). This choice is arbitrary, and perhaps unrealistic at the higher end of our redshift range, but it is not a necessary feature of the method. The target sample is defined as the galaxies that are not selected as tracers. The upper panel of Fig. 1 shows the redshift distributions of both samples. Tracers have a density between 0.015 (at $z = 0.2$) and 0.5 (at $z = 1.2$) times the target density. The redshift binning is chosen to have 20 bins equally spaced in comoving distance χ between the redshift limits of the catalogue, which makes the tracer sample have a constant density per bin per unit comoving surface area $\mathrm{d}A$, $\mathrm{d}N \propto \mathrm{d}A\mathrm{d}\chi \propto \mathrm{d}A$.

It is worth highlighting here some differences between this simulation and a corresponding real data sample, in particular DES. First, the simulation sample used in this work contains about 1/5 of the total area in DES. This is relevant as we expect the clustering information to grow more powerful as area grows, so the simulation is a conservative estimate of the value of clustering, in that sense. Secondly, the galaxy tracers used in adding clustering information in this work are unbiased with respect to the total sample. A real data application is likely to use luminous red galaxies (LRGs) or other highly biased population as tracers. We have not, however, assumed in this simulation that tracers are unbiased, but we have instead marginalized over a biasing relation. The lower tracer bias (relative to mass) in our simulation may be considered a conservative scenario, in the sense that it will increase the impact of shot noise in the density estimates compared to an LRG tracer sample. However, biased tracer populations may also introduce more complicated evolution effects in the biasing relation which may result in an increase in the number of free parameters necessary to describe it, affecting how much clustering information we can extract. The usage of biased tracers for galaxy clustering is left for future work. Finally, we have used a limited redshift range in this work, $0.2 < z < 1.2$, and we have used tracers spanning this entire redshift range. In the application to real data, a more complete redshift range will have to be considered, and tracers may be available just for a limited redshift range, but that can be accommodated naturally in the method and was shown to work as expected in SB19.

3.2 The phenotype approach: Deep and wide SOMs

The phenotype method described in Section 2 is then applied to the simulation. As stated in Section 2.2, the approach can benefit from a deep sample with deeper photometry and extra observed wavelength bands than the target (or 'wide') sample, which helps

[2]The selection is defined as sed_cos $\equiv c \in$ 0, 1, 2, 5, 6, 7, 10, 11, 12, 15, 16, 17, 21, 22, 23, 24, 25, 29, 30, 35, 36, 37, 38, 39, 41, 42, 43.

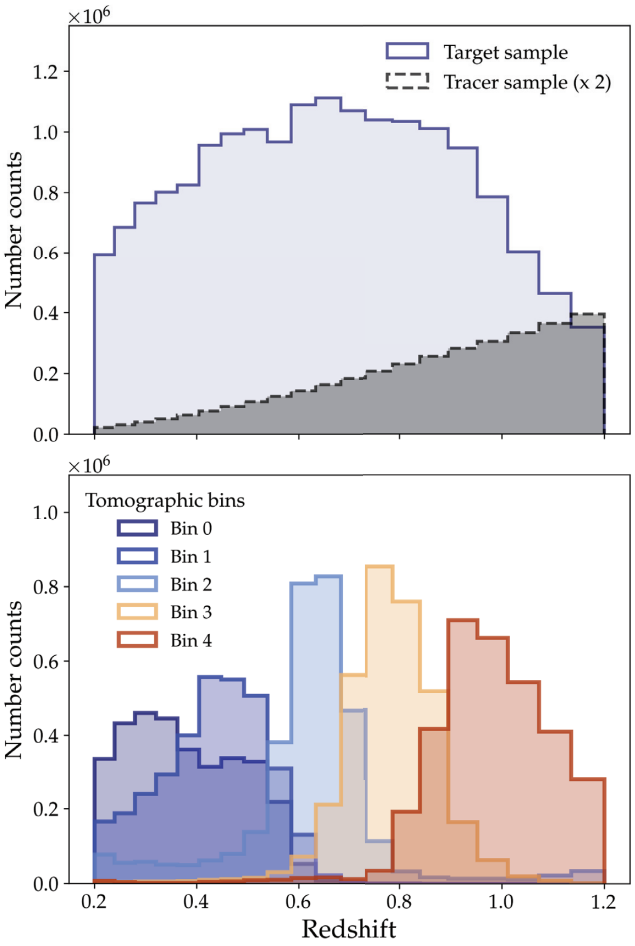

Figure 1. Upper panel: Redshift distributions of the target and tracer samples. The target sample contains the galaxies for which we want to find a redshift distribution. The tracer sample contains galaxies with known redshifts that are used to add the clustering information into the redshift estimation. Lower panel: Redshift distribution of tomographic bins defined as in Section 3.3.

define galaxy phenotypes that individually span narrower redshift ranges. We choose among the available bands in MICE2 the DES g, r, i, z bands for both samples, and the additional CFHT u, DES Y, and VHS J, H, K bands for the deep sample, mimicking the DES wide and deep survey fields. For the deep sample, we use the true fluxes from the simulation for simplicity (in reality these will still be measured with some noise, making the relation between phenotype and colour slightly more broad), while for the wide sample we add Gaussian noise to the true fluxes by fitting a linear relation between magnitude and logarithmic magnitude error for each band using observed noise from the DES Year 1 public data.[3] We produce deep and wide photometries for all galaxies of the target sample. We finally select only galaxies that have a signal to noise above 5 in each wide band, g, r, i, z.[4] We leave for future work an accurate abundance

[3]https://des.ncsa.illinois.edu/releases/y1a1/key-catalogs/key-mof

[4]Before adding the noise, we shift each galaxy's magnitude by -1.2, to increase the number density of our target sample passing *S/N* cuts to 4.7 galaxies arcmin^{-2}, as observed in the DES Y1 Metacalibration source sample (Troxel et al. 2018a). This counteracts the MICE catalogue culling described in Section 3.1.

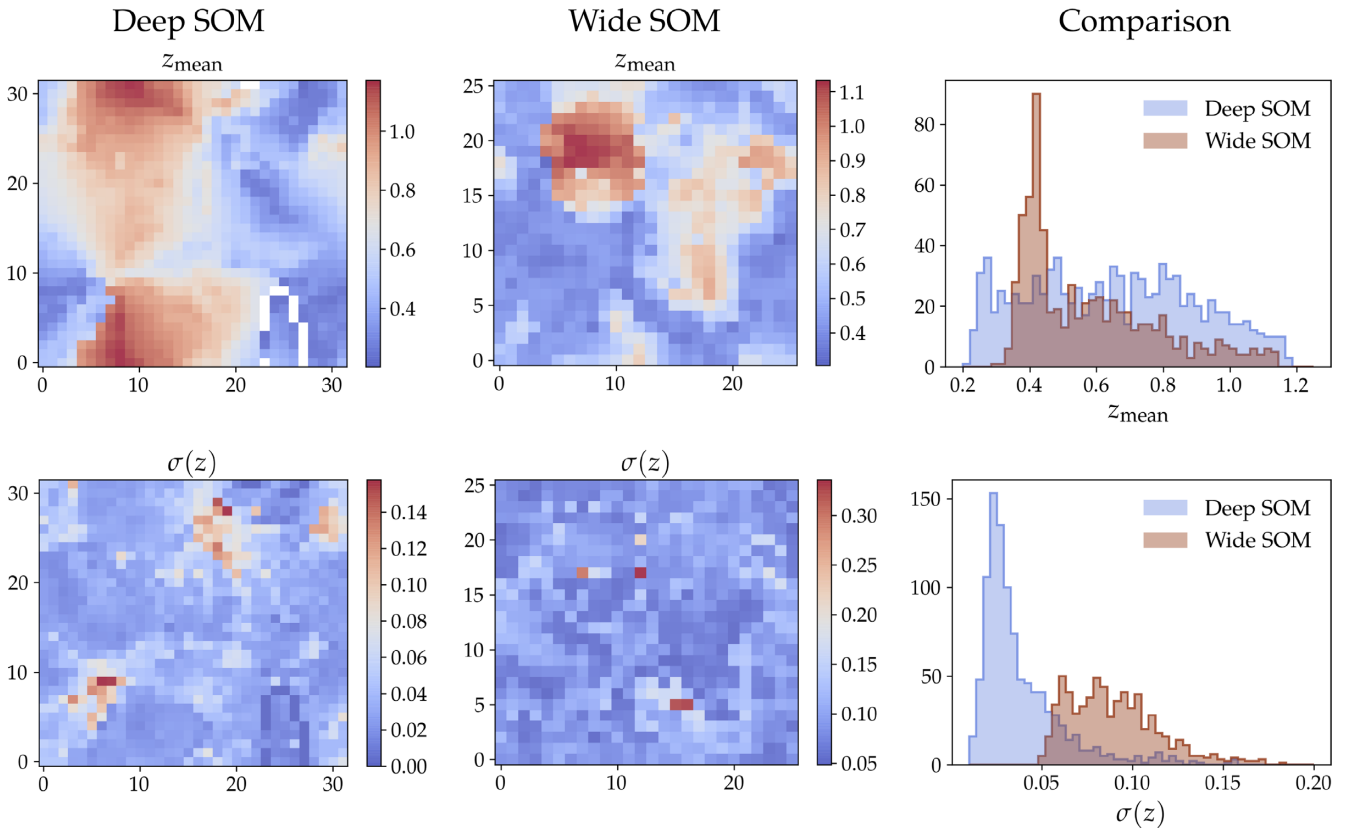

Figure 2. Mean redshift and redshift dispersion of cells in deep and wide SOMs described in Section 3.2. The left and central columns show the SOM maps populated with these quantities, while the plots in the right-hand column show the comparison of these distributions. These show how the deep SOM better samples the redshift space of the simulation test, with a lower redshift scatter per cell.

matching of the colour and magnitude distributions of the simulation to the observed ones from real galaxy surveys.

Following Section 2.2, we create two SOMs on square grids with periodic boundary conditions, each similar to the SOM in Masters et al. (2015). The deep SOM is trained with eight colours, defined as mag $-$ i, where mag $= \{u, g, r, z, Y, J, H, K\}$, in a 32 $\times$ 32 grid. The wide SOM is trained with one magnitude, i, and three colours, $g - i$, $r - i$, and $z - i$, in a 26 $\times$ 26 grid. Each colour is renormalized to span the range [0, 1], while the magnitude spans the range [0, 0.1], i.e., we give colours 10 $\times$ more weight than to the i magnitude in creating the wide SOM (we also tested giving 100 $\times$ more weight and equal weight, which performed worse). Also, to avoid noise influencing the training of the wide SOM, we only use galaxies with an $S/N >$ 10 to do so. In the simulation, we know the truth and an observed magnitude for every galaxy in the target sample, so we can assign a $\hat{c}$ from the wide SOM and a c from the deep SOM to every galaxy (the 'best-matching unit', or BMU, in SOM parlance). From these we can calibrate the probability $p(\hat{c}|c)$. In the application of this method to real data, the 'truth' (low-noise fluxes) are not available for every target but only for a subset, so that only a wide-SOM assignment $\hat{c}$ is available for all targets. But $p(\hat{c}|c)$ can be estimated through repeated injection of deep-sample galaxies into the wide images, serving the same purpose. Both methods should yield an accurate assessment of $p(\hat{c}|c)$, which is essential for success of any photometric approach to redshift estimation from noisy fluxes.

Fig. 2 shows the mean redshift and redshift dispersion of the cells in the deep and wide SOMs described above (left and central columns). From the plots, one can note the smoother redshift distribution and the lower redshift dispersion in the deep SOM compared to the wide SOM. This is even more evident from the comparison plots in the right-hand column of Fig. 2: the distribution of the mean redshift per cell in the deep SOM is more uniform and better samples the redshift space of the simulation ($0.2 < z < 1.2$), and the redshift dispersion per cell in the deep SOM is significantly lower (median $\sigma(z)$ of 0.030 for the deep SOM versus 0.086 for the wide SOM).

3.3 Tomographic bins

Tomographic redshift bins are defined as groups of wide-SOM cells. We first find the mean expected redshift for each wide cell as

$$z_{\rm mean}(\hat{c}) = \int {\rm d}z\, z \left[\sum_c p(z|c)p(c|\hat{c})\right], \qquad (11)$$

where $p(z|c)$ is also estimated using all galaxies in the target sample. We sort the wide SOM cells by their $z_{\rm mean}$, then split them into five contiguous redshift bins with equal number of galaxies. The true redshift distribution of each tomographic bin is shown in the lower panel of Fig. 1. The estimation of $n(z)$ presented in this work can be applied to any subset of the target galaxies defined by the features F_i, but here we will use as an example the determination of $n(z)$ for Bin 3 of this scheme. To do so, we first retrain the wide and deep SOMs using only those target galaxies whose noisy photometry places them into this bin. This choice potentially avoids some biases that can arise from differential bin selection within the finite range of redshifts in individual deep-SOM cells, as highlighted by Wright et al. (2020) and other works. This step can be executed on real data using the deep sample.

3.4 Spectroscopic sample

To determine a prior $p(t, z)$, we will make use of a spectroscopic sample for which both t and z are assumed to be known definitively for each galaxy passing target-sample cuts. In the simulation, the truth values are known exactly; in reality, they will typically come from a spectroscopic or high-quality photo-z sample, and span a small area of the sky and are hence subject to sample variance. They are intended to be representative of the full galaxy population, but can be subject to incompleteness and biases. Our simulated spectroscopic sample consists of all target galaxies from one HEALPIX sky pixel (Górski et al. 2005; Zonca et al. 2019) of the simulation (with NSIDE $= 2^5$), which has an area of $\sim$3.5 deg^2. The same tomographic bin selection as made on the target sample is applied to the noisy versions of the photometry for the spectroscopic galaxies, leaving around 11 000 objects having spectra, in comparison to 3.3×10^6 galaxies in this bin from the full 1000 deg^2 target sample.

In Section 6, we will investigate sample variance by choosing different regions for the spectroscopic sample, and also investigate the effects of placing measurement biases on the redshifts assumed for this sample.

4 ADDING THE CLUSTERING INFORMATION

As described in Section 2, we will work under the approximation that we can replace the latent density field of the tracer population with a set of deterministic estimators $\hat{\delta}_z(\theta)$ discretized in redshift space. We also assume that these tracers are drawn from the same generative model as the targets, up to some local biasing relation $\mathcal{B}$ with parameters b, so that we are assuming $p(\theta|z) \propto \mathcal{B}_z[\hat{\delta}_z(\theta), b_z]$.

Before proceeding to describe the density estimators and biasing functions used in this simulation, we pause to note that we do not require the resultant $p(\theta|z)$ to be perfect or unbiased. The correlation redshift method uses the density estimator $p(\theta|z)$ to inform us whether galaxies are more likely to truly be at z than to be at some $z' \neq z$. In the latter case, the target galaxies are distributed essentially randomly in θ with respect to $p(\theta|z)$. A useful figure of merit (FoM) for our density estimator is therefore the mean boost in (log) likelihood that a galaxy gets if it is assigned to its true redshift:

$$\mathrm{FoM}_z = \langle \log p(\theta_i|z) \rangle_{i\in z} - \langle \log p(\theta_i|z) \rangle_{i\notin z} \tag{12}$$

$$= \left\langle \log \mathcal{B}_z \left[\hat{\delta}_z(\theta_i), b_z\right] \right\rangle_{i\in z} - \left\langle \log \mathcal{B}_z \left[\hat{\delta}_z(\theta_i), b_z\right] \right\rangle_{i\in r}, \tag{13}$$

where the first term is evaluated for galaxies truly at z, and the second term is for a population of galaxies randomly distributed across the footprint. In the simulations we can evaluate this FoM over the full footprint, as a guide for good choices to make for the KDE and bias parameters. In real data, this estimation is possible only over the smaller spectroscopic sample.

4.1 Density estimation

The tracer population, described in 3.1, is split in 20 redshift bins equally spaced in comoving distance in the range $z \in [0.2, 1.2]$ using the true redshift from the simulation. The redshift bins are wider than the typical RMS redshift uncertainty of photometric LRGs in DES, which have, $\sigma_z \sim 0.015(1 + z)$ (Rozo et al. 2016; Vakili et al. 2019), and also wide enough to make their projected density fields nearly independent from each other. We will defer to future work any attempt to include photo-z errors in the tracer sample.

Several methods exist to reconstruct the surface density of galaxies (see e.g. Cautun & van de Weygaert 2011; Darvish et al. 2015) from a point sample. In this work, we will use a KDE to estimate the density field at any position of the field, using a circular kernel function $K(r)$:

$$\hat{\delta}_z(x) \equiv \frac{\frac{1}{N_T}\sum_T K(\theta_{xT})}{\frac{1}{N_R}\sum_R K(\theta_{xR})} - 1. \tag{14}$$

Here θ_{xT} runs over the distances between our sample point x and each of the N_T tracers at redshift z, while θ_{xR} runs over the pairs with a random sample of size N_R that describes the selection function of the tracer sample. The KDE is seen to be equivalent to the weighted two-point functions used in conventional clustering-z redshift techniques. We presume $N_R \gg N_T$ such that the dividing term can be considered a measure of the area surrounding x, taking into account the selection function and mask effects.

Choosing the shape and extent of the kernel K is important. Fig. 3 shows the effect that different kernel shapes have on the field estimate. The top left-hand panel shows a top-hat kernel of size $r_{\max} = 30$ Mpc. Such a large kernel smooths the density field too much and cannot resolve massive structures well, underestimating the density in cluster regions. The top right-hand panel shows a small top-hat KDE with $r_{\max} = 3$ Mpc. This KDE can better resolve dense structures, although it will still underestimate high-density regions, is more affected by shot noise, and indicates zero density in a large fraction of the sky. SB19 show that, in simplified limits, the most informative kernel will match the angular correlation function of the galaxies, so that $K \propto r^{-0.8}$.

Many cosmological applications of redshift inference will also use statistics of the tracer sample as part of their constraining data. Allowing large scales into the KDE can improve its estimation, but will also correlate our resultant $n(z)$ with the observables being used for cosmology, which will complicate the derivation of cosmological parameter constraints. Yet using only very small scales (< 3 Mpc) lowers the *S/N* of the density estimator and the accuracy of $n(z)$ inferences. We compromise by using a kernel that is zero for $r > r_{\max} = 15$ Mpc, although we also explore $r_{\max} = 10$ Mpc for comparison. Note that DES Y1 cosmological analyses used correlations only above $8 - 12$ Mpc h^{-1} (e.g. Krause et al. 2017). The bottom left-hand panel shows a power law $K(r) \propto r^{-0.8}$, truncated at 15 Mpc.

4.2 Biasing relation

In the simulation, we can calculate the true relation between galaxy density at some redshift z and the KDE estimator $\hat{\delta}_z(\theta)$ by calculating the true source density:

$$\mathcal{B}_z^{\mathrm{true}}(\hat{\delta}) \equiv \frac{\frac{1}{N_T} n_T(\hat{\delta})}{\frac{1}{N_R} n_R(\hat{\delta})}, \tag{15}$$

where $n_T(\hat{\delta})$ and $n_R(\hat{\delta})$ are the number of galaxies and randoms in sky regions with some (small range of) KDE value.

Fig. 4 shows, for each redshift bin (colour coded), the relation between this average density of targets as a function of KDE value with a power-law KDE with $r_{\max} = 10$ Mpc. If the KDE delivered a perfectly unbiased field estimation, this would yield the dashed line. In general, the KDE estimate will not deliver such an estimate, both because the KDE yields a biased estimate of tracer density, and because the tracer will be a biased tracer of the target galaxies. There is always a $\mathcal{B}^{\mathrm{true}}$ which will optimize the performance of a given KDE. In real data we will not know this function in advance, so we propose a parametric form for the true probability $p(\theta|z)$ of a target

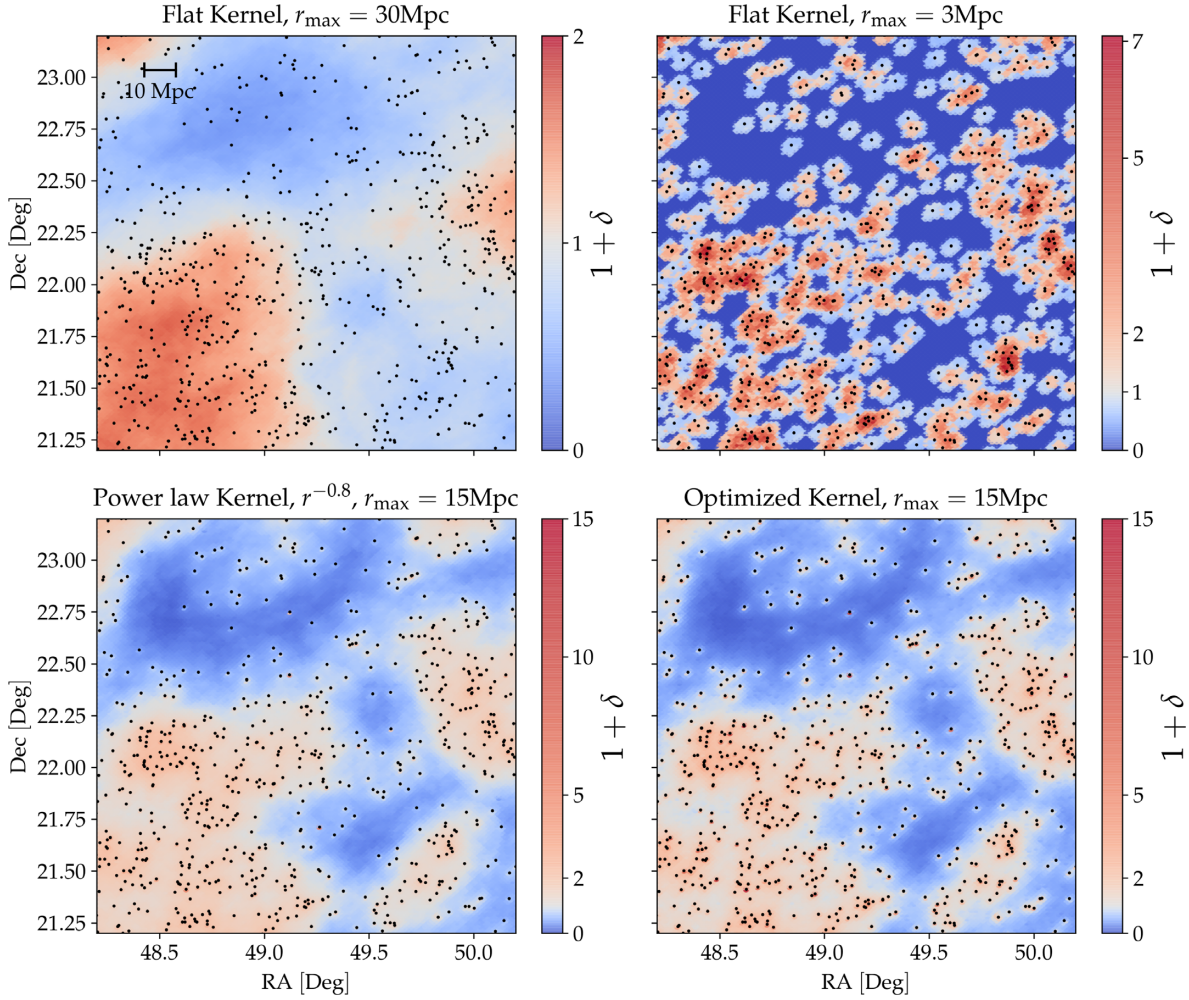

Figure 3. Density field estimation using different kernel density estimators from a tracer sample population. Shows the field estimate for a small patch in the highest redshift bin. The black dots show the position of the tracer galaxies, and the background colours show the estimated value of the density field at different positions. The top panels show a flat kernel with a large size ($r_{\rm max} = 30$ Mpc, left-hand panel) and a small size ($r_{\rm max} = 3$ Mpc, right-hand panel). The bottom left-hand panel shows the density with a power-law kernel that better resolves the structures. The bottom right-hand panel shows a field estimated with an optimized kernel, which is our default density field estimate. Note the change in colour scales in different panels, with white always corresponding to the mean density.

galaxy being at position θ_i and redshift z,

$$p(\theta_i|z) = \mathcal{B}\left(\hat{\delta}_z(\theta_i), \{b_k^z\}\right), \tag{16}$$

where $\{b_k^z\}$ are the parameters of $\mathcal{B}$ at redshift z. This is an approximation of a more general approach where the density field is updated locally by the targets as part of the hierarchical model. The parameters $\{b_i^z\}$ are part of the framework parameters (see Section 2) and they will be sampled along with the other parameters in the HBM (see Section 5). We choose a polynomial of degree four as our mapping function $\mathcal{B}$, such that

$$\log_{10}(p(\theta|z)) = \log_{10}\mathcal{B}_z\left[\hat{\delta}_z(\theta)\right] = \sum_{k=0}^{4} b_k^z \log_{10}(1.1+\hat{\delta}_z)^k, \tag{17}$$

with the additional constraints that $\int p(\theta|z)\mathrm{d}\theta = 1$ and that the derivative must always be positive. Note the use of $(1.1+\hat{\delta}_z)$ on the right-hand side to avoid singularities when the KDE yields $\hat{\delta} = -1$.

The use of a parametric biasing function adds another criterion to the choice of KDE kernel, because we will prefer a kernel which yields a more linear, less complex biasing function which we can expect to require fewer parameters and less variation with redshift. These characteristics will improve our ability to fit optimal biasing functions to the KDE output.

While the biasing relation in general depends on both redshift and phenotype (see Section 2), we are neglecting the phenotype dependence throughout this work. The redshift determination could potentially be improved by, for example, allowing red galaxies a distinct bias from blue galaxies. There will be potential degradation, though, as more free parameters are introduced into the model. We defer an attempt at using this information for a future work.

4.3 Optimizing the estimator

We can go one step further and try to optimize the shape of the KDE kernel, assuming we have a small calibration patch where the redshifts of the target galaxies are known. For this purpose, we define a KDE with shape

$$\mathrm{KDE} \propto r^{\alpha} \exp\left[-\left(\frac{r}{r^*}\right)^{\gamma}\right], \tag{18}$$

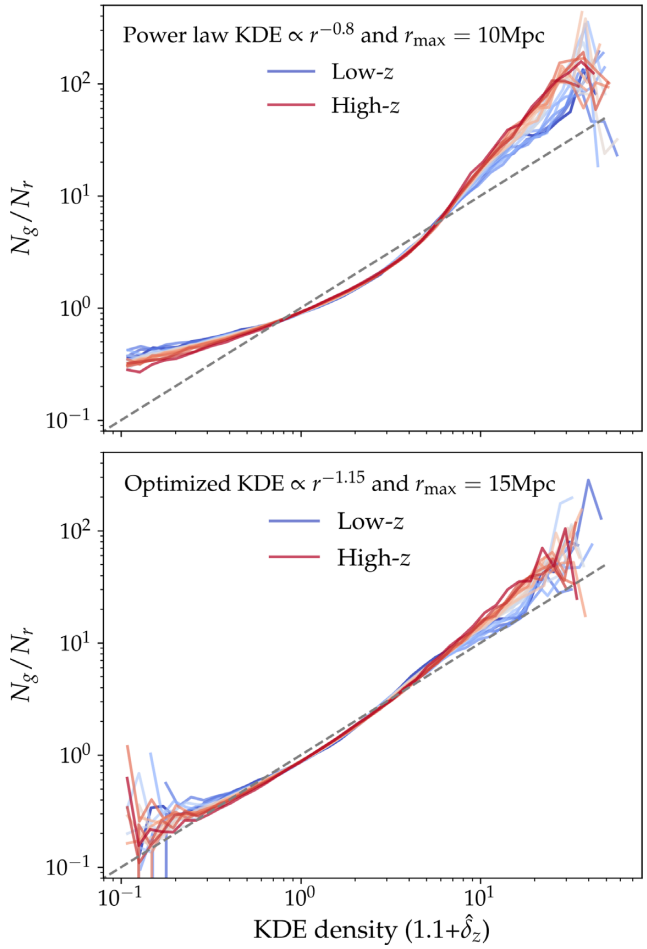

Figure 4. Upper panel: Ratio between the abundance of target galaxies and random points as a function of estimated KDE density, for a power law KDE $r \propto r^{-0.8}$ and $r_{\rm max} = 10$ Mpc. The different redshift bins are colour coded. If the KDE delivered a perfectly unbiased field estimate of the target galaxies, we would expect to find the dashed line relation. All galaxies have been used without tomographic bin selection to obtain a better estimate. The true redshift of all the target galaxies was used, while in a real data scenario one could only estimate this relation in the smaller calibration fields. Lower panel: Same as upper panel, but using an optimized KDE with $r_{\rm max} = 15$ Mpc. The KDE is optimized from a function that combines a power law and an exponential truncation at small scales to deal with shot noise effects (see Fig. 5). The optimal parameters are found from a calibration field from $\sim$3.5 deg^2 where redshifts for the target galaxies are known. It shows a more linear relation, although remains substantially non-linear at the extremes of density.

which combines a power law with exponent α and an exponential truncation of the power law at scale r^* with width parameter γ. Fig. 5 compares this kernel shape to a power law. The motivation for allowing a truncation at small scales is to reduce the effect of shot noise for sparse tracer samples.

The optimization of the KDE works as follows. We write the probability of the optimized KDE parameters for redshift z as

$$p(\alpha_z, r_z^*, \gamma_z|\theta, z) \propto p(\theta|z, \alpha_z, r_z^*, \gamma_z)\; p(\alpha_z, r_z^*, \gamma_z|z), \tag{19}$$

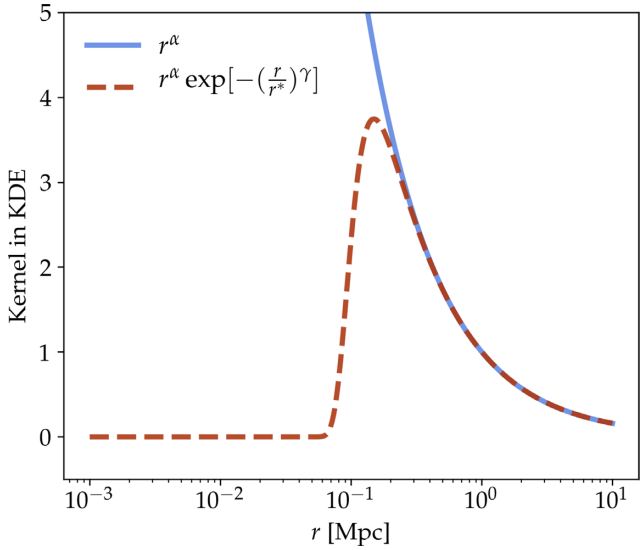

Figure 5. Comparison between a power-law KDE and a KDE with a power law that truncates at some scale r^*. Such truncation reduces the impact of shot noise in smaller scales and naturally adds a small exclusion region around the positions of tracers.

where the last term is the prior on the parameters. Given a sample of targets with known redshifts from a calibration field,

$$p(\theta|z, \alpha_z, r_z^*, \gamma_z) \propto \prod_{i\in z} p(\theta_i|z, \alpha_z, r_z^*, \gamma_z)$$
$$= \prod_{i\in z} \mathcal{B}\left(\hat{\delta}_z(\theta_i, \alpha_z, r_z^*, \gamma_z), \{b_k^z\}\right), \tag{20}$$

where $p(\theta_i|z, \alpha_z, r_z^*, \gamma_z)$ is the probability of the ith galaxy at redshift z from the calibration sample. We obtain this probability by biasing the KDE estimate $\hat{\delta}_z(\theta_i)$ using the $\mathcal{B}_z^{\rm true}$ from equation (15), estimated using only the galaxies from the calibration sample. Note that since we know the true redshifts in the calibration sample, we do not need to use the parametric form from equation (16) but directly use the estimate from equation (15).

When using a small patch of $\sim$3.5 deg^2 to optimize the kernel parameters, we take the average of the maximum-posterior parameters across all redshift bins as an estimate for the optimized KDE, since the constraining power in each redshift bin is weak. We use top hat priors $\alpha_z \in [-2, -0.5]$, $r_z^* \in [0.001, 0.1]$, and $\gamma_z \in [-10, -2]$. The γ parameter has little effect on the posterior so we fix it to its mean value of $\gamma = -4$ and run again. For a kernel limited to $r_{\rm max} = 15$ Mpc we find $<\alpha_z> = -1.15$ and $< r_z^* >= 0.018$Mpc. For a KDE limited to $r_{\rm max} = 10$ Mpc we find $<\alpha_z> = -1.0$ and $< r_z^* >= 0.010$Mpc. Note how a more aggressive power law is preferred when the size of the KDE is larger. The lower panel of Fig. 4 shows the biasing relation (estimated using all target galaxies in the simulation) for the optimized kernel with $r_{\rm max} = 15$ Mpc, which is much closer to the ideal relation than the $\alpha = -0.8$, $r_{\rm max} = 10$ Mpc power-law kernel shown in the upper panel. This is both a consequence of having 2.25 $\times$ more area and of optimizing the kernel shape. The bottom right-hand panel of Fig. 3 shows the density field estimated with the optimized KDE with $r_{\rm max} = 15$ Mpc. This will be our default kernel for further testing.

We compute the FOM of equation (13) for several choices of kernel. For this purpose (but not for the results in Section 6), we use the $\mathcal{B}^{\rm true}$ biasing function estimated using all galaxies. The median FOM value across redshift for the optimized kernel with $r_{\rm max} =$

15 Mpc is 0.263, while for the power law we measure 0.240. For $r_{\rm max} = 10$ Mpc, we find a median FOM of 0.266 for the optimized kernel and 0.219 for the power law. The average information gain per galaxy from optimizing the KDE is $2-5$ per cent, and we find the information to be similar for both $r_{\rm max}$ limits once the kernel shape has been optimized. We select as our default the optimized kernel with $r_{\rm max} = 15$ Mpc since it has a more linear and easier-to-model $\mathcal{B}$ shape. In general, the shape of the optimal KDE and the shape of the biasing relation (Fig. 4) depend on the tracer sample density per unit comoving surface area, among other factors. Here we choose a tracer sample with constant comoving density, which minimizes the variation from this effect across redshift.

5 SAMPLING

Now we turn to the problem of sampling over the redshift and type probability distributions of populations of galaxies and their individual constituents, in the framework of the hierarchical Bayesian model described in the previous sections. It is complicated to simultaneously sample all variables from the joint posterior $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{t}, \boldsymbol{b}|\boldsymbol{F}, \boldsymbol{\theta})$ in equation (8). We will show, however, that it is feasible to draw samples from this posterior using a three-step Gibbs sampler. This is because the conditional posterior distributions of interest can be easily written and sampled. In SB19, the true values of the density field at each position were known, and hence there was no need to sample over the parameters $\{b_k\}$ defining the biasing function $\mathcal{B}\left(\hat{\delta}_z(\theta_i), \{b_k^z\}\right)$ relating the true galaxy density to the KDE (see Section 4). This paper's implementation executes sampling over bias parameters, including the development of some key sampling strategies that will enable a future application to real data.

We use information from all galaxies in the target sample to constrain the redshift and type probability distributions of the galaxy population. Additionally, the fully Bayesian nature of this scheme allows us to make use of prior information on the relevant quantities, when available. In this work, we will assume that we have access to a 'spectroscopic sample' of the galaxies with known z, t, e.g. from a complete spectroscopic survey of a random subsample of targets in a small region of the sky. We will also assume that we can identify a tracer population among the spectroscopic sample, with the same selection as the corresponding tracers in the full sample. These subsamples will place an informative prior on the coefficients $\boldsymbol{f}$, and will also be important in sampling over the biasing function parameters described in Section 4.

5.1 Three-step Gibbs sampler

Each iteration of the Gibbs sampler comprises three steps which are (i) drawing a sample of $\boldsymbol{f}$ from $p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{b}, \boldsymbol{F}, \boldsymbol{\theta})$, (ii) drawing pairs of z_i, t_i for each galaxy i from $p(z_i, t_i|\boldsymbol{f}, \boldsymbol{b}, F_i, \theta_i)$ using the newly drawn $\boldsymbol{f}$, and (iii) drawing a sample of the biasing function parameters $\boldsymbol{b}$ for each redshift bin from $p(\boldsymbol{b}_z|\boldsymbol{f}, z, \boldsymbol{t}, \boldsymbol{F}, \boldsymbol{\theta})$ given the z_i assignments in step (ii). The conditional distributions can be derived from the joint distribution in equation (8). The first two steps of the sampler are as in SB19 and hence we skip the full derivation for brevity (see SB19 for more details), and the third step is new and is considered in more detail.

(i) The conditional posterior on $\boldsymbol{f}$ depends on the counts of sources of $\boldsymbol{z}$ and $\boldsymbol{t}$ (in the last iteration), $\boldsymbol{N} = \{N_{zt}\}$ where N_{zt} is the number of sources assigned to redshift z and phenotype t, and it also depends on the prior information on $\boldsymbol{f}$, $p(\boldsymbol{f})$:

$$p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{b}, \boldsymbol{F}, \boldsymbol{\theta}) \propto p(\boldsymbol{f}) \prod_{z,t} f_{zt}^{N_{zt}}. \tag{21}$$

The prior condition that $\sum f_{zt} = 1$, and $0 \leq f_{zt} \leq 1$, allows us to write the conditional posterior on $\boldsymbol{f}$ as a Dirichlet distribution. Following the derivation in SB19, if $\boldsymbol{M} = \{M_{zt}\}$ are the counts of the prior sample found at each z, t pair, and we assume that each spectroscopic galaxy has been drawn independently from the distribution, then the prior distribution of $\boldsymbol{f}$ follows a Dirichlet distribution with parameters $\boldsymbol{M}$. In this case the conditional posterior follows a Dirichlet on the data counts from the last iteration plus the prior counts:

$$p(\boldsymbol{f}|\boldsymbol{N}) \sim \mathrm{Dir}(\boldsymbol{N} + \boldsymbol{M}), \tag{22}$$

$$\text{with } \mathrm{Dir}(\boldsymbol{N}) \equiv (N + N_z N_t - 1)!\, \delta_D \left(1 - \sum_{zt} f_{zt}\right) \times \prod_{z=1}^{N_z} \prod_{t=1}^{N_t} \frac{\Theta(f_{zt}) f_{zt}^{n_{zt}}}{n_{zt}!}. \tag{23}$$

An important shortcoming of our scheme is that the spectroscopic sample will not usually satisfy the condition that all galaxy draws are independent, because it is taken from a limited sky area and thus subject to large-scale-structure variance. The posterior will therefore not sample this form of variance. The addition of sample variance uncertainties into the prior sampling will be explored in a future publication.

(ii) For each galaxy, the posterior for the z_i, t_i pair conditioned on $\boldsymbol{f}$ and $\boldsymbol{b}$ is

$$p(z_i, t_i|\boldsymbol{f}, \boldsymbol{b}, F_i, \theta_i) \propto \mathcal{L}_{it_i}\ f_{t_i z_i}\ \mathcal{B}\left(\hat{\delta}_{iz_i}(\theta_i), \boldsymbol{b}_{z_i}\right), \tag{24}$$

where apart from using the $\boldsymbol{f}$ obtained in the first step of the sampler (i), we make use of the measurement likelihood $\mathcal{L}_{it_i}$ and the clustering terms $\mathcal{B}$ discussed above. The sampling in this step (ii) will produce pairs of z, t for each galaxy that constitute the next realization of $\boldsymbol{N} = \{N_{zt}\}$, to be used in the step (i) of the next iteration of the Gibbs sampler.

(iii) After we have z assignments for all galaxies in the sample from step (ii), we can now separate galaxies into redshift bins according to those assignments. Then, for each redshift bin, the posterior on the biasing function of that bin conditioned on all other variables looks like:

$$p(\boldsymbol{b}_z|\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{t}, \boldsymbol{F}, \boldsymbol{\theta}) = p(\boldsymbol{b}_z|\boldsymbol{z}, \boldsymbol{\theta}) \propto \prod_{i:z_i=z} \mathcal{B}\left(\hat{\delta}_{iz_i}(\theta_i), \boldsymbol{b}_{z_i}\right). \tag{25}$$

With the choice of parametric biasing function in equation (17), there is no direct sampling algorithm for this conditional posterior. We therefore use the following procedure: first, we run a Metropolis–Hastings (MH) Markov Chain Monte Carlo (MCMC) sampler for the conditional posterior in equation (25) for each redshift bin where we restrict the galaxies to the spectroscopic sample. Since the spectroscopic sample have fixed z_i, this chain can be run once, before the Gibbs sampling commences, and yields a sampling of the prior on bias parameters inferred from the spectroscopic sample (see Appendix A). Next, at each iteration of the Gibbs sampler, we return the 5000th sample from an MH MCMC chain run on equation (25) using all target galaxies currently assigned to a given redshift (we have performed this step with MCMC chains longer than 5000 steps, with consistent results). The proposal distribution

for this MH sampler is to draw at random from the output sampling of the prior. Effectively we are using the target sample for importance-sampling of the prior sample. This procedure is a robust way to combine the prior and target conditionals without the need to tweak the proposal distributions or the parameter limits of the MCMC chains. It is also very fast compared to step (ii) of the Gibbs sampler.

6 RESULTS

We use the simulation described in Section 3 to test the methodology developed throughout this work. The target sample for this section is the third tomographic bin in Fig. 1, which contains $\sim 3.3 \times 10^6$ objects. The spectroscopic sample, for which redshift and type are assumed known, consists of all 11 000 target galaxies from one patch of sky with area $\sim 3.5\,\mathrm{deg}^2$. These objects are used to estimate the prior probability $p(z, c)$ and obtain the sampled prior on the mapping function parameters $\mathcal{B}(\hat{\delta}, \{b_i\})$ (see Section 5 for details about the sampling).

The HBM method yields samples of the redshift and type posterior for each individual galaxy; the redshift and type posterior of the population; and the posterior of the biasing function parameters. We focus on the redshift population posterior, marginalizing over all other parameters, since this is what is usually needed in cosmological analyses of galaxy surveys. In particular, current and future weak lensing analyses are very sensitive to small biases in the mean redshift of the distribution, which can become the leading systematic uncertainty. Therefore, in analysing our results, we define one quality metric to be the difference between the mean of each sample j of our redshift posterior and the true mean from all the target galaxies,

$$\Delta z_j = \langle z_{\mathrm{est,j}} \rangle - \langle z_{\mathrm{true}} \rangle \,. \tag{26}$$

Since we draw samples of the full redshift distribution posterior f_z, another useful metric that is sensitive to the distribution shape is the Kullback–Leibler divergence (D_{KL}) between each sample and the true redshift distribution,

$$D_{\mathrm{KL}}(f^{\mathrm{est}}_{z,j} || f^{\mathrm{true}}_{z}) = \sum_z f^{\mathrm{est}}_{z,j} \log\left(\frac{f^{\mathrm{est}}_{z,j}}{f^{\mathrm{true}}_{z}}\right) . \tag{27}$$

This is a measurement of the relative entropy between the true distribution and the recovered distribution, and can be used to see how much information the photometry and density estimates are adding with respect to the prior knowledge. A Kullback–Leibler divergence of 0 indicates that the two distributions in question are identical, and the lower and closer to 0 its value gets the more similar the two distributions are, as it shows the expected value of the log differences between two distributions. Therefore, if the distributions have an expected divergence of one order of magnitude (i.e. are really different), the D_{KL} will have a value of $\log(10) \approx 2.30$, whereas if they differ by 0.1 orders of magnitude, $D_{\mathrm{KL}} \approx 0.23$.

For each case we investigate, we sample $n(z)$ from three distributions: (1) the prior only; (2) the posterior from an HBM that only includes photometry information; and (3) the posterior from an HBM that includes both photometry and clustering information, marginalizing over the biasing parameters. We denote the HBM with photometry as F (feature) and the HBM with photometry and clustering as $F + \delta$. The F inference is essentially a rigorous application of the reweighting method of Lima et al. (2008).

In the first part of this section, we look into the impact of sample variance in the prior coming from the calibration sample. In the second part, we study how the method performs when the prior on the $p(z, t)$ probability from a calibration sample is modified and biased. For each case, we will show a violin plot of the posterior redshift distribution compared to the true distribution, the distribution of Δz_j differences, and the distribution of D_{KL} divergences.

6.1 Sample variance in the prior

As noted in Section 5.1, we have adopted a Dirichlet prior on $p(z, c)$ that assumes that galaxies drawn from the small spectroscopic-sample patch of sky have independent phenotypes and redshifts. This neglects sample variance from large-scale structure (hereafter just 'sample variance'), which adds noise to the estimated relative density of galaxies at given redshift and type f_{tz} (Cunha et al. 2012). This effect is larger at lower redshifts, where the volume is smaller.

Sample variance most importantly affects the density of types $p(t)$, where $p(z, t) = p(z|t)p(t)$, since the same phenotype would yield the same redshift regardless of where it is observed, provided the redshift distributions of phenotypes are narrow. However, we have seen in Fig. 2 that there are some phenotypes (deep cells) with wider redshift distributions, mostly due to colour–redshift degeneracies. As a result, the redshift distribution $p(z|t)$ of these phenotypes is also affected by sample variance. The Dirichlet sampling of the prior, as presented in Section 5, neglects sample variance uncertainty, but we expect the HBM method to reduce the effect of sample variance in the prior since the target population is much larger than the prior sample. Nevertheless, limited sampling or shot noise from the prior in any of the phenotypes can lead to a noise bias of $p(z|t)$, and make the HBM reconstruction imperfect.

To assess this sample variance, we randomly choose 11 calibration samples of $\sim 3.5\,\mathrm{deg}^2$ each, and apply the HBM method to each, with and without using clustering information. In Fig. 6, we show the results of these runs in the two metrics defined above, i.e. the mean of the redshift distribution and the KL divergence compared to the truth. For each method of inference (prior-only, F, and $F + \delta$), we show the mean of both metrics over the 11 distinct spectroscopic patches, with three different uncertainty estimations: (1) the total standard deviation among all MCMC samples of all spectroscopic patches; (2) the standard deviation of the means of the 11 different prior patches; and (3) the standard deviation within the MCMC samples of one patch. The figure shows that:

(i) The sample variance among patches (2) dominates the total uncertainty budget (1) in every case.

(ii) The HBM (F) reduces the uncertainty in the estimation of the mean redshift, i.e. lessens the impact of spectroscopic sample variance, and also improves the $N(z)$ shape reconstruction (lower KL divergence values) compared to the prior-only inferences.

(iii) The addition of the clustering further reduces the uncertainty in the mean redshift and improves the $N(z)$ reconstruction.

The HBM mean redshift uncertainty goes from $(0.0 \pm 4.2) \times 10^{-3}$ in the prior to $(1.0 \pm 1.6) \times 10^{-3}$ for HBM (F) and $(0.8 \pm 1.2) \times 10^{-3}$ for HBM ($F + \delta$). The shape improves from a $\log_{10}(D_{\mathrm{KL}})$ divergence of 4.69 ± 0.17 in the prior to 4.40 ± 0.17 and 4.11 ± 0.23 for HBM (F) and HBM ($F + \delta$), respectively.

In Fig. 7 we randomly choose one of the spectroscopic-sample patches for the prior, and compare the posterior from running an HBM with photometry alone (F, blue), an HBM with photometry and clustering ($F + \delta$, red) and samples drawn from the Dirichlet prior on $p(z, t)$ (orange). The prior $p(z, t)$ has a mean redshift bias of $\Delta z = (-1.0 \pm 0.1) \times 10^{-2}$, arising from LSS sample variance in this single sky patch. When running the HBM, we find the bias reduced to $\Delta z = (-7.2 \pm 4.4) \times 10^{-4}$ with photometry alone and

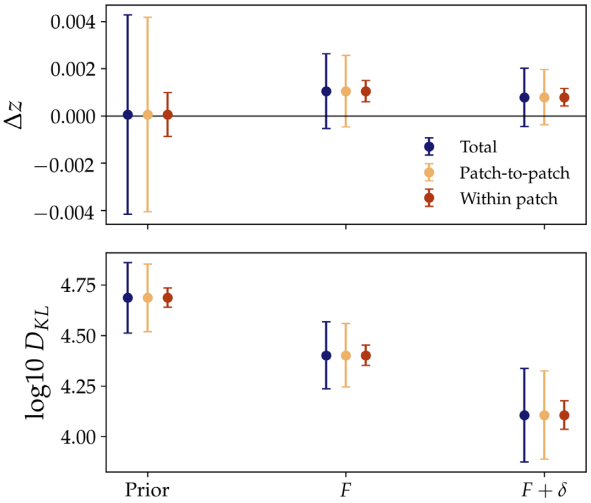

Figure 6. Performance on the posterior redshift probability distribution for a hierarchical Bayesian model (HBM) with photometry and clustering information. Two metrics are shown, the bias in mean redshift distribution Δz (upper panel) and the Kullback–Leibler divergence $D_{\rm KL}$ between the posterior samples and the true distribution (lower panel). The prior information comes from a small patch of $\sim$3.5 deg^2. We show results grouped in three blocks which show the results from drawing Dirichlet samples directly from the prior (labelled as Prior), from drawing samples using an HBM with only photometry (F) and from an HBM with both photometry and clustering ($F + \delta$). The total error budget (blue) is estimated from the standard deviation of samples drawn from HBM chains run in 11 randomly distributed patches of the same size. We also show the contribution to the total error of the sample variance (yellow, Patch-to-patch) and the mean internal variance of each chain (red, Within patch), finding the former one dominates the error budget in every case. The HBM reduces the sample variance uncertainty from the prior and significantly improves the recovered shape when also adding the clustering.

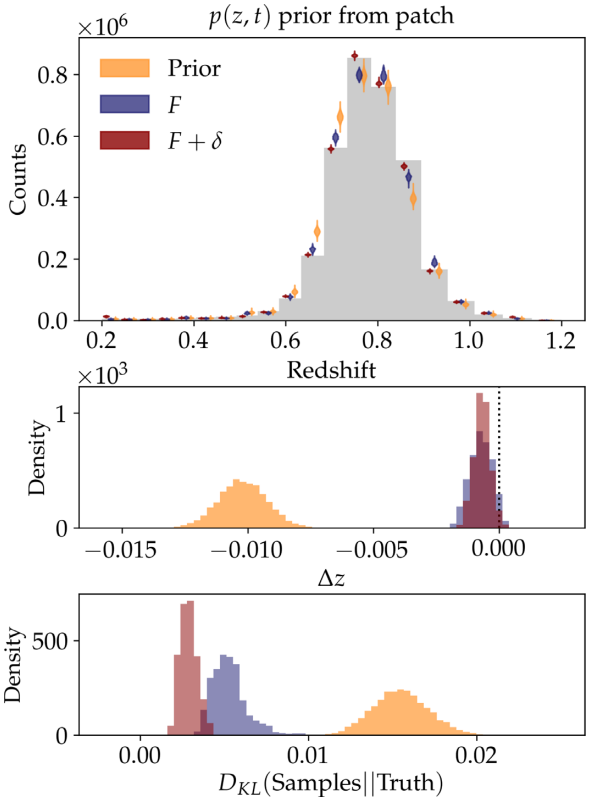

Figure 7. Posterior redshift probability distribution, marginalized over type and when including clustering marginalizing over mapping function parameters. The prior is obtained from a small calibration patch with 10 758 objects over an area of $\sim$3.5 deg^2. The three plotted distributions are obtained from: the prior; the posterior for an HBM with photometry only, F; and the posterior for an HBM with photometry and clustering $F + \delta$. Top: Shows violin plots for each distribution compared to the true redshift distribution (grey). Middle: Shows the posterior distribution of redshift bias Δz values. Bottom: Shows the distribution of Kullback–Leibler divergence ($D_{\rm KL}$) between each sample and the true redshift distribution. The HBM (F) removes most of the redshift bias, since in this case the prior's redshift bias is primarily caused by biases in the type density $p(t)$ caused due to the sample variance of the calibration patch. The addition of clustering sharpens the distribution and improves the overall shape, reducing the $D_{\rm KL}$ divergence.

a bias of $\Delta z = (-6.7 \pm 3.2) \times 10^{-4}$ when adding clustering. In agreement with Fig. 6, we find the HBM with photometry alone, i.e. reweighting (Lima et al. 2008; Sánchez et al. 2014), to be able to correct redshift biases that come from an LSS-biased type probability $p(t)$ (SB19). Since sample variance mostly changes $p(t)$, having feature information is enough to remove most of the redshift bias. In this case that an unbiased spectroscopic sample is available for $\approx 10^4$ galaxies, the addition of clustering information has little impact on the overall redshift bias. Adding the clustering information does, however, further tighten the Δz posterior distribution and also improves the shape of the redshift posterior, leading to a smaller $D_{\rm KL}$ divergence.

6.2 Biases in the prior

So far we have assumed our prior is an unbiased estimate of the underlying distribution in the spectroscopic patch, so it was only affected by sample variance. We now introduce several possible biases in the spectroscopic prior, mimicking some effects that we could find in real data, and analyse the ability of the HBM to overcome these biases. We will use same spectroscopic patch used in creating Fig. 7.

6.2.1 *Prior* p(*z*, t) *with a redshift bias*

We add a systematic redshift bias for each phenotype/deep cell by altering its redshift distribution to

$$p'(z|t) \propto p(z|t) * (21 - z), \quad z = 1, 2, \ldots, 20. \tag{28}$$

Therefore, the prior $p(z, t) = p'(z|t)p(t)$ now has a systematic bias towards low redshift. Fig. 8 shows the HBM results for such a prior. Drawing only from the prior, the mean redshift bias is $\Delta z = (-1.4 \pm 0.1) \times 10^{-2}$. The HBM with only photometry has a mean posterior redshift bias of $\Delta z = (-4.3 \pm 0.4) \times 10^{-3}$, while an HBM with photometry and clustering yields $\Delta z = (-1.8 \pm 0.3) \times 10^{-3}$. Note that the F-only HBM has corrected the same amount of redshift bias as in the previous case with unbiased prior ($\sim$0.01 in Δz), i.e. the sample variance, but cannot correct any of the systematic bias introduced in $p(z|t)$. The $F + \delta$ HBM, however can use the clustering information to further improve the $p(z|t)$ probability and reduce the total redshift bias. It also reduces the $D_{\rm KL}$ divergence, improving the overall shape.

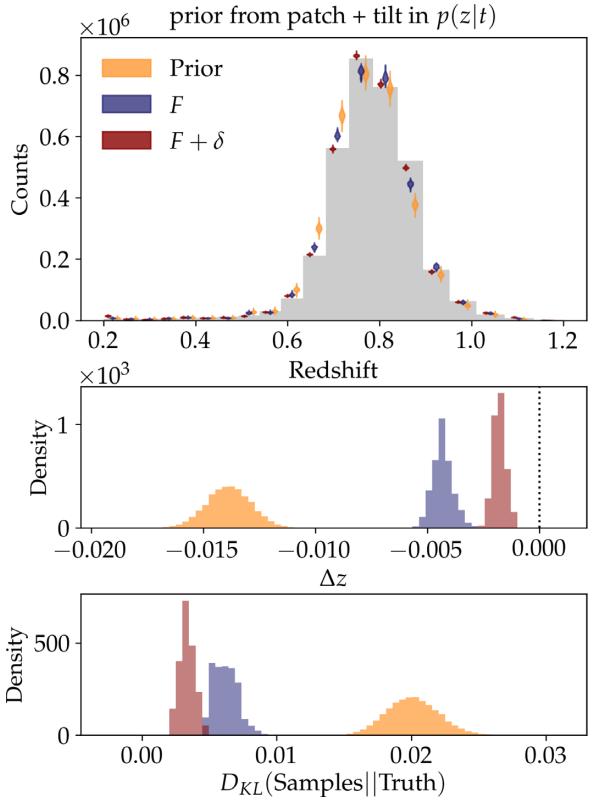

Figure 8. Similar to Fig. 7. The prior, which is obtained from the same small calibration patch, is systematically biased in the conditional redshift probability of each type $p(z|t)$ towards low redshift as per equation (28). The HBM with photometry alone reduces the redshift bias by the same amount as in Fig. 7, since it only corrects redshift biases produced by a bias in $p(t)$. The remaining bias can only be corrected with the addition of clustering, which further reduces this bias and improves the redshift posterior shape.

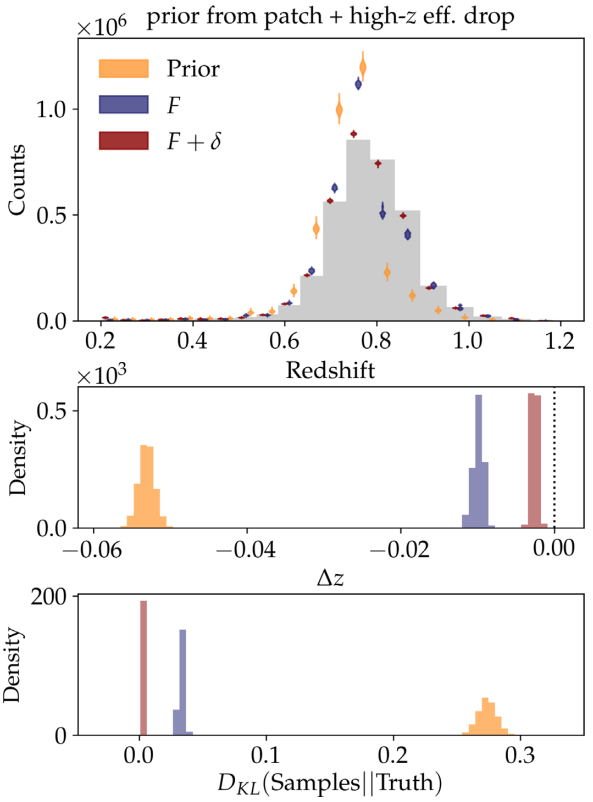

Figure 9. Similar to Fig. 7. The prior mimics a hypothetical spectroscopic efficiency drop above redshift $z > 0.8$ by reweighting the prior with a factor 0.2 in the seven highest-redshift bins. The HBM with photometry is able to correct the redshift posterior in redshift bins far away from $z \sim 0.8$, where the drop happens, by changing the density of deep cells whose redshift probability $p(z|t)$ does not cross $z \sim 0.8$. It increasingly fails to correct the redshift distribution around $z \sim 0.8$ since it cannot modify $p(z|t)$. Adding clustering significantly improves the redshift distribution, removing most of the redshift bias and largely improving the redshift distribution shape.

6.2.2 *Prior* p(*z*, t) *with a redshift efficiency drop*

Spectroscopic surveys usually present sharp selection effects in redshift due to their limited wavelength coverage of the spectra. Using such surveys to estimate the prior probability can bias the whole posterior redshift distribution of the weak lensing samples. In this section we use a prior $p(z, t)$ from a hypothetical spectroscopic survey with an efficiency drop above redshift $z > 0.8$ (the seven highest-redshift bins). We assume only 20 per cent of the galaxies in the last seven redshift bins have been successfully measured with the failed measurement being simply discarded from the catalogue, which we implement by multiplying by 0.2 the prior $p(z, t)$ in those bins.

Fig. 9 shows that this efficiency drop creates a huge redshift bias in the prior of $\Delta z = (-5.3 \pm 0.1) \times 10^{-2}$. For the F HBM we find a redshift bias of $\Delta z = (-\ 9.9 \pm 0.7) \times 10^{-3}$, while for the $F + \delta$ HBM we find $\Delta z = (-2.6 \pm 0.4) \times 10^{-3}$. The F HBM is able to successfully correct redshift bins which are far away from where the efficiency drop happens ($z \sim 0.8$) since there are many deep cells with a very tight redshift-type relation. However, it has more difficulty recovering the redshift distribution closer to the drop, since it cannot update $p(z|t)$. Adding the clustering significantly improves the recovered shape, finding a much better $D_{\rm KL}$ divergence, and eliminates 95 per cent of the redshift bias from the prior.

6.2.3 *Prior* p(*z*, t) *with degraded* $z -$ t *correlation and biased*

So far we have assumed we have a calibration field with spectroscopic data that provide a tight redshift–colour relation. Now we explore what happens if we loosen this assumption and pretend that the redshift information in the prior does not come from spectroscopy but from a hypothetical photometric redshift sample. This can be of interest in real data when spectroscopic redshifts can only sparsely populate the prior on $p(z, t)$. To mimic this effect, we convolve the conditional redshift probability for each type $p(z|t)$ with a top hat function with width of seven redshift bins, which smooths the redshift probability. The median redshift dispersion of the deep cells goes from $\sigma(z) = 0.025$ to $\sigma(z) = 0.1$, significantly reducing the correlation between types and redshift. Furthermore, we add the same systematic redshift bias to each $p(z|t)$ as in Section 6.2.1. Note the sample variance in $p(t)$ is left unchanged.

Fig. 10 shows the broadening effect in the prior, which now has a redshift bias of $\Delta z = (-3.9 \pm 0.1) \times 10^{-2}$. The HBM with photometry alone, which can only modify the density of types, is barely able to change the redshift distribution since the correlation between redshift and type has been degraded, finding a redshift bias of $\Delta z = (-3.0 \pm 0.1) \times 10^{-2}$, and a very similar $D_{\rm KL}$ divergence. In contrast, adding the clustering remarkably improves the redshift bias and shape, leading to a very large decrease in both $D_{\rm KL}$ and Δz

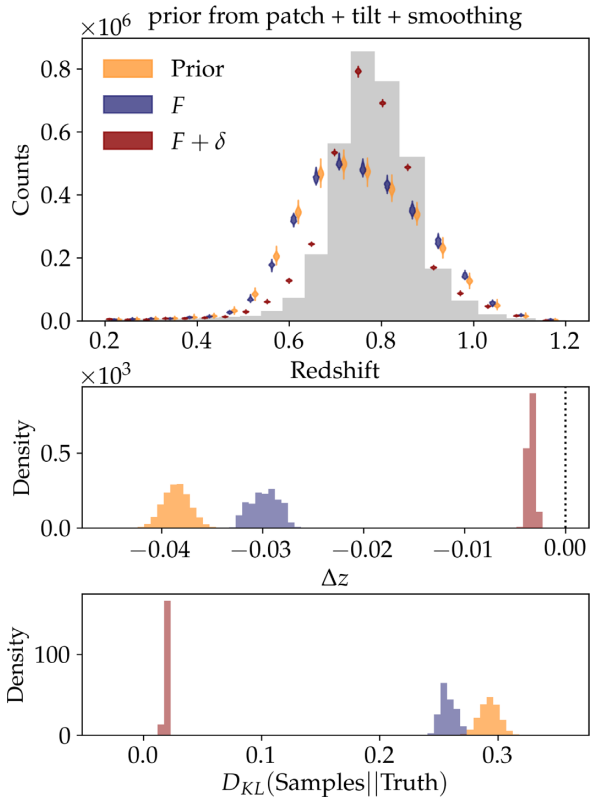

Figure 10. Similar to Fig. 7. The prior is smoothed by convolving $p(z|t)$ with a top hat function of size seven redshift bins, increasing the median redshift dispersion of the deep cells goes from $\sigma(z) = 0.025$ to $\sigma(z) = 0.1$, which reduces the correlation between type and redshift for all deep cells. In this case, the HBM with photometry alone can barely modify the redshift distribution, since there is little correlation between type and redshift. In contrast, adding the clustering information remarkably improves the redshift distribution recovery and reduces most of the redshift bias. This shows that photometric redshift surveys with wider $p(z|t)$ estimation can be used instead of spectroscopic surveys when clustering is available.

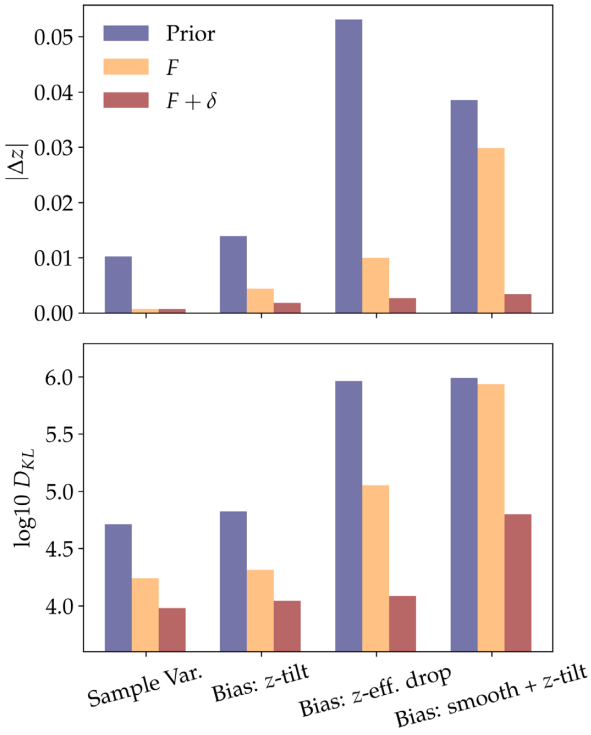

Figure 11. A summary of the quality of $n(z)$ inferences obtained in this work. Upper: the absolute redshift bias in the mean posterior redshift $|\Delta z|$. Lower: the Kullback–Leibler divergence $D_{\rm KL}$ between the posterior samples of $n(z)$ and the true distribution. We show the performance metrics are grouped in blocks of three, showing samples from: (1) the Dirichlet prior (labelled as 'Prior') obtained directly from a spectroscopic survey of $\sim$3.5 deg^2; (2) from an HBM MCMC with only photometric information ('F'); and (3) from an HBM with both photometry and clustering ('$F + \delta$'). The four cases studied are shown, one where the prior only has sample variance from the small patch, and three where the prior is further modified to introduce redshift biases. In all cases, the HBM remarkably improves both the bias and shape of the posterior, and the best results are found with the addition of clustering information.

metrics. In this case, we find a redshift bias of $\Delta z = (-3.4 \pm 0.3) \times 10^{-3}$. This result shows that, when clustering information is used in the HBM, photometric redshift estimates can be used instead of spectroscopic measurements, even if such photo-z estimates are imprecise and are systematically biased.

7 DISCUSSION

Fig. 11 presents a visual comparison of the two performance metrics (Δz and $D_{\rm KL}$) obtained with three different inferences: (spectroscopic) prior from a small patch on the sky; the F HBM with photometric information on the full sample; and the $F + \delta$ HBM including photometric information and clustering against a tracer population. In the first case ('Sample Var.' in the plot), where the prior has no biases but just sample variance, the F and $F + \delta$ HBM methods show comparable results in terms of the mean redshift bias, but the clustering method performs better in recovering the shape of the redshift distribution (lower $D_{\rm KL}$ metric). In the other three cases, where biases are introduced in the prior, the HBM method with clustering always performs better in both metrics. Remarkably, for that method, the mean of the redshift distribution is always recovered with a precision of around 3×10^{-3} or better, even when the redshift biases in the prior are larger than 5×10^{-2}. That is a very important result since accurate characterization of the mean of redshift distributions is critical to cosmological analyses of weak gravitational lensing in imaging surveys. Furthermore, the addition of clustering in the method always improves the reconstructed shape of the redshift distribution (lower $D_{\rm KL}$), which can also be very important for cosmology analyses: mischaracterization of the width or tails of a redshift distribution can be a source of systematic error for both weak lensing and galaxy clustering studies.

Our results demonstrate the robustness of this method to several types of biases in the prior, chosen to mimic known shortcomings in real calibration samples. There is an ongoing discussion among the imaging surveys community about the reliability of different redshift samples and how biases in them are propagating into cosmological analyses and creating artificial tension with other cosmological probes (Troxel et al. 2018b; Asgari et al. 2019; Joudaki et al. 2019; Wright et al. 2020). Some groups have relied on spectroscopic samples for their redshift calibration while others have used high-quality, many-band photometric redshifts instead. Spectroscopic samples provide accurate redshift information but can suffer from selection effects and efficiency problems, while high-quality photometric redshifts can have significant biases, especially at high redshift. We have demonstrated how the $F + \delta$ HBM method is robust to any of these effects, providing a rigorous way to propagate

known priors into the posterior, as well as letting the clustering information overcome the priors and their potential inaccuracies.

The success of this method in estimating redshift distributions to the accuracy needed for large cosmological surveys will still depend on the details of the survey. For reference, weak lensing tomography in Y1 LSST and *Euclid* will require a systematic uncertainty in the mean redshift of the redshift distribution below $<0.002(1+z)$ (Laureijs et al. 2011; The LSST Dark Energy Science Collaboration 2018). It is useful here to discuss how the application of this method to real data might differ from the simulations in this paper. First, in this work we have limited the redshift range of interest to be $0.2 < z < 1.2$, while in reality we will need to consider a larger redshift range (Wright et al. 2020). It could also be possible to consider an additional tracer sample at high redshifts, e.g. an emission line galaxy (ELG) sample or a quasar sample. Secondly, the tracer sample used in this work is idealized in that it spans the entire redshift range of interest, and that we have assumed true redshifts for their galaxies. The latter assumption is not a problem as LRG samples have typical redshift errors of $\sigma_z \sim 0.02$ (Rozo et al. 2016), smaller than the redshift bin size chosen in our work ($\sim$0.05). Having a tracer sample not spanning the entire redshift range will reduce the constraining power of the method at the redshifts where we do not have tracers, but it will not result in any additional redshift bias, as demonstrated in SB19. Thirdly, the photometric noise likelihood function $p(F|t, \theta)$ has been determined comparing the truth to 'observed' values in the full simulated target population, whereas in real data this function might be determined from injection simulations with a smaller number of realizations. Real data might therefore have weaker F-only reconstruction from added shot noise in $p(F|t, \theta)$, which would probably increase the degree of improvement that clustering information would yield. Fourthly, we did not include magnification effects in this work. Density fluctuations in foreground galaxies lens the images of background sources, locally stretching them. The increase of area reduces the galaxy density, but at the same time it increases the apparent flux of individual galaxies, making it more likely to detect intrinsically fainter galaxies. The net effect depends on the slope of the luminosity function of the target and tracer populations, and introduces density fluctuations unrelated to their physical overlap. The effect is typically smaller than the clustering signal, but can become significant at the tails of the recovered redshift distribution (e.g. Gatti et al. 2018) and also impact the recovered mean redshift of the redshift distribution. Therefore, lensing magnification might complicate our density field biasing relation, and we defer any extensive investigation to future work. Finally, one other difference is the area used in the application of the method. In this paper, we have used a sample of 1000 deg^2 of sky, while in the application of the method to data we can expect larger areas (e.g. 5000 deg^2 in DES). A larger area overlap between the target and the tracer population will increase the constraining power of the method in the redshift range of overlap between these two samples, driving the biases of the $F + \delta$ HBM to lower levels than in our simulations.

Finally, we discuss the details of our implementation and the corresponding computational needs. For the run on the simulation, we define a total of 20 redshift bins equally spaced in comoving distance between $z \in [0.2, 1.2]$, as well as 1024 phenotypes defined with an SOM from a 32×32 grid, so f_{zt} has a total of 20 480 free parameters. Then, we have 100 free parameters in a biasing function with five parameters per redshift bin. And furthermore, for each target galaxy i, we have z_i and t_i as parameters, which amounts to $2 \times 3 \times 10^6$ free parameters. To save memory and improve speed, we do not save the individual z, t pairs for each target galaxy at every MCMC sample – the individual (z_i, t_i) samples are aggregated into the number counts N_{zt} necessary for the Gibbs sampling of f_{zt}. We parallelize the sampling of the individual z, t of each galaxy in 334 chunks defined by healpy pixels of NSIDE $= 2^5$, and we parallelize the MCMC chain for the biasing parameters by assigning each redshift bin to its own thread. On average, a full iteration of the chain which samples all parameters using the three Gibbs intermediate steps takes 9 s using 334 parallel jobs, which gives about 400 iterations per hour. The method can be parellelized further for more speed, as that step is the limiting factor. Overall, the Gibbs sampling scheme is simple but has the drawback of yielding long correlation lengths, so that more iterations are needed to get a given number of independent samples. A Hamiltonian Monte Carlo (HMC) implementation is possible, and would yield practically independent samples which would result in a speed up of the method. This HMC implementation may be needed to make the method scalable for next-generation surveys such as LSST.

8 SUMMARY AND CONCLUSIONS

SB19 presented a hierarchical Bayesian model which can naturally combine the three main sources of information for estimating the redshift probability distributions of galaxies and samples of galaxies in a wide-field survey. These three main sources of information are: prior information, which comes from a subset of galaxies with well measured photometric and (typically) spectroscopic properties; broad-band photometry for the galaxies in the wide-field sample; and the clustering of such galaxies against a tracer population with precise and accurate redshift estimates. All these sources of information have been used separately in the past, but this is the first method to combine them in a unified and consistent way. In SB19, the main features and potential advantages of the method were demonstrated on a simple set of simulations, but the actual capabilities of it were not assessed, as they depend upon some important pieces that are needed for its application to real data, like realistic clustering properties and the marginalization over biasing functions in the usage of that clustering information.

In this work, we have expanded the HBM approach of SB19 to include the additional methods needed for its application to the analysis of galaxy survey data. The HBM assumes that the galaxies come from a Poisson sampling of an underlying density field; in this work, we characterize this field as a kernel density estimator $\hat{\delta}(\theta)$ applied to a tracer galaxy population with known redshifts, then modified by some parametric biasing function $\mathcal{B}(\hat{\delta}, b)$. We have detailed here how such a biasing function can be constructed, with appropriate freedom to vary with redshift, and how we can sample and marginalize over it using prior information from spectroscopic information over a limited area of the sky.

Moving beyond the simplistic simulations in SB19, we have now tested the methodology on the public MICE2 simulation, a mock galaxy catalogue created from a light-cone of a dark-matter-only N-body simulation with $\approx$200 million galaxies over an octant of the sky. This simulation features realistic galaxy clustering and galaxy properties, and this allows us to work in a scheme where we can fully employ the phenotype approach proposed in SB19. Under that approach, we assume we have a sample with deep photometry and extra bands to define galaxy phenotypes, and a wide sample with noisier photometry and only a subset of optical bands as observations. We use two SOMs to characterize the properties of these samples, and we use galaxies with best-matching cells in both SOMs to accurately calibrate the likelihood probability that relates wide-field observations and phenotypes, as we would do in real data.

In applying the method to a tomographic bin defined in the simulation, we always assume there is a small region of the sky (of about 3 $\deg^2$) for which the galaxy properties, phenotype, and redshift, are well known. We use this set of galaxies as a prior, both for the phenotype and redshift probability distribution and for the biasing function needed for the addition of clustering information from a tracer population. With this setup, we apply the methodology under different cases, comparing the results obtained with and without clustering information in the method and those from just the prior information. As metrics, we use the difference in the mean of the derived and true redshift distributions for the sample, which is arguably the most important quantity for weak lensing analyses, as well as the Kullback–Leibler divergence, which measures the differences in the shapes of the true and recovered redshift distributions.

When the prior comes with perfect knowledge of a small patch of sky, i.e. unbiased but with sample variance, the HBM method both with and without clustering information perform similarly well in terms of the mean redshift of the population. This is expected, also consistent with SB19, as sample variance mostly changes the phenotype distribution, and that can be recovered in the HBM without the need of clustering information. The shape of the redshift distribution is, however, better recovered when using clustering information.

Clustering information is shown to be very powerful when the redshift information from the small area is biased or incomplete, as is happening in real spectroscopic samples. In such tests, the addition of clustering to the HBM improves both the mean and the shape of redshift distributions. We have demonstrated this with simulations of a gentle coherent bias in the redshift assignments, in the case of uncompensated high-redshift incompleteness of spectroscopy, and in a case with spuriously broad spectroscopic assignments (as one might expect from photometric reference samples). In these cases the HBM with clustering reduced the bias in the sample's mean redshift by a factor of 2–10 compared to photometry-only constraints. The error in the full redshift distribution $n(z)$ is reduced by factors of 3–20, as measured by the Kullback–Leibler divergence.

One shortcoming of the current implementation of the HBM is that we do not account for correlations between the redshifts and phenotypes of the spectroscopic sample induced by large-scale structures in the spectroscopic sample patch, what is also known as sample variance (Cunha et al. 2012). Recently, Sánchez et al. (2020) have developed a way to add sample variance uncertainties into the redshift prior sampling, which will enable the HBM method in this work to also account for that source of uncertainty in future applications. Alternatively, future renditions of the HBM could also be able to treat the density fluctuation field as a stochastic variable and hence include the LSS correlations in a natural way.

The tests performed in this work provide demonstration that the method depicted in SB19, with the generalizations presented here, can be used in realistic conditions, and it can still be very powerful at resolving biases that are potentially present in prior samples, even after marginalizing over biasing functions in the addition of clustering information. The method does not guarantee an unbiased posterior, but it uses all the information at hand to reduce prior biases, and even in all tests performed here, some of which are extreme cases of biased priors, the final biases in the posterior are of the order of 10^{-3} in the mean of redshift distributions. Obtaining a trustworthy $n(z)$ estimation of this accuracy in real survey data would be a milestone for the control of redshift systematic uncertainties in future weak lensing and galaxy clustering analyses.

ACKNOWLEDGEMENTS

The authors thank Boris Leistedt, Daniel Gruen, Justin Myles, and Alexandra Amon for helpful conversations about this topic. AA and EG were supported by MINECO grants CSD2007-00060 and AYA2015-71825, LACEGAL Marie Sklodowska-Curie grant 734374 with ERDF funds from the European Union Horizon 2020 Programme. CS and GMB were supported by grants AST-1615555 from the US National Science Foundation, and DE-SC0007901 from the US Department of Energy. IEEC is partially funded by the CERCA program of the Generalitat de Catalunya. The MICE simulations have been developed at the MareNostrum supercomputer (BSC-CNS) thanks to grants AECT-2006-2-0011 through AECT-2015-1-0013. Data products have been stored at the Port d'Informacio Cientifica (PIC), and distributed through the CosmoHub webportal (cosmohub.pic.es).

DATA AVAILABILITY

The data underlying this article are available at

(i) MICE2 simulations: publicly available at https://cosmohub.pic.es/home under `Catalogs` labelled as `MICECAT`, version 2.

REFERENCES

Arnouts S. et al., 2002, MNRAS, 329, 355
Asgari M. et al., 2019, A&A, 624, A134
Benitez N., 2000, ApJ, 536, 571
Benjamin J. et al., 2013, MNRAS, 431, 1547
Bolzonella M., Miralles J., Pell R., 2000, Astron. Astrophys., 492, 476
Bonnett C., 2015, MNRAS, 449, 1043
Bonnett C. et al., 2016, Phys. Rev. D, 94, 042005
Brammer G. B., van Dokkum P. G., Coppi P., 2008, Astrophys. J., 686, 1503
Buchs R. et al., 2019, MNRAS, 489, 820
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Carretero J., Castander F. J., Gaztañaga E., Crocce M., Fosalba P., 2015, MNRAS, 447, 646
Carretero J. et al. 2017, Proceedings of The European Physical Society Conference on High Energy Physics, EPS-HEP2017: https://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=314, p. 488
Cautun M. C., van de Weygaert R., 2011, Astrophysics Source Code Library, preprint (arXiv:1105.0370)
Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, Astron. J., 132, 926
Colless M. et al., 2001, Mon. Not. R. Astron. Soc., 328, 1039
Collister A. A., Lahav O., 2004, Publ. Astron. Soc. Pacific, 116, 345
Cox D. R., 1955, J. R. Stat. Soc. B, 17, 129
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
Dahlen T. et al., 2013, ApJ, 775, 93
Darvish B., Mobasher B., Sobral D., Scoville N., Aragon-Calvo M., 2015, ApJ, 805, 121
Davis C. et al., 2017, preprint (arXiv:1710.02517)
Dawson K. S. et al., 2013, Astron. J., 145, 10
de Jong J. T. A., Kuijken K., Applegate D., Begeman K., Belikov A., al E., 2013, ESO Messenger, 154, 44
DES Collaboration, 2018, Phys. Rev. D, 98, 043526
DESI Collaboration, 2016, preprint (arXiv:1611.00036)
Drinkwater M. J. et al., 2010, MNRAS, 401, 1429
Elvin-Poole J. et al., 2018, Phys. Rev. D, 98, 042006
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
Gatti M. et al., 2018, MNRAS, 477, 1664
Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, Astrophys. J., 715, 823
Górski K. M., Hivon E., Banday A. J., Wand elt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Hikage C. et al., 2019, Publ. Astron. Soc. Japan, 71, 43
Hildebrandt H. et al., 2010, A&A, 523, A31
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
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, Astron. Astrophys., 457, 841
Ilbert O. et al., 2009, ApJ, 690, 1236
Ivezić Ž. et al., 2019, ApJ, 873, 111
Joudaki S. et al., 2017a, MNRAS, 465, 2033
Joudaki S. et al., 2017b, MNRAS, 471, 1259
Joudaki S. et al., 2019, A&A, 638, L1
Kaiser N., Tonry J. L., Luppino G. A., 2000, Publ. Astron. Soc. Pacific, 112, 768
Krause E. et al., 2017, preprint (arXiv:1706.09359)
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Le Fèvre O. et al., 2005, A&A, 439, 845
Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258
Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118
LSST Dark Energy Science Collaboration, 2012, preprint (arXiv:1211.0310)
Ma Z., Bernstein G., 2008, ApJ, 682, 39
Masters D. et al., 2015, ApJ, 813, 53
Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, preprint (arXiv:1303.4722)
Miyazaki S. et al., 2012, Ground-based and Airborne Instrumentation for Astronomy IV. Proceedings of the SPIE, 8446, 84460Z
Newman J. a., 2008, ApJ, 684, 88
Planck Collaboration VI, 2018, preprint (arXiv:1807.06209)
Rozo E. et al., 2016, MNRAS, 461, 1431
Sánchez C., Bernstein G. M., 2019, MNRAS, 483, 2801 (SB19)
Sánchez C. et al., 2014, MNRAS, 445, 1482
Sánchez C., Raveri M., Alarcon A., Bernstein G. M., 2020, preprint (arXiv:2004.09542)
Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, MNRAS, 431, 3307
Speagle J. S. et al., 2019, MNRAS, 490, 5658
Suchyta E. et al., 2016, MNRAS, 457, 786
The LSST Dark Energy Science Collaboration, 2018, preprint (arXiv:1809.01669)
Troxel M. A. et al., 2018a, Phys. Rev. D, 98, 043528
Troxel M. A. et al., 2018b, MNRAS, 479, 4998
Vakili M. et al., 2019, MNRAS, 487, 3715
Wright A. H., Hildebrandt H., van den Busch J. L., Heymans C., 2020, A&A, 637, A100
York D. G. et al., 2000, Astron. J., 120, 1579
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Softw., 4, 1298

APPENDIX A: PRIOR AND POSTERIOR OF KDE BIASING FUNCTIONS

In this work, we use a galaxy tracer population to estimate the density field from which target galaxies are drawn from, using a kernel density estimation (Section 4). However, as tracer and target populations can be different, and because of effects such as shot noise in the tracer population, we need a mapping function that relates the field estimated from tracers and the field from which target galaxies have been drawn from. As outlined in Section 5, the biasing functions need to be sampled and marginalized over in the Gibbs process of the HBM. For that sampling, we use information from a small set of galaxies with true redshift information as a prior for the Gibbs sampling. In this work, in order to avoid being limited by sample variance in the estimation of this prior for biasing functions, we assume such functions have a smooth redshift dependence and we join four redshift bins from that prior sample at the time of running the corresponding MCMC chains. Then, we effectively use the same prior for four adjacent redshift bins in the Gibbs sampling process. Other than reducing sample variance, this procedure also makes the prior more robust to biases in the redshift estimation of the galaxies used in the prior. Fig. A1 shows an example of the prior and posterior of such biasing functions, parametrized as in equation (17), in one random redshift bin. One can see the posterior given by the HBM chain to be much tighter than the prior, showing how the HBM method is self-calibrating the biasing functions from the wide data in the simulation. The figure here shows one example redshift bin, but this is generally true for all bins considered in this work.

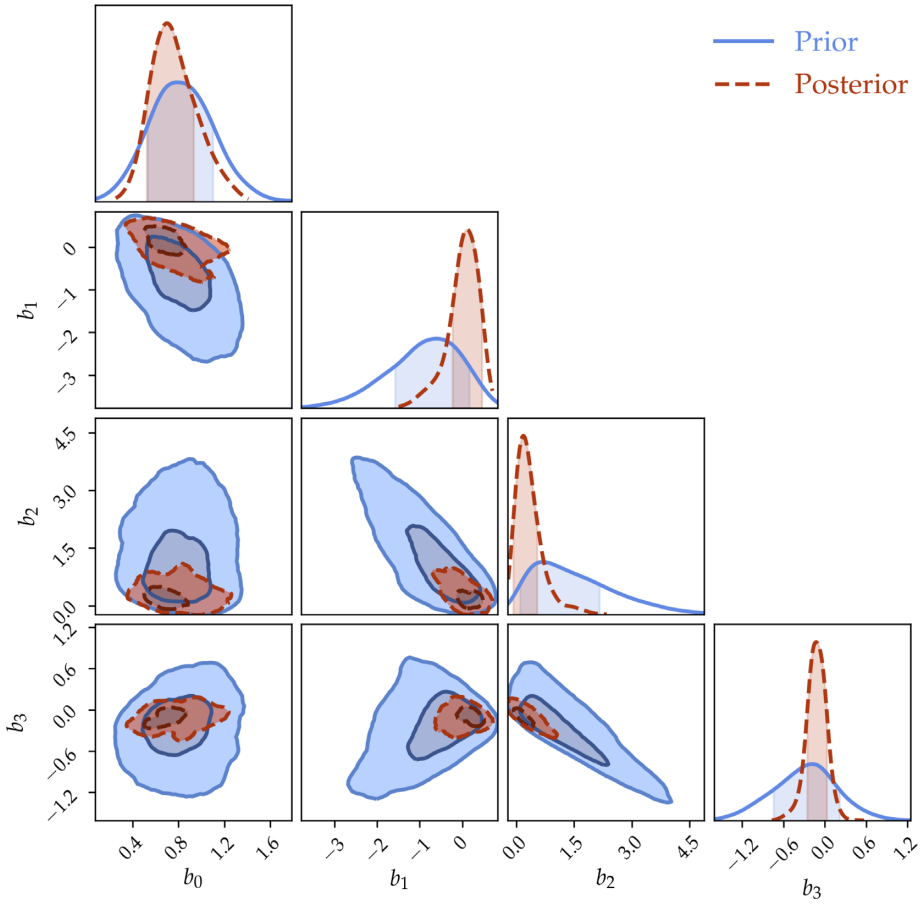

Figure A1. Prior and posterior of the biasing functions, parametrized as in equation (17), in one random redshift bin (bin 4). The posterior appears to be tighter than the prior, showing how the HBM method uses information from the entire sample to characterize these mapping functions.

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