



Redshift inference from the combination of galaxy colours and clustering in a hierarchical Bayesian model

Carles Sánchez  and Gary M. Bernstein

Department of Physics and Astronomy, University of Pennsylvania, 209 S. 33rd St, Philadelphia, PA 19104, USA

Accepted 2018 November 20. Received 2018 November 8; in original form 2018 August 2

ABSTRACT

Powerful current and future cosmological constraints using high-precision measurements of the large-scale structure of galaxies and its weak gravitational lensing effects rely on accurate characterization of the redshift distributions of the galaxy samples using only broad-band imaging. We present a framework for constraining both the redshift probability distributions of galaxy populations and the redshifts of their individual members. We use a hierarchical Bayesian model (HBM) which provides full posterior distributions on those redshift probability distributions, and, for the first time, we show how to combine survey photometry of single galaxies and the information contained in the galaxy clustering against a well-characterized tracer population in a robust way. One critical approximation turns the HBM into a system amenable to efficient Gibbs sampling. We show that in the absence of photometric information, this method reduces to commonly used clustering redshift estimators. Using a simple model system, we show how the incorporation of clustering information with photo- z 's tightens redshift posteriors, and can overcome biases or gaps in the coverage of a spectroscopic prior. The method enables the full propagation of redshift uncertainties into cosmological analyses, and uses all the information at hand to reduce those uncertainties and associated potential biases.

Key words: cosmology: observations – galaxy: surveys – photometric redshifts.

1 INTRODUCTION

Large-galaxy surveys constitute a powerful probe for testing cosmological models through the information they provide about the large-scale structure of the Universe. There exist two main categories of surveys. 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), or *Dark Energy Spectroscopic Instrument* (DESI Collaboration et al. 2016) supply 3D information about the galaxy distribution, but they are expensive in time and resources. Alternatively, 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) are more efficient per galaxy, and enable weak gravitational lensing measurements via galaxy shapes — but

do not provide a complete 3D picture of the visible Universe, as redshifts are estimated using only broad-band spectra of galaxies.

A critical task for cosmological inference using imaging surveys is to establish the redshift distribution $n(z) = dN/dz dA$ of a sample of sources passing some selection criterion s , since errors in the characterization of such distributions can 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; Hoyle et al. 2018). In some cases, the posterior distributions $p(z_i)$ of the redshifts of individual sources are of use as well. Two forms of information commonly form the basis of these inferences. On one hand, *photometric* redshifts compare a set of observed fluxes (or colours or 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 from different techniques, broadly classified as 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)), and training methods (e.g. ANNz, Collister & Lahav (2004); ArborZ, Gerdes et al. (2010); TPZ, Carrasco Kind & Brunner (2013); SkyNet, Bonnett (2015)). Comparisons of different methods have been tried in both simulated and real data (Hildebrandt et al. 2010; Dahlen et al. 2013; Sánchez et al. 2014). On the other hand, with the advent of large-area sur-

* E-mail: carles.sanchez.alonso@gmail.com (CS);
garyb@physics.upenn.edu (GMB)

veys, it has become practical to constrain $n(z)$ using *clustering* information based on the observed sky positions θ_i of the sources. For the latter, most applications are based on two-point statistics of the source population with reference to the positions of a tracer population of galaxies having secure a priori redshift assignments (Newman 2008; Ménard et al. 2013; Schmidt et al. 2013). In addition, there is typically prior information on the redshift distribution of the population coming from a subset of galaxies in the survey, spanning a small area of the sky, for which *spectroscopic* or high-quality photometric redshifts are available. Attempts to use both photometric and clustering constraints on the same survey data in the presence of prior information have been made by the *KiDS* (Hildebrandt et al. 2017) and *DES* (Gatti et al. 2018; Hoyle et al. 2018; Davis et al. 2017; DES Collaboration et al. 2017) surveys. The two methods have been used together either by means of basic visual cross-checks on the two independently derived $n(z)$'s, or to provide cross-checks and joint constraints on some single summary statistic of $n(z)$, such as its mean.

In this work, we develop and test a method to produce a single inference for $n(z)$ (and the z_i) that intrinsically combines prior, photometric, and clustering information, and includes the capability to produce samples of $n(z)$ drawn from the posterior distribution constrained by all the sources of information. This method is robust to some of the practical difficulties of clustering or photometric redshifts, such as the absence of tracers over some portion of the full angular or spectral extent of the survey, or to variation of the selection function or measurement noise across the survey footprint. This is achieved by extending the Bayesian hierarchical method of Leistedt, Mortlock & Peiris (2016, LMP) to include the presence of fluctuations $\delta_z(\theta)$ in the density of sources at redshift z and sky position θ . Other changes with respect to the LMP method include the addition of an informative prior, and the separation of the photometric noise part of the problem from the characterization of the so-called *colour–redshift* relation, which we consider independently. We will demonstrate that our method reduces to the results of the standard two-point clustering method in the limit where the prior and the F_i are uninformative and the clustering is weak.

The combination of clustering, photometric, and spectroscopic information into a single $n(z)$ inference is advantageous for two main reasons. First, it has not been clear to date how one should combine or even check consistency between independent inferences, since each is predicting a function $n(z)$. Second, each method on its own can have failures that render it impossible to generate a usable inference: for example, clustering measures can lack a tracer population at some z ranges; spectroscopic surveys are usually highly incomplete in some way which is biased with redshift; and photometric inferences are often biased by errors in assumed ‘truth’ colours. In such circumstances, an inference based on a single technique can be so uncertain as to be worthless for precision cosmology. But a combined inference could allow each method to resolve the degeneracies or ameliorate the intrinsic biases of another, and be the only route to a viable estimate.

In Section 2, we derive the posterior probability for redshift information when conditioned on observed galaxy features and positions. Section 3 derives the optimal kernel density estimator to be applied to a tracer galaxy population. In Section 4, we outline a Gibbs sampler for the system, and a straightforward method for inclusion of prior constraints on $n(z)$ derived from spectroscopic samples. We then implement the method on a simple simulation in Section 5, demonstrating the ability of the combined inference to robustly reduce noise and biases present in the spectroscopic

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 in the sky
s	indicator of successful detection/selection
\mathcal{L}_{it}	probability of measuring galaxy i with F_i given t
$\mathbf{F}, \mathbf{t}, \mathbf{z}, \boldsymbol{\theta}$	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
$\delta_z(\theta)$	density fluctuation at a given z and θ
π_δ	density fluctuation field hyperparameters
$\boldsymbol{\delta}$	set of $\delta_z(\theta)$ for all redshifts and positions
b_z^t	linear galaxy bias for type t at redshift z
\mathbf{b}	set of b_z^t for all types and redshifts
f_{zt}	joint type and redshift probability $p(z, t)$
\mathbf{f}	set of f_{zt} for all types and redshifts
N_{zt}	number of sources assigned to redshift z and type t
\mathbf{N}	set of N_{zt} for all redshifts and types
M_{zt}	number of sources in the prior at redshift z and type t
\mathbf{M}	set of M_{zt} for all redshifts and types
Δz	difference between the means of estimated and true $n(z)$'s

or photometric information alone. The Appendix shows that our method reproduces the results of standard clustering- z techniques in the absence of photometric information (under some applicable approximations).

2 PROBABILITIES FOR REDSHIFT DISTRIBUTIONS

Our approach closely parallels that of LMP, with the colours of individual galaxies i are seen as being drawn from a pool of possible types t_i , redshifts z_i , and angular positions θ_i with some latent intrinsic mean density $n(t, z)$ on the sky, with observations then yielding a noisy set of observable features which we denote as F_i . The critical extension to the LMP model is to consider that the galaxy densities at redshift z vary by some factor $1 + \delta_z(\theta)$. The tracer population used for correlation redshifts is drawn from the same latent density distribution (up to some bias factor). Note also an important difference from the LMP model and notation, whereby we define galaxy ‘types’ by observed properties rather than rest-frame properties. This *phenotype* approach will allow us to decouple the measurement–noise part of the problem from the colour–redshift relation. Our notation will be that the vector quantities \mathbf{F} , \mathbf{t} , \mathbf{z} , and $\boldsymbol{\theta}$ denote the full set of properties of all selected galaxies, i.e. $\mathbf{F} = \{F_1, F_2, \dots, F_N\}$ (a summary of all the notation can be found in Table 1).

2.1 Generative model

Our fundamental assumption will be that the galaxies are drawn from a *Cox process* (Cox 1955) or *doubly stochastic Poisson process*, whereby a point set is Poisson-sampled from a latent, stochastic density field. We simplify the problem to consider that the redshift z is transformed to an integer indexing a set of finite-width

redshift bins, and each bin has an independent density fluctuation field $\delta_z(\theta)$, with $\langle \delta_{z_i}(\theta) \delta_{z_j}(\theta) \rangle = 0$ for $z_i \neq z_j$.

We first consider a simple case in which the sky is populated with galaxies that have identical appearance on the sky (i.e. constant F) and a mean density per unit solid angle of n . The galaxies have intrinsic redshift distribution specified by $f_z = p(z)$, with $\sum_z f_z = 1$. The galaxies' spatial distribution has linear bias b_z with respect to the underlying density fluctuation δ_z . We also assume there is some selection function s with the probability of a galaxy being selected perhaps depending on sky position, specified as a selection or window function $p(s|\theta)$. We will always assume that we know nothing about the non-selected galaxies, not even that they exist; the observed data are the positions θ and features F of the selected sources. The selected galaxies can be considered now as being a Poisson sampling of a density field

$$\rho(z, \theta | n, f, b, \delta) = n f_z [1 + b_z \delta_z(\theta)] p(s|\theta). \quad (1)$$

Note that the last two terms describe the spatial variation of expected detection rate due to density fluctuations and variable observing conditions, respectively. Then, the probability of selecting a set of galaxies $i \in \{1 \dots N\}$ at positions θ and redshifts z takes the standard Poisson form:

$$\begin{aligned} p(z, \theta | n, f, b, \delta) &= \exp \left\{ - \int d^2\theta \sum_z \rho(z, \theta) \right\} \prod_{i=1}^N \rho(z_i, \theta_i) \\ &= \exp \left\{ -n \int d^2\theta p(s|\theta) \sum_z f_z [1 + b_z \delta_z(\theta)] \right\} \\ &\quad \times \prod_{i=1}^N p(s|\theta_i) \times n f_{z_i} [1 + b_{z_i} \delta_{z_i}(\theta_i)]. \end{aligned} \quad (2)$$

Note that the normalization part of the Poisson likelihood in the exponent contains a sum over z rather than an integral, because we have made z a discrete variable.

Next, we generalize equation (2) to allow for variety among galaxies. We define a *phenotype* of galaxy to specify its noiseless, observed appearance. In other words, all galaxies of phenotype t observed in the same conditions are assumed to have the same selection function $p(s|t, \theta)$ and same probability $p(F, s|t, \theta)$ of being selected and measured to have image features F . Following biological nomenclature, the phenotype strictly describes the manifestation of the galaxy in the image. Galaxies of identical phenotype t could live at distinct redshifts. The typical astronomical definition of galaxy 'type' specifies rest-frame properties, but a fixed type may have observed properties that are redshift dependent. Our 'phenotype' reverses these, and will allow us to use well-observed sources as a library of phenotypes even when knowledge of these sources' redshifts is imprecise. Such galaxy-type definition allows us to decouple the measurement–noise part of the problem from the colour–redshift relation. The latter will be informed by prior spectroscopic observations or models, and by the clustering information.

We will assume that we have a finite set of phenotypes indexed by integer t . Each has a mean sky density of $n^t = n f_t$, 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. We have the constraint $\sum_t f_t = 1$. We write $p(z|t) = f_z^t$ for the redshift distribution of type t , and we will also denote

$$f_{tz} \equiv p(z, t) = p(z|t)p(t) = f_z^t f_t. \quad (3)$$

Letting the bias also depend on phenotype as b_z^t , we can generalize equation (1) to the expected mean density for each phenotype t :

$$\rho(z, \theta, t | n, f, b, \delta) = n f_{tz} [1 + b_z^t \delta_z(\theta)] p(s|t, \theta). \quad (4)$$

Knowing the survey noise properties and the intrinsic (noiseless) appearance of phenotype t , we can also 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 . This function will be independent of z since the phenotype's observables are, by definition, independent of z . Thus, for each observed galaxy i and phenotype t , we can assign a feature/selection likelihood

$$\mathcal{L}_{it} \equiv p(F_i, s|t_i, \theta_i). \quad (5)$$

This function depends on the quality of the observations in direction θ_i and the selection/measurement algorithms. We will assume that this is known a priori, e.g. by the result of injecting artificial copies of the phenotype into the images (Suchyta et al. 2016).

We can now extend the Poisson probability (2) to include a variety of phenotypes and the probability observing the galaxy features:

$$\begin{aligned} p(F, \theta, t, z | n, f, b, \delta) &= \exp \left[-n \sum_t f_t A^t(f^t, b^t, \delta) \right] \\ &\quad \times \prod_i \mathcal{L}_{it} n f_{t_i z_i} [1 + b_{z_i}^{t_i} \delta_{z_i}(\theta_i)]. \end{aligned}$$

The exponentiated quantity is, as required for Poisson distributions, the expected number of detections $\langle N \rangle$ for the entire sample. We express this in terms of an effective survey area for each phenotype t :

$$\begin{aligned} A^t(f, b, \delta) &\equiv \sum_z \int d^2\theta p(s|t, \theta) f_z^t [1 + b_z^t \delta_z(\theta)] \\ &= \int d^2\theta p(s|t, \theta) \left[1 + \sum_z f_z^t b_z^t \delta_z(\theta) \right] \end{aligned} \quad (7)$$

$$\approx \int d^2\theta p(s|t, \theta) \quad (8)$$

which can also be determined from knowledge of the survey properties. In the last line, we make the approximation that the density fluctuation integrated over the mask is small, $\int d^2\theta p(s|\theta) \delta_z(\theta) = 0$. In this case, the A^t are known constants of the survey.

The full generative model for the data must also specify the process $p(\delta|\pi_\delta)$ generating the stochastic density fluctuation fields given hyperparameters π_δ , – e.g. a log-normal distribution with π_δ specifying the power spectrum. We also require priors $p(b)$ and $p(n)$, plus any prior information on $p(f)$ aside from the constraint that $\sum_t f_t = 1$. Fig. 1 is a directed acyclic graph of the model for generating the galaxy data.

2.2 Redshift inference

The principal quantity of interest is the underlying redshift distribution

$$n(z) = n \sum_t f_{tz}. \quad (9)$$

The equation can easily be altered to yield the redshift distribution of chosen subsets of the phenotypes. We will usually be concerned with the shape, not the normalization, of $n(z)$, so will pay little attention to n and focus on the fractions f . We will also be interested

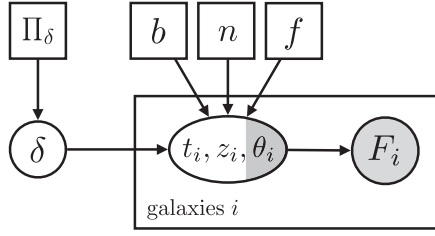


Figure 1. A directed acyclic graph of the hierarchical model for the observable positions θ and features F in the galaxy catalogue. Observed quantities are shaded: only the positions θ_i and image features (e.g. fluxes) F_i of individual sources, not types or redshifts. Squares hold model parameters, circles are stochastic quantities.

in many cases in the probabilities of the individual redshifts z , and to enable a Gibbs sampling scheme we will keep b and t as conditional variables. Using Bayes' theorem, the posterior joint probability of these variables of interest is

$$p(\mathbf{f}, \mathbf{z}, \mathbf{b}, \mathbf{t} | \mathbf{F}, \boldsymbol{\theta}, \pi_\delta) \propto \int d\mathbf{n} d\mathbf{p} p(\mathbf{F}, \boldsymbol{\theta}, \mathbf{t}, \mathbf{z} | \mathbf{n}, \mathbf{f}, \mathbf{b}, \delta) \times p(\delta | \pi_\delta) p(\mathbf{n}) p(\mathbf{f}) p(\mathbf{b}). \quad (10)$$

The first term under the integral is given by equation (6). We take the standard scale-free Jeffreys (logarithmic) prior $p(n)$ on the overall density n . Marginalization over n then yields:

$$p(\mathbf{f}, \mathbf{z}, \mathbf{t}, \mathbf{b} | \mathbf{F}, \boldsymbol{\theta}, \pi_\delta) \propto p(\mathbf{f}) A(\mathbf{f})^{-N} p(\mathbf{b}) \int d\delta p(\delta, \pi_\delta) \quad (11)$$

$$\times \prod_i \mathcal{L}_{it} f_{iz_i} [1 + b_{zi}^t \delta_{zi}(\theta_i)]$$

$$A(\mathbf{f}) \equiv \sum_{zt} f_{iz} A^t. \quad (12)$$

The total effective area of the survey for source detection is $A(\mathbf{f})$, such that $\langle N \rangle = n A(\mathbf{f})$.

The required marginalization of the Cox distribution over the stochastic field(s) in equation (11) is known to be analytically intractable for all but the most trivial processes $p(\delta | \pi_\delta)$. For some processes (including log-normal), there are feasible means to sample over the joint distribution of δ and the other parameters, at which point the marginalization is trivial – an example of log-normal field inference from photo- z data is Jasche & Wandelt (2012). A full treatment of the Cox posterior would enable the clustering redshift inference to make use not only of correlations between a tracer population and the target population, but also the clustering among target populations. We will defer any attempt at sampling δ to further work.

In this paper, we will instead work with the approximation that we can replace the stochastic density fluctuation $\delta_z(\theta)$ with some deterministic estimator $\hat{\delta}_z(\theta)$ of the realization of the density fields in the generative probability of equation (6). In other words, we will replace the $p(\delta | \pi_\delta)$ in equation (11) with a Dirac delta function enforcing $\delta_z = \hat{\delta}_z$, which also renders the hyperparameters π_δ as irrelevant. The estimator $\hat{\delta}_z$ will presumably come from a tracer population; this approximation loses any information that may be present in the clustering amongst the target population. We will refer to this as the ‘deterministic density’ approximation, since it yields a single set of density fields $\hat{\delta}_z(\theta)$ from the observed set of tracers, ignoring the fact that the density fields remain stochastic even after the tracer population is specified. With this simplifying

approximation, the posterior distribution for redshift information in equation (11) becomes

$$p(\mathbf{f}, \mathbf{z}, \mathbf{t}, \mathbf{b} | \mathbf{F}, \boldsymbol{\theta}) \propto p(\mathbf{f}) A(\mathbf{f})^{-N} \prod_i \mathcal{L}_{it} f_{iz_i} (1 + b_{zi}^t \hat{\delta}_{zi}), \quad (13)$$

$$\hat{\delta}_{iz} \equiv \hat{\delta}_z(\theta_i). \quad (14)$$

The roles of the photometric and clustering information are clear and simple in the posterior of equation (13). The photometric information is in \mathcal{L}_{it} , which tells how well galaxy i resembles phenotype t (along with how likely this is to pass selection). We may, for instance, know that our feature vector is equal to some ‘truth’ value F_i associated with phenotype t , with some additive noise determined by the observing conditions in direction θ_i of the survey: $\mathcal{L}_{it} = \mathcal{L}(F_i - F_i | \theta_i)$. Next, is the term f_{iz} , which is an expression of the prior probability that any galaxy is of phenotype t and redshift z . Finally, there is the clustering term, describing the modification of the probability by (our estimator for) the density fluctuation field.

The technique for combined photometric/clustering redshift inference is thus as follows:

(i) Create a library of phenotypes t , which span the range of possible feature vectors and window functions that galaxies can exhibit in the survey. The phenotype library might be developed, for instance, from the galaxies detected in a deep subset of the imaging survey. One possible approach is to discretize the colour space of galaxies using a self-organized map (Masters et al. 2015), in which case the cells of such a map would become our galaxy phenotypes, representing well-localized regions of observed feature space.

(ii) Characterize the window functions $p(s | t, \theta)$. If the selection function is well chosen, then this may be analytically accessible – for example if s is a sharp flux cutoff, and the flux measurement is known to have Gaussian errors, then the selection becomes an error function. Alternatively, one can inject copies of t into the data and empirically determine the selection rate, as in Suchyta et al. (2016). Then, we integrate $p(s | t, \theta)$ over the survey footprint to get the effective areas A^t for each type, and evaluate the $\mathcal{L}_{it} = \mathcal{L}(F_i, s | t, \theta_i)$ for each detection.

(iii) Using the tracer population, develop an overdensity estimator $\hat{\delta}_{iz}$ at each redshift z and the sky position θ_i of each detection.

(iv) Posit a prior $p(b)$ for the clustering biases b_z^t of each phenotype relative to the tracer population. In practice, this prior will need to bound the possible dependence of bias on redshift and type if we want to avoid the bias-density degeneracy inherent to clustering information.

(v) Posit a prior $p(f)$ for the fractions of the galaxy population in each type/redshift combination. This prior would likely limit the z 's available for a given t based on astrophysical knowledge that the colours/features of phenotype t are consistent with a small number of possible redshifts, and/or by knowledge of the redshifts obtained from spectroscopy of some small sample from this phenotype.

(vi) Sample from the posterior distribution of $\mathbf{f}, \mathbf{z}, \mathbf{t}$, and \mathbf{b} defined by equation (13). From these samples, one can construct samples from the posterior distributions of $n(z)$ and/or the individual redshifts z_i as marginalized over type and bias.

One important caveat to the replacement of the Cox process with a deterministic density estimator is that the samples of posterior redshift information that we extract from this process will not include the variance attributable to noise in the density estimator, i.e. the shot noise in the tracer population.

3 DENSITY ESTIMATORS

Clustering redshifts make use of a population of galaxies with reliably known redshifts, either from spectra or from highly accurate photo- z 's. The tracer population must have a known selection function in a given redshift bin so that meaningful clustering statistics can be produced. The proper way to use the tracer galaxies would be to define a phenotype $t = T_z$ for the tracers in each redshift bin z , and include them in the full Cox process sampling. For the tracer phenotypes, we would have $f_{z'}^{T_z} = \delta_K(z, z')$, and could set $b_z^{T_z} = 1$, normalizing the δ_z to the fluctuations of the tracers.

But as noted we will instead pull the tracer galaxies out of the likelihood and use them to produce a deterministic density estimator $\hat{\delta}_z(\theta)$. We can also define our estimator to yield $\langle \hat{\delta}_z \rangle = 0$ when averaged over the survey footprint.

3.1 Optimal kernel density estimator

We investigate a kernel density estimator (KDE) for $\hat{\delta}_z(\theta)$. We will show here that this yields posterior likelihoods based on pair counts. Standard methods for correlation redshifts estimate $n(z)$ using two-point correlation functions, which are also computed from pair counts, but the standard methods never attempt to make point estimates of the fluctuation fields. We will show here and in the Appendix that, in a case where photometric information (\mathbf{F}) is ignored, the use of an optimal KDE in our method yields $n(z)$ estimates that are equivalent to the standard two-point methodologies.

In this section, we will drop the subscript z on the assumption that we will treat distinct redshift bins as having independent mass distributions. With the assumption that the kernel depends only on the distance θ_{xT} between a test point x and a tracer galaxy T , we have

$$\hat{\delta}(x) = \frac{1}{n_T} \sum_T K(\theta_{xT}) - \frac{1}{n_R} \sum_R K(\theta_{xR}) \quad (15)$$

with n_T being the mean density of tracers, and n_R the mean density of a population of unclustered random points having the angular selection function of T . We assume $n_R \gg n_T$ so that the latter term can be considered a deterministic integral over survey area. We assume that the process for δ is stationary and isotropic, with known variance $\sigma^2 = \langle \delta^2 \rangle$ and correlation function $w(\theta_{12}) = \langle \delta(\theta_1)\delta(\theta_2) \rangle$. These conditions are easily met when using cosmological tracer populations.

We seek a kernel that minimizes the error $\langle (\hat{\delta} - \delta)^2 \rangle$ while maintaining no bias, $\langle \hat{\delta} \rangle = \delta$ for any field point. We can construct this minimum-variance estimator straightforwardly under two conditions:

- The field has a function $r(\theta)$ such that $\langle \delta(\theta_1)\delta(\theta_2) \rangle = r(\theta_{12})\delta(\theta_2)$. It is straightforward to show that if there is a function r that satisfies this equation, then it must be $r(\theta) = w(\theta)/\sigma^2 = w(\theta)/w(0)$. For a Gaussian field, the mean $\delta(\theta_1)$ conditioned on $\delta(\theta_2)$ is indeed linear in $\delta(\theta_2)$.

- The noise in the KDE estimate is dominated by Poisson shot noise. This holds in the limits of low tracer density or weak clustering. More generally, the variance of correlation estimators involves sample-variance terms (Bernstein 1994), but we will ignore this.

In the real cosmological applications, the first condition is not necessarily true because the density field is non-Gaussian at the smaller (~ 10 Mpc) scales where most of the kernel's power will be. The Poisson-dominated limit will typically be attained for tracer populations at the relevant scales. Should either assumption fail, it merely

means that the KDE we derive here could be non-optimal; it does not invalidate the approach. Future work will examine these effects with simulations of the real sky.

When these assumptions are true, the ideal kernel will minimize

$$\text{Var}[\hat{\delta}(x)] = \frac{1}{n_T} \int 2\pi\theta d\theta K^2(\theta) \quad (16)$$

subject to the constraint

$$\int 2\pi\theta d\theta K(\theta)r(\theta) = 1, \quad (17)$$

which is optimized by

$$K(\theta) = \frac{r(\theta)}{\int 2\pi\theta' d\theta' r^2(\theta')} = \sigma^2 \frac{w(\theta)}{\int 2\pi\theta' d\theta' w^2(\theta')}. \quad (18)$$

Note that the kernel is unchanged if we rescale the fluctuation field, which rescales w and σ^2 by identical factors. Thus, the kernel choice depends only on the shape, not the amplitude, of $w(\theta)$.

Our density estimator at field point x thus becomes

$$\hat{\delta}(x) = \sigma^2 \frac{\frac{1}{n_T} \sum_T w(\theta_{xT}) - \frac{1}{n_R} \sum_R w(\theta_{xR})}{\frac{1}{n_R} \sum_R w^2(\theta_{xR})}. \quad (19)$$

We caution the reader that the normalization in the denominator of equation (19) yields an unbiased estimator only when the kernel is chosen to match the true $w(\theta)$ for all θ . There are also practical issues to address with this KDE when the tracers are sparse and the value of $\hat{\delta}$ is not small, because the $(1 + b\hat{\delta})$ terms in the likelihood can yield non-physical (and non-mathematical!) probabilities ≤ 0 . These and other complications of the application of this method to real data will be explored in future work.

3.2 Missing tracers

One practical difficulty with clustering- z estimation has been that tracers are not available over the full-redshift range of the targets, which leaves $n(z)$ indeterminate in these redshift gaps, and makes it impossible to apply a normalization condition to the derived $n(z)$.

In the HBM method herein, a region of redshift, or of the sky, where we have no tracers is equivalent to having to marginalize equation (11) over the unknown values of $\delta_z(\theta)$. For a single source galaxy, this marginalization is equivalent to setting $\delta = 0$, since $\langle \delta \rangle = 0$ by construction. We therefore can proceed by simply setting $\delta_{iz} = 0$ in equation (13) for the posterior within any angular or redshift gaps in the tracer coverage. When there are multiple galaxies in such gaps, marginalizing (11) over δ is not necessarily equivalent to setting $\delta = 0$ because of the influence of the prior $p(\delta, \pi_\delta)$ on correlations between different δ_{iz} . But we are already ignoring such effects by using the estimator $\hat{\delta}$ in place of marginalization over the δ process.

In short, where we do not have tracer information, we do not try to use it. In these regions, the posterior probability simply reverts to the photometric-only form. Thus, while our HBM formalism does not eliminate degeneracies in pure clustering- z inference from gaps in the tracers, it does admit a combination with photo- z in a principled fashion in which the photometric data resolves the clustering degeneracies to the extent it can. A similar statement can be made about the bias-density degeneracy inherent to clustering- z measurements; the HBM does not eliminate this degeneracy, but does break this degeneracy to the extent possible by combination of the two methods.

3.3 Relation to standard clustering redshifts

The methods of Newman (2008), Ménard et al. (2013), and Schmidt et al. (2013) are all predicated on the assumption that, when both data and tracers are restricted to a redshift bin z , the cross-correlation $w_{DT}(\theta)$ between the data (targets) and the tracers is related to the tracer autocorrelation by

$$w_{DT}(\theta) = b_z w_z(\theta). \quad (20)$$

[In this subsection, there is no division of the target sample into phenotypes.] If this is true, and we consider the density fields in disjoint bins to be uncorrelated, then the cross-correlation between the full data sample D and the targets T_z in bin z becomes

$$w_{DT_z}(\theta) = f_z b_z w_z(\theta) \quad (21)$$

$$\Rightarrow f_z b_z = \frac{w_{DT_z}(\theta)}{w_z(\theta)}. \quad (22)$$

Each angular bin θ can thus provide an estimator for the quantities $q_z \equiv f_z b_z$. If we define XY_θ as the pair counts between population X and Y in a bin θ of angular separation, divided by the densities n_X and n_Y , then a standard estimator using equation (22) is

$$\hat{q}_{z\theta} = \frac{DT_{z\theta} - DR_\theta}{w_z(\theta) DR_\theta} \quad (23)$$

using the populations data galaxies (D), randoms (R), and tracers at redshift z (T_z).

All three of the above papers effectively produce a single estimator for q_z through a weighted sum over the $\hat{q}_{z\theta}$. [In these papers, the estimation of q_z is sometimes described as doing a least-squares fit of a model to the observed $w(\theta)$ data. When the free model parameter is the overall clustering amplitude, the least-squares solution can be expressed as an equivalent weighted sum.] An optimized estimator is easily derived under the assumption that errors are dominated by shot noise in the DT_z pair counts. The weights are proportional to $w_z(\theta) \cdot DR_z$ and the estimator can be rewritten as

$$\hat{q}_z = \frac{\sum_\theta w(\theta) (DT_{z\theta} - DR_\theta)}{\sum_\theta w^2(\theta) DR_\theta} \quad (24)$$

$$= \frac{\frac{1}{n_{T_z}} \sum_{i \in D, j \in T_z} w(\theta_{ij}) - \frac{1}{n_R} \sum_{i \in D, j \in R} w(\theta_{ij})}{\frac{1}{n_R} \sum_{i \in D, j \in R} w^2(\theta_{ij})}. \quad (25)$$

It is clear that clustering methods alone can never break the degeneracy between redshift distributions f_z and bias b_z . The similarity between equation (25) from the clustering- z methods and the optimal KDE estimator for $\hat{\delta}_z$ in equation (19) suggests that these quantities are related. In the Appendix, we demonstrate that in the limit where the photo- z 's are uninformative, and the clustering is in the Poisson limit, that maximization of the posterior using equation (13) and the optimal KDE in equation (19) yields the same result as the standard two-point methods. This strongly suggests that our joint constraint method makes use of all the information that is used by the standard clustering- z methods. Both methods are subject to the same degeneracy between b_z and f_z , though our method allows the photo- z information to potentially break this degeneracy.

4 SAMPLING AND PRIORS

Now we consider the problem of simultaneously constraining the redshift and type probability distributions of populations of galaxies

and their individual constituents. It is difficult to sample all variables simultaneously from the joint posterior $p(\mathbf{f}, \mathbf{z}, \mathbf{t}, \mathbf{b} | \mathbf{F}, \boldsymbol{\theta})$ in equation (11). It is possible, however, to draw samples from this posterior using a three-step Gibbs sampler because the conditional posterior distributions can be sampled. In the simulated data used to illustrate the method in this paper, we will assume that the biases $b_z^t = 1$, i.e. assume that the tracers and targets have the same fluctuations. Sampling of the \mathbf{b} will be demonstrated in a future publication, and hence we will be describing a two-step sampler in this section.

Each iteration of the Gibbs sampler comprises two steps which are (i) drawing a sample of \mathbf{f} from $p(\mathbf{f} | \mathbf{z}, \mathbf{t}, \mathbf{F}, \boldsymbol{\theta})$ and (ii) drawing pairs of z_i, t_i for each galaxy i from $p(z_i, t_i | \mathbf{f}, F_i, \theta_i)$ using the newly drawn \mathbf{f} . The conditional distributions can be read directly from the joint distribution in equation (11). We will make the further simplification that the effective area A' is independent of type and therefore $A(\mathbf{f})$ is constant. Next, we detail the expressions used in each of the two steps:

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

$$p(\mathbf{f} | \mathbf{z}, \mathbf{t}, \mathbf{F}, \boldsymbol{\theta}) \propto p(\mathbf{f}) \prod_{z,t} f_{zt}^{N_{zt}}. \quad (26)$$

It also depends on the prior information on \mathbf{f} , $p(\mathbf{f})$. The prior condition that $\sum_{z,t} f_{zt} = 1$, and $0 \leq f_{zt} \leq 1$, allows us to write the conditional posterior on \mathbf{f} as a Dirichlet distribution, as we will explore next in Section 4.1. This allows us to draw a realization of \mathbf{f} that we will use in the next step of the Gibbs sampler.

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

$$p(z_i, t_i | \mathbf{f}, F_i, \theta_i) \propto \mathcal{L}_{it_i} f_{it_i} (1 + \delta_{it_i}), \quad (27)$$

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

The two-step Gibbs sampler allows us to explore the joint posterior distribution and hence to get samples of the redshift distribution of the full galaxy population as well as the individual redshift probability distributions.

4.1 Introducing a prior

Let us now revisit the conditional posterior distribution on step (i) of the Gibbs sampling, $p(\mathbf{f} | \mathbf{z}, \mathbf{t}, \mathbf{F}, \boldsymbol{\theta})$. Such a distribution depends only on the number counts of \mathbf{z} and \mathbf{t} pairs for the sources, so we can write $p(\mathbf{f} | \mathbf{z}, \mathbf{t}, \mathbf{F}, \boldsymbol{\theta}) = p(\mathbf{f} | \mathbf{N})$. Now, we can use Bayes' theorem to write the posterior on the distribution parameters \mathbf{f} as a likelihood of the binned data given the \mathbf{f} probabilities times a prior on those probabilities:

$$p(\mathbf{f} | \mathbf{N}) \propto p(\mathbf{N} | \mathbf{f}) p(\mathbf{f}). \quad (28)$$

The likelihood of the binned counts \mathbf{N} given the probabilities \mathbf{f} follows a multinomial distribution:

$$p(\mathbf{N} | \mathbf{f}) = N! \prod_{z=1}^{N_z} \prod_{t=1}^{N_t} \frac{f_{zt}^{n_{zt}}}{n_{zt}!}. \quad (29)$$

The Dirichlet distribution is the conjugate prior of the multinomial distribution that has the great advantage that if we choose the prior to follow a Dirichlet distribution, which appropriately fulfills $0 \leq f_{zt} \leq 1$ and $\sum_{zt} f_{zt} = 1$, then the posterior distribution will also be Dirichlet-distributed.

In getting the conditional posterior distribution of interest, $p(\mathbf{f}|N)$, we will distinguish between two cases, uninformative and informative priors.

4.1.1 Uninformative prior

The Dirichlet distribution corresponding to an uninformative, uniform prior is:

$$p(\mathbf{f}) = (N_z N_t - 1)! \delta_D \left(1 - \sum_{zt} f_{zt} \right) \prod_{z=1}^{N_z} \prod_{t=1}^{N_t} \Theta(f_{zt}). \quad (30)$$

Combining with the likelihood in equation (29), we get the following posterior:

$$\begin{aligned} p(\mathbf{f}|N) &= (N + N_z N_t - 1)! \delta_D \left(1 - \sum_{zt} f_{zt} \right) \\ &\times \prod_{z=1}^{N_z} \prod_{t=1}^{N_t} \frac{\Theta(f_{zt}) f_{zt}^{n_{zt}}}{n_{zt}!} \\ &\equiv \text{Dir}(\mathbf{N}). \end{aligned} \quad (31)$$

This is a Dirichlet distribution parameterized by the counts N_{zt} from the last iteration of the sampler. For brevity, we will write all subsequent Dirichlet distributions using the compact form of equation (31).

4.1.2 Informative prior

Alternatively, an informative prior on the coefficients \mathbf{f} may come from a representative (random) subset of galaxies with known z, t – e.g. from a complete spectroscopic survey of a random subsample of targets. If $\mathbf{M} = \{M_{zt}\}$ are the counts of this prior sample found at each z, t pair, then the prior distribution of \mathbf{f} follows a Dirichlet distribution with parameters \mathbf{M} , and hence the conditional posterior follows a Dirichlet on the data counts from the last iteration plus the prior counts:

$$p(\mathbf{f}|N) \sim \text{Dir}(\mathbf{N} + \mathbf{M}). \quad (32)$$

In this way, it is very clear how the number of galaxies in the prior, M_{zt} , which will be small relative to the number of galaxies in the full sample, N_{zt} , will determine the relative importance of the prior in the HBM. The effect of the prior will be explored further in the next section, with a demonstration on simulations.

5 DEMONSTRATION ON SIMULATIONS

We now present a simple simulation of galaxy survey data and test our methodology by exploring some of its main features. We adopt simple galaxy properties and noise distributions as the method can be easily adapted to account for realistic effects.

In the simulation, we choose to work in a space where $z \in [0, 1]$ with 50 equally spaced bins, and we have 50 different galaxy types, which for convenience will be defined between 0 and 1 like the redshifts, so that $t \in [0, 1]$ and $N_z = N_t = 50$. For each redshift bin (or slice), we generate δ_z from a Gaussian Random Field (GRF)

with a resolution of 1024×1024 pixels that we will use to define galaxy positions in a way to simulate galaxy clustering. For every redshift slice, the fields are drawn from a power spectrum $P(k) \sim k^{-3}$, where k is a given scale in Fourier space, and they are chosen to be uncorrelated between different redshifts. The rms value of the GRF is around 2.5 for each slice, and hence we clip the GRF to maintain non-negative density. If $\rho(x, y, z) = \rho(\theta, z)$ is the value of the field at each line of sight and redshift, we define the overdensity field as:

$$\delta(\theta, z) = \frac{\rho(\theta, z)}{\bar{\rho}(z)} - 1, \quad (33)$$

where the mean field $\bar{\rho}(z)$ is taken, for each redshift slice, over all the 1024×1024 pixels in it. In reality, a (biased) estimate of this field can be obtained from a population of tracers with well-known redshifts, as described in Section 3. For this demonstration, we will take the field $\delta(\theta, z)$ as known, rather than create a tracer population and KDE estimator. This demonstration will also assume known bias ($b = 1$) and perfect detection/selection ($p(s) = 1$).

The simulation consists of $N = 10^5$ galaxies with types t , redshifts z , features \mathbf{F} , and positions θ produced as follows:

- We draw a galaxy type t for each galaxy from the following distribution:

$$p(t) \propto t^a \exp^{-(t/t_0)^a}, \quad (34)$$

with $a = 1.5$, $t_0 = 0.3$.

- We assign redshifts to these galaxies depending on their type by using a simple model whereby a given galaxy type can only span, at most, three possible redshift values:

$$p(z|t) = \begin{cases} \text{if } t = 0 & \begin{cases} 0.8 & z = t \\ 0.2 & z = t + 0.02 \end{cases} \\ \text{if } t = 1 & \begin{cases} 0.8 & z = t \\ 0.2 & z = t - 0.02 \end{cases} \\ \text{else} & \begin{cases} 0.6 & z = t \\ 0.2 & z = t \pm 0.02 \end{cases} \end{cases} \quad (35)$$

- We generate an observable feature F for each galaxy using a very simplified model: Based on each galaxy's type t , we create a $1Dflux$ such that the measurement likelihood is Gaussian with some variance σ_F^2 , as:

$$p(F|t) = \mathcal{L}_{Ft} = \mathcal{N}(t - F, \sigma_F^2). \quad (36)$$

- Finally, given each galaxy's redshift, we draw a position θ for each of them following:

$$p(\theta|z) \propto (1 + \delta(\theta, z)), \quad (37)$$

where the field δ is taken exactly using the GRF as in equation (33).

Fig. 2 presents a graphical description of the simulation described here, showing some example GRFs and galaxies drawn.

Next, we will use this simulation to test the methodology described earlier in this paper. In applying the sampling methods described in Section 4, we will obtain samples of the full posterior redshift distributions of galaxy populations and their individual members. These samples can be directly used in cosmological analyses to propagate the uncertainties coming from photometric redshift estimation. However, in analysing the results, we will also find it useful to define a simple metric that allows us to easily compare between different cases or variations of the scheme. For that, we compute the difference in the mean between the estimated and the true redshift distributions, where using the former we can compute for each sample j of the redshift distribution coming from the

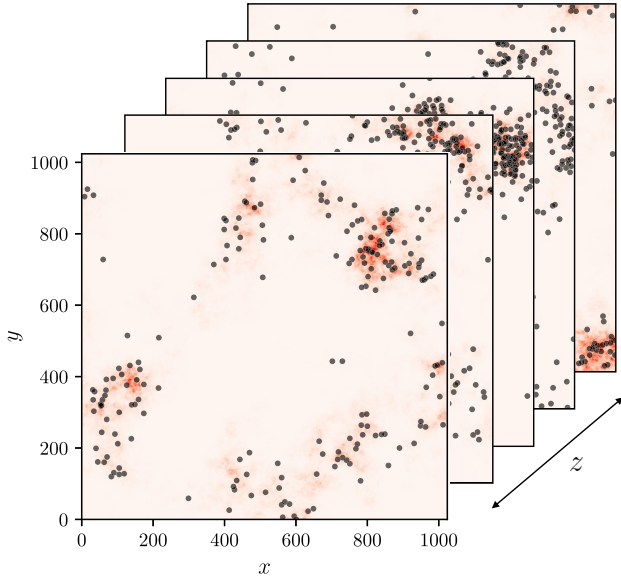


Figure 2. A graphical representation of the clustering present in the simulation used in this work. The plot shows five random redshift slices of the simulation, each of them showing the Gaussian Random Field (GRF) δ , in red, and a subset of the galaxies placed in the slice (a 5 percent random subset). As explained in the main text, there are no correlations between δ fields of distinct slices. The axis in the plot correspond to the 1024×1024 resolution of the GRFs.

HBM implementation (or the Dirichlet prior) and the latter is a fixed quantity:

$$\Delta z_j = \langle z_{\text{est},j} \rangle - \langle z_{\text{true}} \rangle. \quad (38)$$

We will show both the distribution of these Δz_j metric for the samples j from the posterior, together with the median and standard deviation of such distributions.

In all the tests performed in this section, we run the chains of the Gibbs sampler using four different walkers, each with 2500 samples, for a total of 10^4 samples per chain, and require the Gelman & Rubin convergence metric (Gelman & Rubin 1992) to be $R < 1.03$ for considering the chain converged. The walkers are initialized at different realizations of the Dirichlet prior. All the chains run in less than an hour in a two-core laptop when using a Python implementation.

5.1 Fiducial results

Now we present the results of a fiducial case in which we use a set of 8×10^4 galaxies from our simulation as our photometric sample (we only know F and θ about these) and a random subset of 10^3 galaxies as our prior (for these, we know z and t). In this and other cases, we will use ‘ F ’ to refer to analyses that use only photometric (feature) information, and ‘ $F + \delta$ ’ to refer to the combination of photometric information with clustering as per equation (11).

Fig. 3 shows the redshift and type distributions of the simulation together with the recovered distributions obtained with samples of the full posterior distributions, as violin plots. For the recovered distributions, both the F only and the $F + \delta$ cases are shown. In the two cases, the redshift and type distributions of the photometric sample are effectively recovered, within the uncertainties displayed by the violins in each case. However, strong differences appear between

the two cases in the magnitude of the uncertainties, especially for the redshift distribution. The observed features F in the photometric sample are used in the HBM scheme to tighten the posterior distributions compared to the prior, but these features F mainly inform the types t of galaxies, through the likelihood $\mathcal{L}_{F,t}$. In contrast, when adding clustering to HBM that informs the redshift part of the problem, and can tighten the redshift distribution posterior significantly, this can be appreciated in the left-hand panel of Fig. 3.

Fig. 4 shows the distribution of the Δz metric computed from the redshift posteriors shown in the left-hand panel of Fig. 3. Again, the addition of clustering to the HBM method considerably sharpens the recovered Δz distribution, without causing any apparent bias. Moreover, this plot includes the distribution of Δz from samples of the Dirichlet prior. The median and standard deviation of the Δz distributions shown in that figure can be found in Table 2.

So far we have looked at the sampling of the redshift and type distributions for the entire population. However, in the HBM approach, we are sampling both the probability distributions of the populations together with those of their individual members. In particular, we are sampling the redshift probability distribution of each individual galaxy in the sample. Fig. 5 shows the sampling of the redshift probability distributions of two random galaxies in the sample, for the HBM method with and without clustering information, showing the true redshift of the galaxies, for comparison. We can see how the addition of the clustering information sharpens the individual redshift probability distributions as well.

5.1.1 Clustering only up to a given z

The framework for the addition of clustering information in this work enables the consistent usage of that information even if it is only available in a part of the redshift range of interest (see Section 3.2). Here, we explore this aspect of the method by using the exact knowledge of the δ -field only in half the redshift range of the simulation. Fig. 6 shows the redshift distribution posteriors when using the knowledge of δ in the range $z < 0.5$. The effect of clustering is apparent in the lower half of the redshift range, and the posteriors in the other half appear consistent with the HBM method without clustering information. More quantitatively, Fig. 4 and Table 2 show how the Δz metrics behave when using clustering information in only a part of the redshift range, demonstrating that it significantly helps constraining that metric over the HBM without clustering case, without being as constraining as using the clustering information in the entire redshift range.

5.2 Biases in the prior

So far in this Section we have shown how the method described in this work can provide an unbiased sampling of the redshift distribution of a galaxy population when we use a representative subsample of it as a prior. However, in reality, having a representative galaxy subsample with both spectroscopy and accurate photometry, so we know z and t for those galaxies, can be very difficult. For this reason, we will study how the method is sensitive to biases in the subsample used for the prior and, especially, if the method is able to overcome such biases for the estimated distributions. In the analysis of the possible biases that can plague our prior sample, we will split the prior expression as in equation (3), $p(z, t) = p(z|t)p(t) = f_z^t f_t$, and we will explore the effects of biasing the prior separately in f_t and f_z^t , and then in both simultaneously.

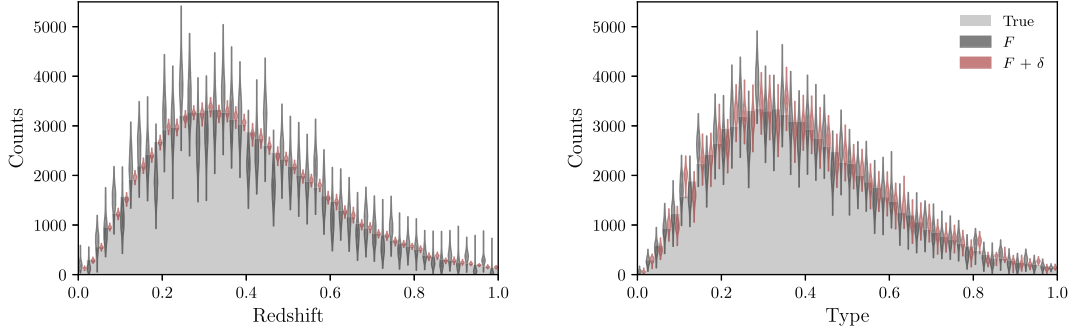


Figure 3. (Left panel): Redshift posterior distribution, marginalized over type, for the HBM method with and without the inclusion of clustering in the posterior. (Right panel): Posterior distribution of galaxy type, marginalized over redshift, for the same cases. In both plots, the true distributions of redshift and type, respectively, are shown for comparison with the recovered distributions. Note the strong effect of the clustering addition in recovering a tighter redshift distribution.

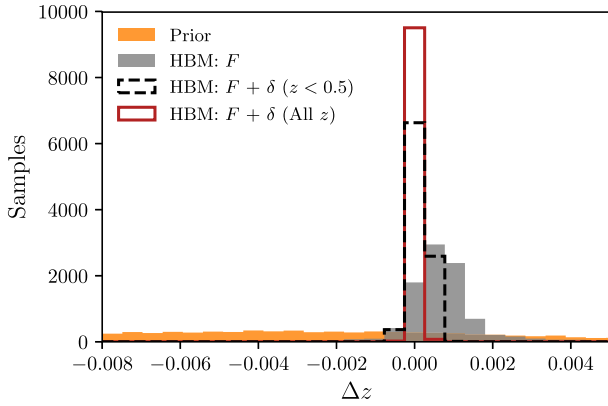


Figure 4. Distribution of the Δz metric, the difference between the mean of the recovered and true redshift distributions, for the redshift distribution posteriors displayed in Figs. 3 and 6. The effect of the clustering addition is clear again. The constraining power of the prior is also shown in this plot.

Table 2. Median and standard deviation of the Δz distributions presented in Fig. 4, showing the constraints for the prior and the HBM runs, with and without clustering information, for the fiducial case.

Fiducial	Median (Δz)	Standard Deviation (Δz)
Prior	-0.00356	0.00608
HBM: F	0.00068	0.00235
HBM: $F + \delta$ ($z < 0.5$)	0.00012	0.00022
HBM: $F + \delta$ (All z)	0.00003	0.00010

5.2.1 Bias in f_t

We first introduce a bias in the $f_t = p(t)$ part of the prior. The $p(t)$ distribution gives the abundance of galaxies in the different regions of the observed feature space, for instance it would correspond to the density of galaxies in each cell of a self-organized map of galaxy colours. In presenting the details of the simulation, we showed that a type t for each galaxy is drawn from the $p(t)$ distribution in equation (34), with $a = 1.5$, $t_0 = 0.3$. Now, in order to select galaxies for the prior, we draw a subset of 1000 galaxies from the simulation following the distribution in equation (34), with $a = 1.3$ and $t_0 = 0.25$, while keeping the f_z^t relation as in equation (35). Because of the relation between z and t , this bias in f_t produces a bias in redshift, of the order of $\Delta z \lesssim -0.02$.

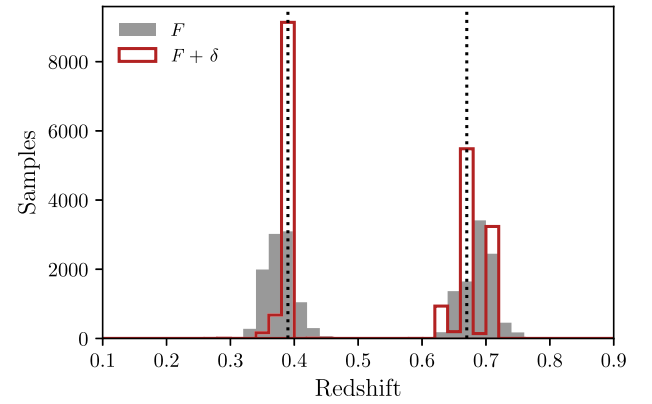


Figure 5. Posterior redshift probability distributions for two random galaxies in the sample, at redshifts $z = 0.39$ and $z = 0.67$ (dotted lines), using the HBM with and without clustering information. The addition of clustering can significantly sharpen the redshift posteriors of individual galaxies as well as that of the full population $n(z)$.

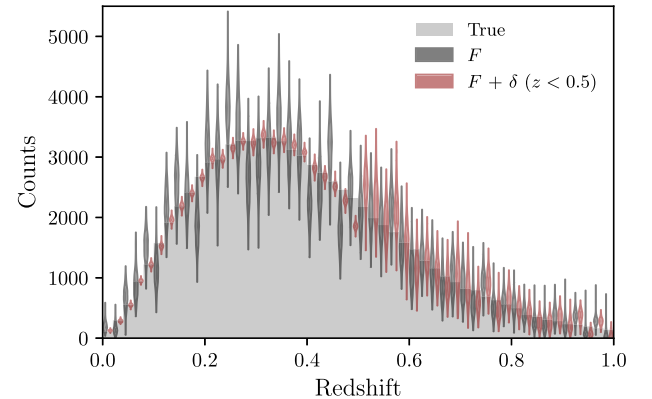


Figure 6. Redshift posterior distribution, marginalized over type, for the HBM method and the addition of clustering to it. Here, we can see that the method works properly when density information is available for only half of the redshift range ($z < 0.5$), with the expected result that the posterior at redshifts with $z > 0.5$ resembles the photometry-only case.

The Δz constraints for this case coming from the prior and the HBM method with and without clustering information are shown in the upper panel of Fig. 7 and in Table 3. In there, we can clearly see the bias in the prior, and we can also see how the HBM method,

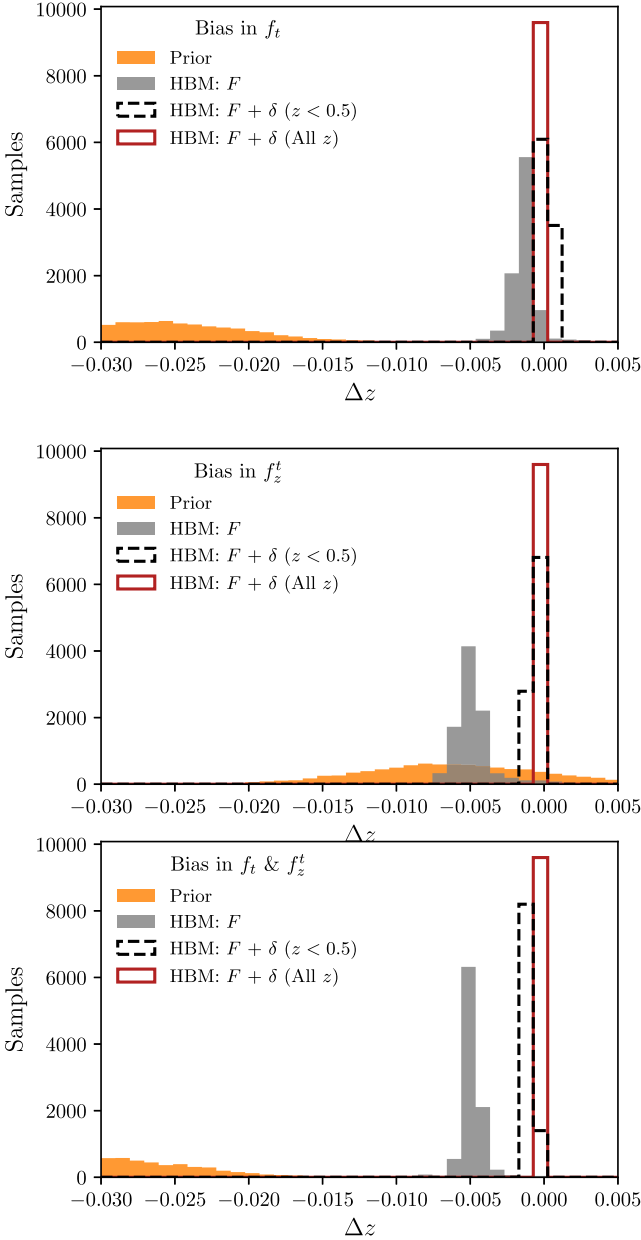


Figure 7. Distribution of the Δz metric for the redshift distribution posteriors coming from the prior and the HBM methods, with and without using clustering information, in the presence of biases in the prior: bias in f_t (upper), bias in f_z^t (centre), and bias in both f_t and f_z^t (lower). One can see how the HBM method without clustering information can correct biases in the f_t part of the prior, but biases in the f_z^t part can only be corrected by adding clustering information.

both with and without the usage of clustering information, is able to correct for that bias in the derived redshift distribution for the sample. The F -only analysis reduces redshift bias by $20 \times$ in this case, because the bias is due to a misestimate of the distribution of galaxy phenotypes, which can be corrected with photometry from the full sample of sources, even if the photometry is noisy.

5.2.2 Bias in f_z^t

We introduce a bias in the $f_z^t = p(z|t)$ part of the prior, which tells us about the relation between galaxy types (or observed features)

Table 3. Median and standard deviation of the Δz distributions presented in Fig. 7, showing the constraints for the prior and the HBM runs, with and without clustering information, for the cases with biases in the prior.

	Median (Δz)	Standard Deviation (Δz)
<i>Bias in f_t</i>		
Prior	−0.02644	0.00614
HBM: F	−0.00132	0.00181
HBM: $F + \delta$ ($z < 0.5$)	0.00015	0.00024
HBM: $F + \delta$ (All z)	−0.00010	0.00009
<i>Bias in f_z^t</i>		
Prior	−0.00632	0.00628
HBM: F	−0.00493	0.00338
HBM: $F + \delta$ ($z < 0.5$)	−0.00061	0.00021
HBM: $F + \delta$ (All z)	−0.00024	0.00010
<i>Bias in f_t & f_z^t</i>		
Prior	−0.03083	0.00623
HBM: F	−0.00496	0.00124
HBM: $F + \delta$ ($z < 0.5$)	−0.00094	0.00020
HBM: $F + \delta$ (All z)	−0.00026	0.00010

and redshift, which is known as the *colour–redshift* relation if we are only using galaxy colours as features. We randomly draw galaxies from the simulation to be treated as ‘spectroscopic,’ and for a spectroscopic source of type t , we assign a redshift with a biased modification of the relation in equation (35):

$$p(z|t) = \begin{cases} \text{if } t = 0 & \begin{cases} 0.85 & z = t \\ 0.15 & z = t + 0.02 \end{cases} \\ \text{if } t = 1 & \begin{cases} 0.75 & z = t \\ 0.25 & z = t - 0.02 \end{cases} \\ \text{else} & \begin{cases} 0.5 & z = t - 0.02 \\ 0.4 & z = t \\ 0.1 & z = t + 0.02 \end{cases} \end{cases} \quad (39)$$

The bias in this relation produces a shift in the mean redshift of the spectroscopic prior, $\Delta z \sim -0.005$. The Δz constraints for this case coming from the sampling of the prior and the HBM method, with and without clustering information, are presented in the middle panel of Fig. 7 and in Table 3. In this case, the HBM method without clustering information is unable to correct for the bias in the prior, as the observed features of the galaxies in the sample do not inform the f_z^t ’s, only the f_t ’s, which are unbiased in this case. Without the usage of clustering information, the colour–redshift part of the problem is only informed by the prior. However, when using clustering in the HBM, we are adding direct information about redshift, and hence the method is able to significantly reduce the redshift biases in the prior.

5.2.3 Bias in f_t and f_z^t

Now we combine the two types of biases analysed before in a more realistic case where the prior is biased in both the types – redshift relation and the abundance of the different galaxy types, in the same way, we introduced each of such biases above. The results are shown in the lower panel of Fig. 7 and in Table 3. Again, the HBM method without clustering information is able to correct the bias coming from the prior having a different galaxy-type distribution, but is unable to correct for a wrong – redshift relation in the prior. On the other hand, the addition of clustering allows the HBM method to significantly reduce biases coming from the prior,

as it is using the galaxy sample-observed features to inform the galaxy-type distribution and the clustering information to inform the redshift part.

6 CONCLUSIONS

There are three main sources of information for estimating the redshift probability distributions of galaxies and ensembles of galaxies in a wide-field survey. First, we have prior information, which comes from a subset of galaxies, typically spanning a smaller area in the sky, for which we have both reliable redshift estimates and precise photometry. Then, using all the galaxies in the wide-field sample, there exist methods for estimating those distributions from either the photometry of such galaxies or the clustering of them against a tracer population with precise and well-characterized redshift estimates, which overlaps with the wide-field sample in the sky. However, one encounters important complications when trying to combine the independent answers from these last two approaches, since, for instance, the tracer population used in clustering methods typically only covers a subset of the redshift range of interest, and hence so do the derived redshift distributions. Also, it is very important to correctly propagate the uncertainties due to the limited size of the prior sample compared to the full population into the inference.

The method presented in this paper is the first to combine these three sources of information in a unified and consistent scheme, using all the information available. The method consists of a hierarchical Bayesian model which allows us to sample both the redshift distribution of a galaxy sample and the individual redshifts of galaxies therein. It can be seen as a generalization of the approach in Leistedt et al. (2016) to include clustering information as well as an informative prior. In the case where one marginalizes over the latent density fields δ , the method gives a rigorous posterior probability for the redshift distribution $n(z)$, as well as individual z_i 's, if the galaxies are a Poisson sampling of a density field with known statistical properties. The approximation that we make here to enable a straightforward sampling process is that the stochastic density field can be perfectly estimated from a kernel density estimator using the positions of a tracer population. In the limit of weak photometric information and weak clustering, the maximum-posterior $n(z)$ estimator is shown to be equivalent to standard clustering- z methods using two-point correlation functions. With the KDE approximation, the posterior $n(z)$ from our HBM no longer fully samples the effects of shot noise in the tracer population; this will be addressed in future improvements.

This paper's simple simulation, which is given knowledge of the true density field, demonstrates the benefits of adding clustering information into the photometric redshift estimation. This inclusion not only sharpens the posterior redshift probability distributions of galaxies and populations of galaxies but, more importantly, it can overcome biases in the prior which are not possible to resolve when only using galaxy photometry. That demonstrates the advantages of the combination of clustering and photometry into the same inference, where the usage of the two techniques separately would yield different answers which would be hard to combine or result in a failed validation cross-check. We also show how clustering information can be naturally used even if it is only available for a subset of the redshift range of interest, without producing any additional biases.

The methodology described herein does not resolve redshift degeneracies that are intrinsic to the photometric and clustering methods: for photo- z 's, there remains the possibility that the map of colour to redshifts is non-unique, i.e. galaxies at distinct z can have

identical features F . For clustering, it remains true that the redshift dependence of bias $b(z)$ is degenerate with the redshift distribution. Our method does, however, exploit any cases in which one method might resolve the degeneracies of the other – for instance if the photometric features of some galaxy-type t restrict $p(z|t)$ to a small enough range of z that variation in $b'(z)$ can be assumed to be fairly small. The success of this method in producing $n(z)$ to the accuracy desired for large cosmological surveys will therefore still depend on the details of the survey and the galaxy population; but we now have a means to harness all of the available information.

One of the attractive aspects of the hierarchical Bayesian model for photometric redshift estimation is that the sampling of the posterior distribution of $n(z)$ for the galaxy population allows us to correctly propagate the uncertainties in the estimation of that distribution into cosmological analyses of galaxy clustering and weak gravitational lensing. The model can also be extended to correct for other observational effects, such as the calibration of photometric zero points (Leistedt et al. 2018), and can trivially accommodate more complicated noise likelihood functions. Finally, the method presented here can eventually be incorporated as part of a fully Bayesian analysis of galaxy surveys, which would use all the available information to reconstruct the matter density field (Jasche & Kitaura 2010; Jasche & Wandelt 2013).

ACKNOWLEDGEMENTS

The authors thank Àlex Alarcón, Romain Buchs, Chris Davis, Daniel Gruen, Boris Leistedt, and Justin Myles for helpful conversations about this topic. This work was supported by grants AST-1311924 and AST-1615555 from the US National Science Foundation, and DE-SC0007901 from the US Department of Energy.

REFERENCES

- Arnouts S. et al., 2002, *MNRAS*, 329, 355
- Benítez N., 2000, *Astrophys. J.*, 536, 571
- Benjamin J. et al., 2013, *MNRAS*, 431, 1547
- Bernstein G. M., 1994, *Astrophys. J.*, 424, 569
- Bolzonella M., Miralles J., Pell R., 2000, *Astron. Astrophys.*, 492, 476
- Bonnett C. et al., 2016, *Phys. Rev. D*, 94, 042005
- Bonnett C., 2015, *MNRAS*, 449, 1043
- Brammer G. B., van Dokkum P. G., Coppi P., 2008, *Astrophys. J.*, 686, 1503
- Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, 432, 1483
- Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, *Astron. J.*, 132, 926
- Colless M. et al., 2001, *MNRAS*, 328, 1039
- Collister A. A., Lahav O., 2004, *Publ. Astron. Soc. Pac.*, 116, 345
- Cox D. R., 1955, *J. Roy. Stat. Soc. Ser. B (Methodological)*, 17, 129
- Cunha C. E., Huterer D., Busha M. T., Wechsler R. H., 2012, *MNRAS*, 423, 909
- Dahlen T. et al., 2013, *Astrophys. J.*, 775, 93
- Davis C. et al., 2017, preprint ([arXiv:1710.02517](https://arxiv.org/abs/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., et al., 2013, *ESO Messenger*, 154, 44
- DES Collaboration et al., 2017, preprint ([arXiv:1708.01530](https://arxiv.org/abs/1708.01530))
- DESI Collaboration et al., 2016, preprint ([arXiv:1611.00036](https://arxiv.org/abs/1611.00036))
- Drinkwater M. J. et al., 2010, *MNRAS*, 401, 1429
- Flaugher B. et al., 2015, *Astron. J.*, 150, 150
- Gatti M. et al., 2018, *MNRAS*, 477, 1664
- Gelman A., Rubin D. B., 1992, *Stat. Sci.*, 7, 457
- 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

- Hildebrandt H. et al., 2010, *Astron. Astrophys.*, 523, A31
Hildebrandt H. et al., 2012, *MNRAS*, 421, 2355
Hildebrandt H. et al., 2017, *MNRAS*, 465, 1454
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
Jasche J., Kitaura F. S., 2010, *MNRAS*, 407, 29
Jasche J., Wandelt B. D., 2012, *MNRAS*, 425, 1042
Jasche J., Wandelt B. D., 2013, *MNRAS*, 432, 894
Kaiser N., Tonry J. L., Luppino G. A., 2000, *Publ. Astron. Soc. Pac.*, 112, 768
Le Fèvre O. et al., 2005, *Astron. Astrophys.*, 439, 845
Leistedt B., Mortlock D. J., Peiris H. V., 2016, *MNRAS*, 460, 4258
Leistedt B., Hogg D. W., Wechsler R. H., DeRose J., 2018, preprint (arXiv:1807.01391)
LSST Dark Energy Science Collaboration, 2012, preprint (arXiv:1211.0310)
Masters D. et al., 2015, *Astrophys. J.*, 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, p. 84460Z, preprint <http://proceedings.spiedigitallibrary.org/proceeding>
Newman J. a., 2008, *Astrophys. J.*, 684, 88
Sánchez C. et al., 2014, *MNRAS*, 445, 1482
Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, *MNRAS*, 431, 3307
Suchyta E. et al., 2016, *MNRAS*, 457, 786
York D. G. et al., 2000, *Astron. J.*, 120, 1579

APPENDIX: EQUIVALENCE TO CORRELATION-FUNCTION METHODS

Consider the case when there is no photometric information available on our detected galaxies, 100 per cent selection rate, and the window function is independent of redshift so that $A(\mathbf{f})$ becomes some constant value A . Equation (13) for the posterior reduces to

$$p(\mathbf{f}, \mathbf{z}, \mathbf{b}|\theta) \propto p(\mathbf{f}) \prod_i f_{z_i} (1 + b_{z_i} \hat{\delta}_{iz}). \quad (\text{A1})$$

If we marginalize over \mathbf{z} we obtain

$$p(\mathbf{f}, \mathbf{b}|\theta) \propto \sum_{\mathbf{z}} p(\mathbf{f}, \mathbf{z}, \mathbf{b}|\theta) \quad (\text{A2})$$

$$\propto p(\mathbf{f}) \prod_i \left[\sum_{\mathbf{z}} f_{z_i} (1 + b_{z_i} \hat{\delta}_{iz}) \right] \quad (\text{A3})$$

$$= p(\mathbf{f}) \prod_i \left[1 + \sum_{\mathbf{z}} f_{z_i} b_{z_i} \hat{\delta}_{iz} \right], \quad (\text{A4})$$

$$\hat{\delta}_{iz} \equiv \hat{\delta}_z(\theta_i). \quad (\text{A5})$$

A minimal prior is that the redshift distribution fractions f_z must be non-negative and sum to unity. In this case, we can effectively drop $p(\mathbf{f})$ from this equation, and we find that the posterior probability depends only upon the quantities $q_z \equiv f_z b_z$, i.e. we recover the degeneracy between bias and $n(z)$ that is well known for pure clustering redshifts. The log of the posterior is now

$$\log P(\mathbf{f}, \mathbf{b}|\theta) = \sum_i \log \left[1 + \sum_{\mathbf{z}} q_z \hat{\delta}_{iz} \right] + \text{const.} \quad (\text{A6})$$

We seek the estimated values \hat{q}_z which maximize this quantity. Writing the estimated projected density along θ_i as $y_i \equiv \sum_{\mathbf{z}} \hat{q}_z \hat{\delta}_{iz}(\theta_i)$,

the condition for posterior maximization at $\{\hat{q}_z\}$ is the set of simultaneous equations

$$0 = \frac{\partial}{\partial \hat{q}_z} \sum_i \log(1 + y_i) \quad (\text{A7})$$

$$= \sum_i \frac{\hat{\delta}_{iz}}{1 + y_i}. \quad (\text{A8})$$

We examine the limit $|y_i| \ll 1$, when the projected mass fluctuations are weak. Note that this does *not* require the spatial fluctuations δ to be weak. The solution becomes

$$0 = \sum_i [\hat{\delta}_{iz}(1 - y_i + y_i^2 - \dots)] \quad (\text{A9})$$

$$\approx \sum_i \hat{\delta}_{iz} - \sum_{z'} q_{z'} \sum_i \hat{\delta}_{iz} \hat{\delta}_{iz'} + \sum_{z', z''} q_{z'} q_{z''} \sum_i \hat{\delta}_{iz} \hat{\delta}_{iz'} \hat{\delta}_{iz''} \quad (\text{A10})$$

$$\approx S_{1z} - q_z S_{2z} + q_z^2 S_{3z}, \quad (\text{A11})$$

where we have defined

$$S_{nz} = \sum_i \hat{\delta}_{iz}^n. \quad (\text{A12})$$

In going from (A10) to (A11), we have simplified a matrix equation to a series of linear equations by assuming that the independence of the density fields in distinct bins drives terms like $\sum_i \hat{\delta}_{iz} \hat{\delta}_{iz'}$ to zero unless $z = z'$.

If we make the approximation that the density fluctuations $\hat{\delta}$ are weak or symmetric in the sense that $S_{3z} S_{1z} \ll S_{2z}^2$, then the posterior is maximized at values \hat{q}_z solving equation (A11):

$$\hat{q}_z \approx \frac{S_{1z}}{S_{2z}} \left(1 + \frac{S_{3z} S_{1z}}{S_{2z}^2} \right). \quad (\text{A13})$$

At this point, it is useful to calculate the expectation value of S_{nz} under Poisson sampling of the density fields $\hat{\delta}_z$:

$$\langle S_{nz} \rangle = \int d^2\theta \left\{ n \sum_{z'} f_{z'} [1 + b_{z'} \hat{\delta}_{z'}(\theta)] \hat{\delta}_z(\theta)^n \right\} \quad (\text{A14})$$

$$= nA (\mu_{nz} + f_z \mu_{(n+1)z}), \quad (\text{A15})$$

where n is the mean source density, A is the survey area, and we use the central moments of the density estimators

$$\mu_{nz} \equiv \langle \hat{\delta}_z^n(\theta) \rangle_{\theta}. \quad (\text{A16})$$

We have $\mu_{1z} = 0$ and $\mu_{2z} = \sigma_z^2$, the variance of the density estimator.

At this point, we will make an approximation that the clustering is weak, $\sigma_z \ll 1$, and that the central moments of the field satisfy $\mu_{nz} \lesssim \sigma_z^n$, so that we can retain only leading terms in δ . In this limit, we can approximate the denominator in equation (A13) by its expectation value, $\langle S_{2z} \rangle = nA(\sigma_z^2 + f_z \mu_{3z}) \approx nA\sigma_z^2$. It can be shown that the correction term in parentheses in equation (A13) will cancel the influence of the μ_{3z} term to first order. We therefore arrive at a very simplified estimate of the maximum posterior solution \hat{q}_z :

$$\hat{q}_z \approx \frac{1}{\sigma_z^2} \sum_i \hat{\delta}_{iz} \quad (\text{A17})$$

$$= \frac{1}{N_D} \sum_{i \in D} \frac{\frac{1}{n_{Tz}} \sum_{j \in Tz} w_z(\theta_{ij}) - \frac{1}{n_R} \sum_{j \in R} w_z(\theta_{iR})}{\frac{1}{n_R} \sum_{j \in R} w_z^2(\theta_{ij})} \quad (\text{A18})$$

$$\approx \frac{\frac{1}{n_{Tz}} \sum_{i \in D, j \in Tz} w_z(\theta_{ij}) - \frac{1}{n_R} \sum_{i \in D, j \in R} w_z(\theta_{iR})}{\frac{1}{n_R} \sum_{i \in D, j \in R} w_z^2(\theta_{ij})}. \quad (\text{A19})$$

Equation (A18) makes use of the KDE for $\hat{\delta}$ from equation (19). In the final line, we approximated that the denominator of (19) is nearly constant for all galaxies i in the data sample D . The sets D , Tz , and R are the target galaxies (data), the tracer population in redshift shell z , and a randomly distributed sample, respectively. The resultant maximum-posterior estimate for q_z is seen to be the same, under these weak-field approximations, as the Poisson-limited optimal estimator for q_z from using two-point functions, equation (25).

This demonstrates that in the weak-field, no-photometry limit, our posterior probability has a maximum at the same value for $q = fb$ as the standard correlation-function-based clustering- z methodology. We have not demonstrated, however, that the variances of the two estimators coincide. Nonetheless the equivalence of the maximum-posterior estimator to the standard two-point estimator demonstrates that the latter is a limiting case of our method, and we can expect our principled approach to the $n(z)$ posterior to incorporate all the information of the two-point functions – especially if one can truly marginalize over the density fields.

This paper has been typeset from a \LaTeX file prepared by the author.