MNRAS 000, 1–22 (2015) Preprint 6 January 2021 Compiled using MNRAS LATEX style file v3.0

A Composite Likelihood Approach for Inference under Photometric Redshift Uncertainty

M. M. Rau[1]★, C. B. Morrison[2], S. J. Schmidt[4], S. Wilson[3], R. Mandelbaum[1], Y.-Y. Mao[5] for the LSST Dark Energy Science Collaboration

[1]McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213
[2]Department of Astronomy, University of Washington, Box 351580, Seattle, WA 98195, USA
[3]School of Computer Science and Statistics, Lloyd Institute, Trinity College, Dublin, Ireland
[4]Department of Physics, University of California, Davis, CA 95616, USA
[5]Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Obtaining accurately calibrated redshift distributions of photometric samples is one of the great challenges in photometric surveys like LSST, Euclid, HSC, KiDS, and DES. We combine the redshift information from the galaxy photometry with constraints from two-point functions, utilizing cross-correlations with spatially overlapping spectroscopic samples. Our likelihood framework is designed to integrate directly into a typical large-scale structure and weak lensing analysis based on two-point functions. We discuss efficient and accurate inference techniques that allow us to scale the method to the large samples of galaxies to be expected in LSST. We consider statistical challenges like the parametrization of redshift systematics, discuss and evaluate techniques to regularize the sample redshift distributions, and investigate techniques that can help to detect and calibrate sources of systematic error using posterior predictive checks. We evaluate and forecast photometric redshift performance using data from the CosmoDC2 simulations, within which we mimic a DESI-like spectroscopic calibration sample for cross-correlations. Using a combination of spatial cross-correlations and photometry, we show that we can provide calibration of the mean of the sample redshift distribution to an accuracy of at least $0.002(1+z)$, consistent with the LSST-Y1 science requirements for weak lensing and large-scale structure probes.

Key words: keyword1 – keyword2 – keyword3

1 INTRODUCTION

With ongoing and future large area photometric surveys like the Dark Energy Survey (DES; e.g., Abbott et al. 2018b), the Kilo-Degree Survey (KiDS; e.g., Hildebrandt et al. 2017), the Hyper Suprime-Cam (HSC; e.g., Aihara et al. 2018), the Rubin Observatory Legacy Survey of Space and Time (LSST; e.g., Ivezić et al. 2019), the Roman Space Telescope (e.g. Spergel et al. 2015) and Euclid (e.g. Laureijs et al. 2011) modern cosmology has entered the era of precision cosmology, where it becomes increasingly important to accurately account for sources of systematic bias and uncertainty (e.g. Mandelbaum 2018). Large area photometric surveys constrain cosmological parameters and the growth of structure using two-point statistics of galaxy and shear fields (see e.g. Hildebrandt et al. 2017; Uitert et al. 2017; Abbott et al. 2018a; Joudaki et al. 2018; Hikage et al. 2019; Heymans et al. 2020). Using only the broadband photometry of galaxies allows for a limited accuracy in the estimated redshifts. In photometric surveys, we therefore typically consider two-point statistics of density fields that have been projected along the line-of-sight, i.e., in the redshift direction. They are then subsequently compared with the corresponding weak lensing (WL) and large scale structure (LSS) theory predictions in a likelihood framework. These theory predictions have to account for the line-of-sight projection, and therefore depend on the redshift distribution of the galaxies in the sample that have to be accurately modelled and calibrated (see e.g. Huterer et al. 2006; Hoyle et al. 2018; Tanaka et al. 2018; Hildebrandt et al. 2020; Joudaki et al. 2020).

A primary goal of large area photometric survey programs is to map the growth of structure and expansion history of the Universe, and thereby constrain the dark energy equation of state via the distance-redshift and growth-redshift relations (see e.g., Albrecht et al. 2006, p. 31) which both enter the WL and LSS modelling. Note that these fundamental relationships within our cosmological model are redshift dependent, as are some key sources of theoret-

★ E-mail: markusr@andrew.cmu.edu

ical uncertainty, such as the galaxy-dark matter bias model (see e.g. Matarrese et al. 1997; Clerkin et al. 2015; Chang et al. 2016; Simon & Hilbert 2018; Prat et al. 2018). The inferred ensemble redshift distributions for samples of galaxies can therefore exhibit a degeneracy with cosmological or astrophysical parameters. Inaccurate distance, or redshift, measurements based on the photometry of the galaxies are therefore important modelling systematics in these surveys (e.g. Ma et al. 2006; Bernstein & Huterer 2010). We therefore must exploit all data sources that have the potential to break these degeneracies to perform efficient and accurate inference.

The two methods available to constrain the redshift of galaxies in the absence of accurate spectroscopic measurements are 'template fitting' methods and empirical methods that 'learn' the mapping between photometry and redshift (for a recent review, see Salvato et al. 2019). SED fitting methods fit the galaxy photometry with models of the galaxy spectral energy distribution (SED; e.g., Arnouts et al. 1999; Benítez 2000; Ilbert et al. 2006; Feldmann et al. 2006; Greisel et al. 2015; Leistedt et al. 2016; Malz & Hogg 2020). Machine Learning-based methods infer photometric redshifts by constructing a density estimate for the conditional distribution of the galaxy redshifts given their photometry (Tagliaferri et al. 2003; Collister & Lahav 2004; Gerdes et al. 2010; Carrasco Kind & Brunner 2013; Bonnett 2015; Rau et al. 2015; Hoyle 2016). Combinations of both these techniques have also been investigated (Speagle & Eisenstein 2015; Hoyle et al. 2015). Unfortunately the accuracy of these techniques is limited since they suffer from different sources of systematic error. Template fitting approaches can be systematically biased, if fits are constructed using sets of spectral energy distributions that are not representative of all galaxies in the sample. In contrast, photometric redshift techniques that require a training set can produce systematically biased results due to incomplete spectroscopic training samples. It is particularly difficult to obtain representative spectroscopic data due to the long exposure times that are necessary to obtain accurate spectroscopic redshifts for faint sources (see e.g. Huterer et al. 2014; Newman et al. 2015).

Instead of inferring photometric redshifts by fitting models for the spectral energy distribution, we can also infer photometric redshift infromation using spatial cross-correlations between photometric samples and spectroscopic samples (e.g. Newman 2008; Ménard et al. 2013; McQuinn & White 2013; Scottez et al. 2016; Raccanelli et al. 2017; Morrison et al. 2017; Davis et al. 2017; Gatti et al. 2018). Cross-correlation methods measure the spatial cross correlation between a reference sample with accurate redshift information, typically spectroscopic galaxy catalogs, and photometric samples that do not have accurate redshift information. Ignoring cosmic magnification effects (see e.g. Scranton et al. 2005) the expected spatial cross correlation is only nonzero for samples at the same redshift. By cross correlating subsamples of spectroscopic samples that are selected in thin redshift slices with these photometric catalogs and comparing the resulting signals, we can reconstruct the redshift distribution of the unknown photometric sample.

It is important to highlight the different sources of systematic uncertainty in these two approaches: the measurement of spatial cross correlations requires that the sample with unknown redshift information and the reference sample overlap spatially and cover the same redshift range. However, the spectroscopic calibration sample does not have to cover the same color/magnitude space as the unknown photometric sample. It is, however, important to accurately model the redshift-dependent galaxy-dark matter bias of the photometric sample and the spectroscopic calibration sample, since the redshift-dependent ratio between these two functions is completely degenerate with the photometric redshift distribution to be inferred. In contrast, template-based redshift inference requires a complete set of templates but no calibration sample. Checking a fitted model can also, in principle, use the color space alone, by comparing the photometry generated by the fitted templates with the measurements. In practice this approach has limitations. The generation of SED model templates is challenging and often requires spectroscopic reference data for some galaxies. Furthermore, degeneracies between galaxy type and galaxy redshift can make the aforementioned color-based approach ill-defined. Thus, while template fitting does not require spectroscopic data to infer redshifts of galaxies, in practise it is often necessary for building and evaluating models. Finally, empirical techniques that construct photometric redshift estimates by 'learning' from a spectroscopic calibration dataset require reference data that does not have to spatially overlap, but needs to be representative in color-redshift space.

Besides spatial correlations of galaxy clustering, we can also use other two-point statistics from e.g., weak gravitational lensing (e.g. Benjamin et al. 2013; Stölzner et al. 2020). There also exists a considerable literature in how photometric redshift uncertainty can be treated in the individual cosmological probes (McLeod et al. 2017; Hoyle & Rau 2019) or how one can combine template fitting and cross correlation measurements (Alarcon et al. 2020b; Sánchez & Bernstein 2019; Jones & Heavens 2019; Rau et al. 2020). Shortly before this paper was submitted for publication Myles et al. (2020); Gatti et al. (2020); Cawthon et al. (2020) presented the redshift inference scheme for the DES Y3 analyses, that combines a cross-correlation and shear ratio data vector with redshift information derived using an empirical mapping of broad band 'Wide field' photometry to spatially smaller calibration fields with narrow-band photometric and spectroscopic redshift information.

This paper presents a composite likelihood approach to jointly constrain photometric redshift distributions using information from both the available photometry and the clustering of galaxies. We focus on statistical challenges in this inference. In particular, the parts of the model that utilize the photometry of galaxies can pose computational challenges, since the likelihood depends on measurements of all galaxies in the sample. We therefore derive an efficient methodology that facilitates inference of redshift distributions within this computationally expensive part of the model. Redshift inference based on noisy photometry is an inverse problem and the inference scheme requires careful regularization to achieve good probability coverage. We therefore describe several regularization techniques and evaluate their respective merits in numerical experiments. Information from the spatial distribution of galaxies can then be incorporated within the composite likelihood framework by efficient MCMC sampling. We test our methodology using data from the CosmoDC2 (Korytov et al. 2019) simulated extragalactic catalog. While some of the inference techniques developed in this paper can also be used in the context of an empirical mapping to a small-area calibration field, our primary goal is to facilitate inference using physical SED modelling that utilizes a likelihood that jointly describes photometry and spatial information for all observed galaxies. Inference under spatial variations in photometry or redshift information will be addressed in the course of the paper and in § 10.

The paper is structured as follows: § 2 describes the simulated galaxy samples used in this work, while § 3 gives a brief introduction into inverse problems and deconvolution by discussing a simple toy model for photometric redshift inference. The following sections describe our inference methodology in detail: § 4 starts with a description of the photometric likelihood, where we also discuss several regularization schemes, and § 5 formulates the

cross-correlation likelihood. Both of these parts are then combined in a composite likelihood framework in § 6. § 7 discusses aspects of model evaluation and parametrization of systematics. We then apply our methodology to the simulated data in § 8. § 9 summarizes our findings. § 10 closes the paper with a discussion of future work.

2 SIMULATED GALAXY SAMPLES

We use data from the CosmoDC2 simulated extragalactic catalog (Korytov et al. 2019) in this work. CosmoDC2 is a mock extragalactic catalog based on a trillion particle N-body simulation with a box size of 4.225 Gpc^3, the 'Outer Rim' run (Heitmann et al. 2019). The simulated catalog covers 440 deg^2 of sky area and spans a redshift range $0 < z \leq 3$. Galaxies are assigned to the halo catalog and supplemented with additional galaxies based on the assumption of a power law extrapolation of a power law sub-halo mass function at lower masses. The resulting catalog exhibits a number count slope consistent with that of the Hyper SuprimeCam Deep survey (Aihara et al. 2018) down to an r-band magnitude of r $\sim$ 28, well beyond the apparent magnitudes that will be utilized in this paper. The galaxy catalog uses a combination of empirical and semi-analytic modelling, utilizing the Galacticus (Benson 2012) and GalSampler codes (Hearin et al. 2020). For more details on the catalog generation and properties we refer the reader to Korytov et al. (2019).

In § 2.1 we will describe the particular selection of photometric data and the photometric redshift catalog used in this work. § 2.2 describes the generation of the reference spectroscopic sample.

2.1 Photometric Sample and Photometric Redshift Catalog

The photometric sample consists of mock galaxies from the LSST-DESC "CosmoDC2" synthetic sky catalog (Korytov et al. 2019). The catalogs do not contain stars or AGN, so star-galaxy separation and non-thermal contamination are not an issue in this data set. Observations consist of magnitudes in the six *ugrizy* Rubin Observatory filters. Simulated photometric errors were added to the six bands using a simple model designed to match the expected photometric S/N due to depth, seeing, airmass, and sky brightness at the completion of the full 10-year Wide Fast Deep survey (Ivezić et al. 2019). All galaxies are assumed to be isolated, i.e. blending effects are not modeled. We restrict the sample to galaxies with an i_{LSST}-band magnitude of $i_{\mathrm{LSST}} < 25.0$ that corresponds to a point source i_{LSST}-band signal-to-noise (S/N) of $\sim$ 20. We make this cut because redshift estimates for lower S/N objects degrade rapidly below this S/N level. We reserve a small set of $\sim$ 100000 galaxies for training of the photo-z algorithms; this training set is a random subset of the $i_{\mathrm{LSST}} < 25.0$ sample, and thus completely representative of the underlying galaxy distribution, so no modeling of spectroscopic incompleteness effects is necessary.

Template Fitting Redshifts We use the publicly available Bayesian photometric redshift code BPZ[1] (Benítez 2000) to compute redshift estimates for our simulated galaxies. BPZ is a template-based redshift estimation code that estimates redshift by computing model fluxes from a set of template SEDs and evaluating the resulting χ^2 when compared to observed fluxes. BPZ includes the optional application of a bivariate Bayesian prior over the joint distribution of type/SED and apparent magnitude in the redshift estimation, though we do not employ the prior in this investigation.

To construct a template set we begin with the empirical SED catalog of Brown et al. (2014). We then use the ESP software package (Kalmbach & Connolly 2017), which constructs a principal component basis set from the empirical SEDs and uses photometric training data to construct the final SED template via Gaussian Processes. The final training set used by BPZ consists of the 129 empirical templates and 100 additional templates output from ESP. These templates roughly, but not perfectly, span the observed range of colors for the sample. We compute the likelihoods for all SEDs by comparing the observed fluxes to model fluxes evaluated on a grid of redshift spanning $0 < z < 3$. The 1-dimensional marginalized (over template type) posterior distributions for each galaxy comprise our final template fitting redshift estimate.

Machine Learning-based Redshifts We use the python version of the publicly available FlexCode[2] (Izbicki & Lee 2017) combined with the XGBoost algorithm (Chen & Guestrin 2016) to compute photometric redshifts which we will refer to by the name FlexZ-Boost. FlexZBoost estimates the conditional density in redshift for each galaxy by fitting to an orthonormal set of basis functions (in this case cosines) via regression with XGBoost. To further refine the estimates, 25 per cent of the training data is reserved as a validation set to determine optimal values for trimming extraneous low-level peaks in the likelihood, and a "sharpening" parameter of the form $p(z) \propto p(z)^{\alpha}$ that adjusts the overall width of the density estimates to best match the data. For this analysis we use 35 cosine basis functions, and a sharpening parameter, chosen via cross-validation, of 1.4. Given the representative training data used in this experiment, we expect very accurate redshift estimates from the FlexZBoost algorithm.

2.2 Spectroscopic Sample

The simulated reference spectroscopic sample is selected to mimic, in broad strokes, the sample selections of the Dark Energy Spectrosopic Instrument (DESI, DESI Collaboration et al. 2016, Zhou et al. 2020a, Zhou et al. 2020b). This consists of a set of four samples with increasing mean redshift: a magnitude-limited sample to $r_{\mathrm{LSST}} < 19.5$; a Luminous Red Galaxy (LRG) sample; an Emission Line Galaxy (ELG) sample; and finally a high-redshift Quasar (QSO) sample. We show the redshift and i_{LSST}-band magnitude distributions of these subsamples in Fig. 1. The LRG, ELG, and QSO samples are selected such that their density per redshift matches that of the DESI samples (priv. comm. Rongpu Zhou and Jeffrey Newman). This sample is distinct from the redshift calibration data mentioned in the previous section.

We construct a magnitude-limited sample, by imposing a magnitude cut of $r_{\mathrm{LSST}} < 19.5$. To approximate the LRG, ELG and QSO galaxy samples, we use the values of the stellar mass, star formation rate, and black hole mass times Eddington ratio as proxies for objects that are LRG, ELG, and QSO-like respectively. Our goal with these samples is to select galaxies that will have differing bias properties and mimic the complexities of the DESI sample in this regard, while matching the density and signal-to-noise we would expect with a DESI-like sample. We thus use these simple truth quantities from the simulation rather than recreate the full color selection of a true, simulated DESI sample. The QSO, ELG, and

[1] available at: `http://www.stsci.edu/~dcoe/BPZ/`

[2] available at `https://github.com/tpospisi/flexcode`

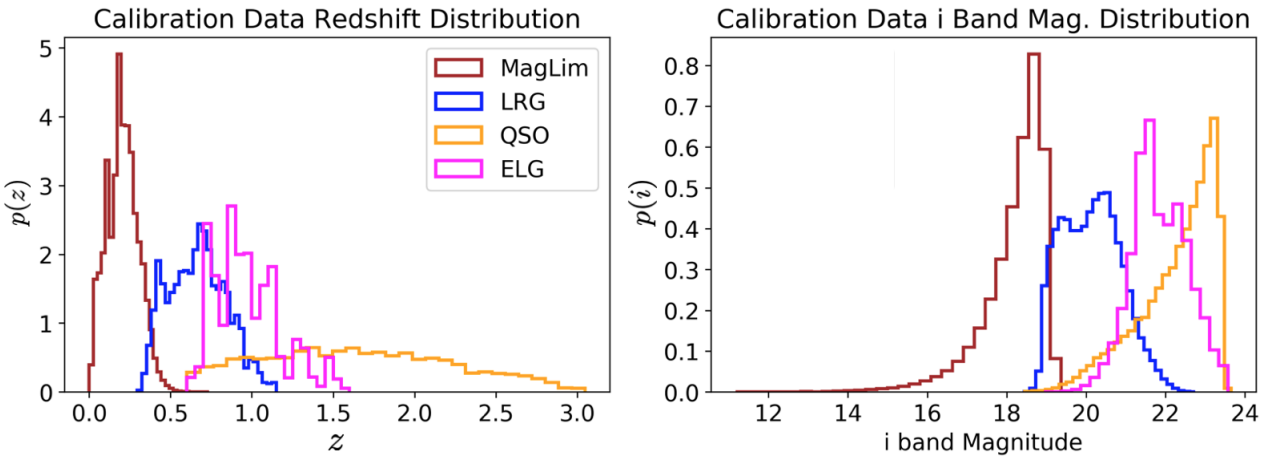

Figure 1. *Left:* Redshift probability density functions of the galaxy populations that constitute the DESI-like spectroscopic reference sample. *Right:* Corresponding i-band magnitude distributions.

LRG samples are all selected with $r_{\rm LSST} > 19.5$, to be independent of the magnitude-limited sample. Additionally, QSOs and ELGs are selected to have $r_{\rm LSST} < 23.4$ and LRGs have the cut $z_{\rm LSST} < 23.0$ applied to them. QSOs are selected by ordering the candidate QSOs in a redshift bin by the product of their black hole mass and black hole Eddington ratio, cutting on the value when the density of QSOs matches the expected DESI density for a given redshift range. This process is repeated for ELGs using their star formation rate in the simulation as a proxy for "ELG-ness". We also impose the condition that the candidate ELGs have a black hole mass times Eddington ratio below what we cut on for the QSOs, to assure that the samples are independent. This process of rank-ordering and selecting the top galaxies until we achieve the expected DESI density is repeated again for the LRGs, this time with stellar mass as our proxy value. Both the ELG star formation and QSO selection are excluded from the LRG selection, ensuring that the samples are independent. We calculate values for these cuts on a ∼50 deg^2 test area in the CosmoDC2 simulations and apply them to the full 300 deg^2 area.

3 INTRODUCTION TO DECONVOLUTION PROBLEMS

As we will see in detail in the following sections, the photometric redshift problem is a deconvolution problem, where the redshift distribution of a sample of photometrically observed galaxies is inferred from their noisy photometric measurements. To give the reader an intuitive understanding of deconvolution problems, we present a short introduction into the classical deconvolution problem. A similar description in the context of photometric redshift estimation can be found in Padmanabhan et al. (2005). We close this section by discussing the limitations of the toy model considered here and motivate the likelihood inference framework presented in the following.

3.1 A Toy Model

Consider three vectors of random variables $\mathbf{Z}$, $\mathbf{Z}^p$ and $\boldsymbol{\epsilon}$ with dimension $N_{\rm gal}$, which denotes the photometric sample size. $\mathbf{Z}$ and $\mathbf{Z}^p$ denote the true and photometric redshifts of the galaxies in the sample and $\boldsymbol{\epsilon}$ the residual error between both quantities. The additive noise model that connects these random variables is given as:

$$\mathbf{Z}^p = \mathbf{Z} + \boldsymbol{\epsilon}\,. \tag{1}$$

The probability densities[3] associated with these random variables are:

$$Z_j \sim p_z \tag{2}$$
$$Z_j^p \sim p_z^p \tag{3}$$
$$\epsilon_j \sim p_\epsilon\,, \tag{4}$$

where $j \in \{1, \ldots, N^{\rm gal}\}$ and '$\sim$' connects the realization of a random variable on the left hand side with the probability density function (PDF) on the right hand side from which this realization is drawn.

The random variable ϵ_j is assumed to be identically and independently distributed, as well as independent of Z_j. These assumptions do not hold in the photometric redshift scenario, as the noise very clearly depends on the color, and therefore redshift, of the galaxy. However, in the following toy model, we adopt these assumptions for simplicity. The theory can be easily extended towards input-dependent noise (see e.g. Meister 2009) without changing the intuition presented in this section.

In order to derive an estimator for p_z, we use the convolution theorem[4] that connects the PDF of the sum of independent random variables with the convolution of their densities. We can therefore write:

$$p_z^p = p_z * p_\epsilon = \int p_z(z^p - z)p_\epsilon(z)\mathrm{d}z = \int p_\epsilon(z^p - z)p_z(z)\,\mathrm{d}z\,, \tag{5}$$

[3] Note that the probability densities p_z and p_z^p are both redshift distributions of samples of galaxies. They differ since $(p_z^p\,,\, p_z)$ denotes the sample distribution of (photometric, true or spectroscopic) galaxy redshifts. Thus p_z^p would be broader, since the error in the redshift, drawn from p_ϵ, is convolved with the true redshift.

[4] A Fourier-based approach is not necessary. Concretely, the likelihood framework presented in the following section works in real space. A Fourier description for the classical deconvolution problem is, however, analytically tractable and provides a clear picture of the nature of the problem and the importance of regularization.

The Fourier transform of a probability distribution is the characteristic function. We will denote the characteristic functions of (p_z, p_z^p, p_ϵ) as $(p_z^{\mathrm{ft}}, p_z^{\mathrm{p,ft}}, p_\epsilon^{\mathrm{ft}})$. Given a sample drawn from a PDF, e.g., the sample of photometric redshifts of N_{gal} galaxies, we can estimate $p_z^{\mathrm{p,ft}}$ as[5]

$$\hat{p}_z^{\mathrm{p,ft}}(t) = \frac{1}{N_{\mathrm{gal}}} \sum_{j=1}^{N_{\mathrm{gal}}} \exp\left(itZ_j^p\right). \tag{6}$$

The argument t of the characteristic function could be interpreted as a kind of redshift-frequency if we treat the redshift of a galaxy as a 'time parameter'. Under the assumption of independence between $\mathbf{Z}$ and ϵ we can write:

$$p_z^{\mathrm{p,ft}}(t) = \hat{p}_z^{\mathrm{ft}}(t)\hat{p}_\epsilon^{\mathrm{ft}}(t) = p_z^{\mathrm{ft}}(t)p_\epsilon^{\mathrm{ft}}(t). \tag{7}$$

Therefore an estimator for $p_z^{\mathrm{ft}}(t)$ is given as:

$$\hat{p}_z^{f\,t}(t) = \frac{1}{p_\epsilon^{\mathrm{ft}}(t) N_{\mathrm{gal}}} \sum_{j=1}^{N_{\mathrm{gal}}} \exp\left(itZ_j^p\right), \tag{8}$$

where we assume $p_\epsilon^{\mathrm{ft}}(t)$ is known and nonzero everywhere. We note that this estimator is consistent and unbiased (Meister 2009). The error term $p_\epsilon^{\mathrm{ft}}(t)$ here acts as a 'filter' to weight down small scale modes in the distribution. However, we note that this term $1/p_\epsilon^{\mathrm{ft}}$ can become large when p_ϵ^{ft} is small.

As a consequence the inverse Fourier transform

$$p_z(z) = \frac{1}{2\pi} \int \exp\left(-itz\right) \hat{p}_z^{f\,t}(t)\,\mathrm{d}t\,, \tag{9}$$

is neither integrable nor square integrable. Loosely speaking this implies that the parameter space that describes the shape of $p_z(z)$ does not have to be bounded. We will see this effect also for the more complex model considered in the later sections of this work. We reiterate that while the estimator of $\hat{p}_z^{\mathrm{ft}}$ has very desirable properties, the inverse transformation is not well defined, hence deconvolution problems are part of a larger class of 'inverse problems'.

In order to obtain well-defined results, we therefore have to perform regularization either by regulating the shape/parametrization of p_z^p (e.g. using Kernel methods), projecting p_z onto a suitable basis like wavelet functions or by directly restricting the $1/p_\epsilon^{\mathrm{ft}}$ term, as implemented in a Ridge method (e.g. Meister 2009, § 2.2.3). We will not discuss the details of these methods and refer to the literature for a more detailed explanation (e.g. Meister 2009). It is, however, instructive to study the functional form of one of these regularized estimators. Making the ansatz of a kernel density estimate for the photometric redshift PDF, one can show that the deconvolved density $\hat{p}_z$ can be estimated as (e.g. Meister 2009):

$$\hat{p}_z(z) = \frac{1}{2\pi} \int \exp\left(-itz\right) \left(\frac{K^{\mathrm{ft}}(tb)}{p_\epsilon^{\mathrm{ft}}(t)}\right) \frac{1}{N_{\mathrm{gal}}} \sum_{j=1}^{N_{\mathrm{gal}}} \exp\left(itZ_j^p\right) \mathrm{d}t\,, \tag{10}$$

where b denotes the bandwidth and K^{ft} the fourier transform of the kernel function that enters the kernel density estimation ansatz for p_z^p. We see that by restricting the shape of the density p_z^p to a kernel density estimate whose smoothness is governed by the parameter b, we regularize the $1/p_\epsilon^{\mathrm{ft}}(t)$ term by a multiplicative factor, that renders the inverse Fourier transformation both integrable and square integrable assuming bounded, compactly supported and non-vanishing K^{ft}. The bandwidth parameter governs the tradeoff between the bias, or 'smoothness', of the density estimate, and its variance. Choosing a larger bandwidth washes out small scale noise in the reconstructed density. In the limit of vanishing bandwidth, we would again obtain an ill-posed inverse Fourier transformation.

In the following sections, we will apply regularization techniques that restrict the functional form of the redshift distribution, following a similar idea as presented in closed form in Eq. (10) for the classical deconvolution problem.

3.2 Towards the Photometric Redshift Problem

We note that inverse problems like the classical deconvolution problem also appear in several other scenarios like the measurements of shapes, where the point-spread function (PSF) of galaxies convolves the galaxies' light profiles and leads to a loss of information. While the considered toy model of the photometric redshift problem is analytically tractable, it does not describe the realistic situation. Besides the relatively simple extension towards a galaxy-dependent photometric noise, the noise distribution p_ϵ is, in photometric redshift estimation, given as a joint likelihood between the photometry of all galaxies in the sample, that depends on the additional parameters that enter the SED modelling. The redshift of each galaxy is a parameter that enters its likelihood and the sample redshift distribution is its prior. Furthermore, the model does not properly account for the spatial distribution of galaxies. The clustering of galaxies does not only constrain their redshift distribution, but connects to SED modelling, with nuisance parameters that describe, e.g., galaxy-dark matter bias.

We structure the discussion of the likelihoods used in this work in practice based on the following roadmap: we first describe our likelihood framework for the photometry of galaxies given a set of templates in § 4. We reiterate that in contrast to the classical deconvolution problem, the 'error distribution' in the full photometric redshift problem is based on the joint photometric likelihood. We will therefore base our estimator and inference on this likelihood instead of the characteristic function. Despite these methodological differences, we note that the necessity for regularization is the same as in the analytically tractable classical deconvolution problem. The considered regularization schemes are described in § 4.2. A particular challenge in the context of large area photometric surveys is the necessity to scale the inference to a large number of galaxies. In Appendices § A and § B we derive an efficient inference framework based on the Laplace approximation that facilitates fast probabilistic deconvolution. We will use this deconvolution methodology in the following sections § 4 to § 8.

4 PHOTOMETRIC LIKELIHOOD

The spectral energy distributions of distant galaxies are a complex superposition of spectral components from their stellar populations.

The SED of the galaxy can be uniquely mapped to a given redshift z, which allows us to predict the galaxy flux as a function of redshift in a given optical filter band $\mathcal{F}(\lambda)$ by

$$f_i(z, \alpha) = \int \mathcal{F}(\lambda)\, \mathrm{SED}_\lambda(\lambda, z, \alpha)\, \mathrm{d}\lambda\,. \tag{11}$$

where $\mathrm{SED}_\lambda(\lambda, z, \alpha)$ is the Spectral Energy Distribution template in units of $\mathrm{erg}/\mathrm{cm}^2/\mathrm{s}/\text{Å}$. The parameter α denotes additional free parameters in the SED template models, such as galaxy age, type, or red continuum slope. For a given set of photometric filters $\mathcal{F}(\lambda)$ we

[5] Here, $\hat{}$ denotes an estimator for the respective function.

obtain a mapping between the redshift z of the galaxy and a vector of fluxes $\mathbf{f}$. We will denote this mapping as $\mathcal{T}(z, \alpha)$.

Assuming that the measurements of photometry for different galaxies are independent[6] we can make the ansatz for the joint likelihood of fluxes of a galaxy sample $\hat{\mathbf{F}}$

$$p(\hat{\mathbf{F}}|\mathbf{z}, \boldsymbol{\alpha}) = \prod_{i=1}^{N_{\rm gal}} \mathcal{N}(\hat{\mathbf{f}}_\mathbf{i}|\mathcal{T}(z_i, \alpha_i), \boldsymbol{\Sigma}_i) \,. \tag{12}$$

Here, $\boldsymbol{\Sigma_i}$ denotes the measurement covariance matrix of the flux measurements $\hat{\mathbf{f}}_\mathbf{i}$, and $\hat{\mathbf{F}}$ denotes the set of all flux measurements of the galaxies. We assume Gaussian uncertainties here, where $\mathcal{N}(x, \mu, \Sigma)$ denotes the Normal distribution. The parameter α can either be a galaxy-specific index that selects a certain template from a pre-specified number of models, or a physical property of galaxies.

The prior on the parameters z and α must account for their correlation. An example for a possible parametrization in the case of a galaxy-specific template index would be a two-dimensional histogram. However, other parametrizations are possible, especially if additional parameters that change the shape of the base templates are included in the template set. In this work, we will consider the simplest case, where we use a multidimensional histogram prior where each histogram cell denotes a combination of redshift bin and discretized α parameter value, that for example could indicate a template selection. The histogram index i runs over all histogram bins $\{i : 0 < i \leq N_{\rm tot}\}$, where $N_{\rm tot} = N_{\rm bins} \times N_{\rm parameters}$. The prior on the corresponding histogram heights, denoted as n_i^B corresponding to the interval I_i in the $z - \alpha$ parameter space, reads:

$$p(z, \alpha) = \sum_{i=1}^{N_{\rm tot}} n_i^B \, [(z, \alpha) \in I_i] \,. \tag{13}$$

Here $[K]$ denotes the Iverson bracket, that is (0, 1) if the proposition K is (false, true). We note that $\mathbf{n^B}$ parametrizes the joint distribution of redshift histograms and $\boldsymbol{\alpha}$ parameter. For simplicity we will in the following omit the marginalization over $\boldsymbol{\alpha}$ and refer to $\mathbf{n^B}$ as the parameters of the sample redshift distribution. The reason is that in this paper we do not add additional parameters to parametrize the SEDs over which we need to marginalize. In applications like weak gravitational lensing and galaxy clustering we are mainly interested in estimating the redshift distribution of a sample of galaxies, here referred to as the base sample and parametrized by the vector $\mathbf{n}^{\rm B}$. It is therefore useful to marginalize over the redshifts of individual galaxies. We note that if the posterior of individual galaxy redshifts is important, we can always post-sample using the final posterior on $\mathbf{n}^{\rm B}$, based on Eq. (34), that then also includes information from galaxy clustering. The posterior distribution of the sample redshift distribution given $\hat{\mathbf{F}}$ is then:

$$p(\mathbf{n}^{\rm B}|\hat{\mathbf{F}}) \propto p(\mathbf{n}^{\rm B}) \prod_{i=1}^{N_{\rm gal}} \int_0^\infty {\rm d}z_i \, p(z_i|\mathbf{n}^{\rm B}) \, \mathcal{N}(\hat{\mathbf{f}}_\mathbf{i}|\mathcal{T}(z_i), \Sigma_i) \,. \tag{14}$$

Discretizing the integral and using Eq. (13) we obtain

$$p(\mathbf{n}^{\rm B}|\hat{\mathbf{F}}) \propto p(\mathbf{n}^{\rm B}) \prod_{i=1}^{N_{\rm gal}} \sum_{j=1}^{N_{\rm tot}} n_j^{\rm B} \int_{z_L^j}^{z_R^j} {\rm d}z_i \, p(\hat{\mathbf{f}}_\mathbf{i}|\mathcal{T}(z_i), \Sigma_i) \,. \tag{15}$$

The histogram heights $n_i^B = \pi_i^B / \Delta z$ can be expressed as the ratio between $\boldsymbol{\pi}$ and the histogram width Δz, assuming equal-sized redshift bins. The vector $\boldsymbol{\pi^B}$ has the properties $\sum_{i=1}^{N_{\rm tot}} \pi_i^B = 1$ and $0 \leq \pi_i^B \leq 1$, and therefore lies on the simplex. Our first choice for a distribution on the simplex for $p(\boldsymbol{\pi^B})$ (and therefore $p(\mathbf{n^B})$) was the Dirichlet distribution[7]. During the course of this project we have applied a mean field variational inference scheme that uses the Dirichlet as the variational distribution as well as a Gibbs sampling scheme based on the Dirichlet-Multinomial cojugacy for posterior inference. We found that the variational inference scheme yielded underestimated error bars, likely due to the restricted covariance structure of the dirichlet. Moreover, the sampling approach did not scale well to the large galaxy samples expected for the first-year LSST observations. Specifically, the computational workload to update redshift variables for $10^6 - 10^{10}$ galaxies seems very large, and while subsampling techniques provide a possible mitigation, they can lead to biased inferences (Quiroz et al. 2018). Furthermore the application of sampling techniques requires a sufficiently large trace to ensure convergence. This can be difficult to ensure in this case. In order to provide a more flexible distributional ansatz than the dirichlet, while still maintaining the computational advantages of a mean field variational inference scheme, we decided to develop a scheme that is based on the logit-normal distribution (Atchison & Shen 1980), as explained in the following section. While these considerations motivate our choice of method, we note that this should not discredit alternative approaches based on sampling or variational inference in general. We will perform a more detailed analysis of convergence and probability coverage of multiple inference techniques in future work.

4.1 Photometric Redshift inference

The problem specified by Eq. (14) is a deconvolution problem that extends the simple toy model considered in § 3. The 'noise' PDF is now given by a joint likelihood that can depend on a complex set of parameters. Furthermore, while the discussion in § 3 focused on deriving an estimator for the deconvolved density, the focus here is to infer posteriors using efficient inference techniques. We present the detailed description and derivation of the inference pipeline in Appendices A and B. The final form of Eq. (14) is then given in the form of a logit-normal posterior:

$$p(\mathbf{n^B}|\hat{\mathbf{F}}) \approx \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_\mathbf{y}|}} \frac{1}{\Delta z^{N_{\rm bins}} \prod_{i=1}^{N_{\rm bins}} n_i^B} \exp\left(-\frac{1}{2}\left(\log\left(\frac{\boldsymbol{n^B}_{-{\rm N_{bins}}}}{n^B_{N_{\rm bins}}}\right) - \boldsymbol{\mu}_{\rm y,ML}\right) \boldsymbol{\Sigma}_\mathbf{y}^{-1} \left(\log\left(\frac{\boldsymbol{n^B}_{-{\rm N_{bins}}}}{n^B_{N_{\rm bins}}}\right) - \boldsymbol{\mu}_{\rm y,ML}\right)\right), \tag{16}$$

where Δz denotes the histogram bin width. The estimation of the covariance $\boldsymbol{\Sigma}_\mathbf{y}$ and mean vector $\boldsymbol{\mu}_{\rm y,ML}$ are detailed in Appendices A and B. However, we note that this formalism derives the hessian $\mathbf{H} = -\boldsymbol{\Sigma}_\mathbf{y}{}^{-1}$ and obtaining the covariance matrix $\boldsymbol{\Sigma}_\mathbf{y}$ requires matrix inversion. The subscript 'y' here refers to the variable transformation:

$$\mathbf{y}(\boldsymbol{\pi}) = \left[\log\left(\pi_1/\pi_{N_{\rm bins}}\right), \ldots, \log\left(\pi_{N_{\rm bins}-1}/\pi_{N_{\rm bins}}\right)\right] \,, \tag{17}$$

[6] This assumption can be violated due to effects such as blending of nearby galaxy light profiles on flux calibration errors..

[7] The Dirichlet is the conjugate prior to the multinominal distribution, which can make sampling and inference easier. Concretely, if a Dirichlet prior is set on the probabilities of the multinomial likelihood (which are its parameters), the posterior over these probabilities is again a Dirichlet. However conjugacy does not imply that the prior is ideal in all circumstances.

that we discuss in more detail in Appendix B. The vector $\mathbf{n^B}_{-N_{\rm bins}}$ denotes here the vector of $\mathbf{n^B}$ excluding the last entry $n^B_{\rm N_{bins}}$, where we assume equal sized redshift histogram bin width.

Like the classical deconvolution problem, the inference of Eq. (14) is an inverse problem. We can therefore expect that there exists a parameter vector $\boldsymbol{\pi} \in \Delta$, where Δ denotes the simplex space, that has a high likelihood (relative to the maximum likelihood value), but a large distance from the true $\boldsymbol{\pi}_{\rm opt}$.

Furthermore, as we have seen in § 3, the solutions of inverse problems do not have to be bounded[8] (or even well defined), which implies that uncertainties can be arbitrarily large (see, e.g., Kuusela 2016). Regularization, detailed in the following section, is therefore a key aspect in our inference pipeline.

4.2 Regularization

In this section we describe techniques that we employ to regularize[9] the deconvolution problem. As instabilities arise when the histogram width is of the same order as the uncertainty in the redshift of the individual galaxies, picking broader bins reduces these artifacts (see e.g. Kuusela 2016). Considering our toy model in Eq. (10), we see that if $K^{\rm ft}(tb)$ is narrower than $p^{\rm ft}_{\epsilon}(t)$, there can be values of t where their ratio, and therefore the integrand in Eq. (10), can become large or even unbounded. Subject to the aforementioned limitations, this behaviour generalizes to the deconvolution problem considered here. In the following, we will denote this scheme as the 'Wide Bin' method. As will be seen in the following section, this simple scheme can lead to posteriors that can be biased and too narrow. We therefore consider alternative approaches.

4.2.1 *Merging Bin Regularization*

The 'Merging Bin Regularization' scheme (Kuusela 2016) uses a very thin initial histogram binning. This will likely result in the aforementioned typical instabilities of inverse problems, but can avoid biases of the Wide-Bin regularization (or other smoothing) schemes. However, we must ensure that the optimization of the maximum-likelihood solution converges to a global maximum. We therefore run multiple optimizations with different initial conditions and pick the best solution. Furthermore, the hessian $\mathbf{H}$ can have a very high condition number. Even though it is possible to sample from the resulting posteriors without the matrix inverse using MCMC sampling (only the inverse covariance enters the χ^2), sampling is more efficient using the standard Box-Mueller method (Box & Muller 1958), which requires an inverse.

Tikhonov Regularization We perform the matrix inversion of the hessian $\mathbf{H}$ using Tikhonov regularization (see e.g. Kress 1998, pp. 86-90). Here, we treat the matrix inversion as a system of linear equations constructed from the hessian and the column-wise inverse hessian/unit matrix respectively. The instability of the problem can lead to very small entries in the hessian that imply large entries (in absolute values) in its inverse. As a regularization, one can add a penalty term and reformulate the problem as a minimization

$$\min \left\{ ||\mathbf{H}\mathbf{a}_i - \mathbf{1}_i||_2^2 + ||\Upsilon \mathbf{a}_i||_2^2 \right\} , \tag{18}$$

where i denotes column i of the inverse hessian and the unit matrix respectively. The matrix $\Upsilon = \alpha \mathbf{1}$ is the Tikhonov matrix, α the regularization parameter and $\mathbf{1}$ the unit matrix. The regularization term penalizes large values for $\mathbf{a_i}$, which regularizes the inversion and reduces the condition number. The analytic solution to this minimization problem is given as:

$$\mathbf{c}_i = \left(\mathbf{H}^{\rm T}\mathbf{H} + \Upsilon^{\rm T}\Upsilon \right)^{-1} \mathbf{H}^{\rm T}\mathbf{1_i} . \tag{19}$$

We note that we introduce the Tikhonov regularization here predominantly as a way to regularize the matrix inversion of the hessian. We recommend selecting the parameter α to be just as large as necessary to perform this inversion accurately. Tikhonov regularization can be used as the main regularization in inverse problems; however, we find that the merging bin regularization scheme performs much better in terms of producing well-calibrated probability $\mathbf{n^B}$ posteriors. We reiterate that the idea of the merging bin regularization scheme proposed by Kuusela (2016) is to deliberately start with histogram bins that are too small and lead to a noisy deconvolved density. We then exploit the characteristic noise pattern in the deconvolved distribution, where bins that overshoot, i.e. are larger than the true value, are immediately followed by those that undershoot. This results in an alternating or 'zig-zag' pattern of the deconvolved density. Merging these neighboring bins then helps to 'stabilize' the deconvolved distribution. We therefore sample from a posterior obtained assuming a finely binned histogram and merge neighboring bins, which compensates the noise effect. We can then in principle directly use these samples in the cross-correlation likelihood. Nonetheless, it is computationally more efficient to remap these samples to a regular grid with the same or very similar resolution than the binning used for the cross-correlations, since treating the finely binned histogram heights as free parameters would not

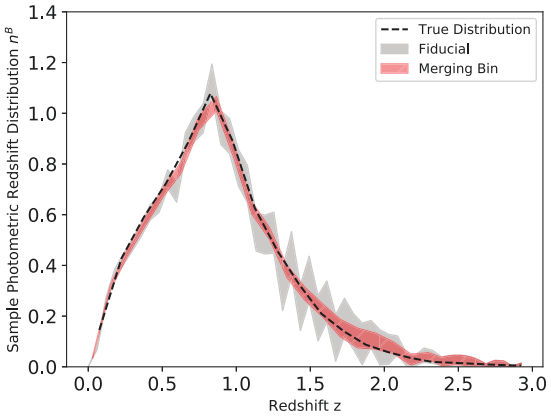

Figure 2. Illustration of the impact of the 'Merging Bin' regularization on the posterior of the sample photometric redshift distribution. We show 1σ intervals. The x-axis shows the redshift value z, the y-axis the value of the $\mathbf{n^B}$ parameters. The errorbars are the [16, 84] percentiles, which would correspond to 1σ intervals for a normal distribution. The black dashed curve shows the spectroscopic redshift distribution. The red contour shows the result of the 'Merging Bin' regularization with 30 bins applied to the 'Fiducial' contours that are binned using 50 bins. We refer for a detailed explanation to § 8 and Fig. 7, which shows and discusses the 'Fiducial' case as 'Small Sample (50k)'.

[8] In our case we note that all parameter values $\boldsymbol{\pi} \in \Delta$ are bounded, because Δ (with a chosen metric) is a bounded metric space. However these solutions in logit space (see § B) do not have to be bounded.

[9] We will use the term 'regularization' not only in the context of Bayesian statistics, where it's often implemented in the form of a prior, but in general to describe methods that restrict the complexity of parameters or functions.

add more information due to the resolution loss. We found that the template fitting posterior after merging bin regularization can again be well-described by a logit normal distribution that we fit using Assumed Density filtering.

Assumed Density Filtering To fit the logit normal distribution to the sampled and merged posterior samples, we work in logit space and make a gaussian ansatz

$$q(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ . \tag{20}$$

We can directly generate samples from the true distribution by sampling from the original, finely binned, logit-normal distribution with subsequent merging, i.e. averaging, of neighboring bins, and then transforming back to logit space. We will denote this true distribution as $p_{\rm true}(\mathbf{y})$. Assumed density filtering then commences by minimizing the Kullback-Leibler divergence between our ansatz $q(\mathbf{y})$ and $p(\mathbf{y})$,

$$\mathrm{KL}(p||q) = \int \mathrm{d}\mathbf{y} \ (p(\mathbf{y}) \log p(\mathbf{y}) - p(\mathbf{y}) \log q(\mathbf{y})) \ . \tag{21}$$

After optimizing $KL(p||q)$ for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we can show that the optimium is reached if

$$\int \mathrm{d}\mathbf{y}\, p(\mathbf{y})\, \mathbf{y} = \boldsymbol{\mu} \,, \tag{22}$$

and

$$\boldsymbol{\Sigma} = \int \mathrm{d}\mathbf{y}\, p(\mathbf{y})\, (\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}} \tag{23}$$

We see that assumed density filtering reduces to moment matching in logit space, when we apply the sample mean and sample covariance estimators to samples from $p_{\rm true}(\mathbf{y})$. We note that this is in general true for distributions of the exponential family (see, e.g., Ranganathan 2004).

To summarize, we perform the inference scheme described in the previous section using a fine histogram binning. Subsequently we sample from the posterior after regularizing the inverse hessian using Tikhonov regularization. We merge neighboring bins from each posterior draw until the noise is reduced and we obtain a smooth probability distribution. We illustrate this process in Fig. 2, which illustrates the impact of the 'Merging Bin' regularization scheme on the posterior of the sample photometric redshift distribution. Comparing the grey contours ('Fiducial') that uses 50 bins with the red contours ('Merging Bin') that merges neighboring bins to a binning of 30 bins, we see the much smoother shape and the elimination of the 'zig-zag' pattern present in the grey contours. We defer a more thorough explanation of the methodology and sample to § 8 and Fig. 7, which discusses the result shown in the grey contours under the abbreviation 'Small Sample (50k)'.

Based on our experience we propose to initially merge neighboring bins until we obtain a bin size of the order of the average $\pm 2\sigma$ range of the individual galaxy redshift distributions. Subsequently we merge fewer bins until the aforementioned characteristic 'zig-zag' noise pattern appears. This can be identified as the limiting resolution we can obtain. We note that it is important to distinguish patterns due to 'real' line of sight structure and due to the aforementioned noise in the deconvolution. If the pattern appears gradually with increasing resolution (merging fewer bins), it is indicative of statistically significant line-of-sight structure. If the deconvolved density suddenly becomes unstable in a 'zig-zag' pattern when fewer bins are merged, we have reached a resolution limit. Using assumed density filtering under the ansatz of a logit normal distribution, we finally reparametrize our model on the final redshift grid.

The merging process described above is largely based on inspecting when the instabilities vanish. There are certainly more principled alternatives. In a classical deconvolution problem, like the one presented in § 3, one could use a bootstrap estimate of the bias and variance of the reconstruction with respect to the oversmoothed photometric redshift distribution. This is consistent with the approach taken by Padmanabhan et al. (2005) based on the recommendation in Craig & Brown (1986). We use a joint likelihood between the photometry and spatial information to produce posteriors for the photometric sample redshift distribution and not a 'point prediction'. Furthermore our 'measured data' is the photometry and spatial information of galaxies. Accordingly our model selection, of which regularization is a part, must reproduce the measured photometry and spatial distribution, e.g., measured by the correlation functions of galaxies. In the Bayesian context, this would translate into the usage of posterior predictive checks (PPC) discussed in § 7.1. While the path to development of a more principled selection of the hyperparameters of regularization is known, it will require a thorough investigation. We defer this to future work, using the aforementioned 'rule-of-thumb' methodology as an interim solution.

5 CLUSTERING LIKELIHOOD

In order to include information about the spatial distribution of galaxies into the likelihood, we consider spatial cross-correlations between photometric and spectroscopic samples. Spatial correlations measure the excess probability over random to find two galaxies separated by a certain distance. This can be exploited to extract redshift information for galaxy samples (see e.g. Newman 2008; Ménard et al. 2013; McQuinn & White 2013; Scottez et al. 2016; Raccanelli et al. 2017; Morrison et al. 2017; Davis et al. 2017; Gatti et al. 2018) for which we do not have accurate redshift information, i.e. photometric galaxy samples, using spatially overlapping spectroscopic catalogs.

The idea is to select the spectroscopic samples in thin redshift slices and estimate the cross-correlation between these redshift-selected samples and the full photometric galaxy sample. As discussed in the following, the resulting signal will then be proportional to the photometric redshift distribution at that redshift.

Fig. 3 illustrates the basic idea of cross-correlation redshift inference. We consider two galaxy samples: a reference sample 'R' and a base galaxy sample 'B'. The reference sample contains galaxies with accurate, often spectroscopic, redshift measurements; the base galaxy sample consists of galaxies observed in broad band photometric filters. As the base/reference samples are typically photometric/spectroscopic samples, we use these terms interchangably in text[10]. The redshift distribution of the base sample is illustrated by the red distribution, while the binned reference sample redshift distributions for simplicity are shown as tophat functions (unlike the simulated samples we use to test our methodology). A single cross-correlation is then obtained by cross-correlating a single tophat selection with the full base sample. Multiple measurements

[10] We note, however, that the reference sample does not have to be a spectroscopic dataset, as multi-band, narrow filter photometric observations (Alarcon et al. 2020a), or photometric redshifts of redMaGiC samples (see e.g. Gatti et al. 2018), also allow for reasonable redshift accuracy.

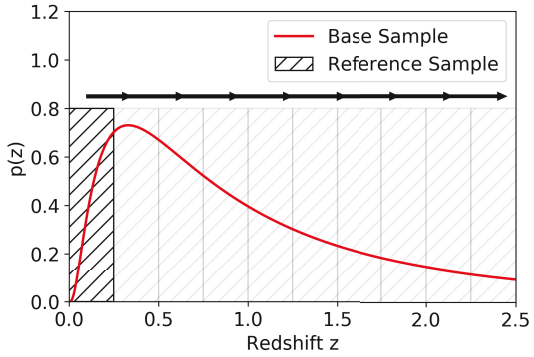

Figure 3. Illustration of the construction of the cross-correlation data vector. A 'Reference Sample' that can be precisely selected in redshift is moved over a 'Base Sample', which can either be a sample without redshift information (photometric sample) or the 'Reference Sample' itself. In this illustration 10 cross-correlations would be estimated, constituting the cross-correlation data vector.

therefore 'slice' through the redshift distribution of the base sample, illustrated here by the arrows and the grey hatched tophat slices.

As described in detail in Schmidt et al. (2013), Morrison et al. (2017) and Ménard et al. (2013), we measure the over-density, compared with a spatially random distribution of points, of photometric galaxies around each galaxy in the spectroscopic sample, within an annulus of physical scale $\Delta\chi = [\chi_{\rm min}, \chi_{\rm max}]$. The theoretical model for a cross-correlation function between the spectroscopic reference sample in tophat bin i and the photometric base sample is given as:

$$w_i^{\rm RB} \propto b_i^R b_i^B \left(\frac{\pi_i^{\rm B}}{z_H^i - z_L^i}\right) \overline{w}_{\rm DM,i}\,. \tag{24}$$

Here, b_i^R and b_i^B denote the value of the redshift-dependent galaxy-dark matter bias of the reference (R) and base (B) samples. The normed histogram bin heights of the base, or photometric, sample redshift distribution are denoted as π_i^B, where $\sum_i \pi_i^B = 1$, and the size of a redshift bin is given as $z_H^i - z_L^i$. The term $\overline{w}_{\rm DM,i}$ denotes the contribution of dark-matter clustering to the cross-correlation signal, which depends on the cosmological model.

We see that the modelling of the cross-correlation signal depends on the product of two redshift-dependent galaxy-dark matter bias functions that are completely degenerate with the set of parameters $\pi^{\mathbf{B}}$ that parametrizes the redshift distribution of the base sample. Furthermore, since $\overline{w}_{\rm DM,i}$ depends on the cosmological model, it will be computationally expensive to sample over these parameters.

To reduce the impact of the cosmological model, we want to combine the cross-correlations $\mathbf{w}^{\rm RB}$ with the correlations $\mathbf{w}^{\rm RR}$ of the spectroscopic sample. We therefore correlate the spectroscopic sample with itself in a manner analogous to what was just described, i.e., by correlating tophat selected spectroscopic samples with the full spectroscopic sample. The corresponding theory prediction then reads

$$w_i^{\rm RR} \propto (b_i^R)^2 \left(\frac{\pi_i^R}{z_H^i - z_L^i}\right) \overline{w}_{\rm DM,i} \tag{25}$$

where π_i^R is the normalized histogram height of the spectroscopic (full) sample redshift distribution. Both correlation function measurements ($\hat{\mathbf{w}}^{\rm RB}$ and $\hat{\mathbf{w}}^{\rm RR}$) just described are assumed to individually follow a Gaussian likelihood[11] .

Based on these definitions and approximations and the considerations in the previous section, we construct a likelihood based on the ratio between $\hat{\mathbf{w}}^{\rm RB}$ and $\hat{\mathbf{w}}^{\rm RR}$. Under the assumption of a diagonal covariance matrix[12] for $\hat{\mathbf{w}}^{\rm RB}$ and $\hat{\mathbf{w}}^{\rm RR}$, we can construct the random variable $\mathbf{\Gamma}$ for bin i with components

$$\Gamma_i^{\rm meas} = \left(\frac{\hat{w}_i^{RB}}{\hat{w}_i^{RR}}\right). \tag{26}$$

We reiterate that both $\hat{w}_i^{RB}$ and $\hat{w}_i^{RR}$ are described by a Gaussian Likelihood. Their respective means and standard deviations are given as $\mu_{\rm RB,i}, \mu_{\rm RR,i}, \sigma_{\rm RB,i}, \sigma_{\rm RR,i}$ respectively. The theoretical prediction for the transformed random variable $\Gamma_i^{\rm meas}$ is then

$$\Gamma_i^{\rm theo}(b_i^B, b_i^R, \pi_i^B, \pi_i^R) = \frac{b_i^B}{b_i^R}\frac{\pi_i^B}{\pi_i^R}\,, \tag{27}$$

and its likelihood:

$$\begin{aligned} p(\Gamma_i^{\rm meas}|\Gamma_i^{\rm theo}) = {} & \frac{b(\Gamma_i^{\rm theo})d(\Gamma_i^{\rm theo})}{a^3(\Gamma_i^{\rm theo})}\frac{1}{\sqrt{2\pi}\sigma_{\rm RB,i}\sigma_{\rm RR,i}}\times \\ & \left(\Phi\left(\frac{b(\Gamma_i^{\rm theo})}{a(\Gamma_i^{\rm theo})}\right) - \Phi\left(-\frac{b(\Gamma_i^{\rm theo})}{a(\Gamma_i^{\rm theo})}\right)\right) \\ & + \frac{1}{a^2(\Gamma_i^{\rm theo})\pi\sigma_{\rm RB,i}\sigma_{\rm RR,i}}\exp\left(-\frac{c}{\Gamma_i^{\rm theo}}\right). \end{aligned} \tag{28}$$

Here $\Phi(z)$ denotes the cumulative distribution function of the zero mean unit variance normal distribution and

$$a(\Gamma_i^{\rm theo}) = \sqrt{\frac{1}{\sigma_{\rm RB,i}^2}(\Gamma_i^{\rm theo})^2 + \frac{1}{\sigma_{\rm RR,i}^2}} \tag{29}$$

$$b(\Gamma_i^{\rm theo}) = \frac{\mu_{\rm RB,i}}{\sigma_{\rm RB,i}^2}\Gamma_i + \frac{\mu_{\rm RR,i}}{\sigma_{\rm RR,i}^2} \tag{30}$$

$$d(\Gamma_i^{\rm theo}) = \exp\left(\frac{b(\Gamma_i^{\rm theo})^2 - ca(\Gamma_i^{\rm theo})^2}{2a(\Gamma_i^{\rm theo})^2}\right) \tag{31}$$

$$c = \frac{\mu_{\rm RB,i}^2}{\sigma_{\rm RB,i}^2} + \frac{\mu_{\rm RR,i}^2}{\sigma_{\rm RR,i}^2}\,. \tag{32}$$

For the following discussion we define $\mathbf{n} = \frac{\boldsymbol{\pi}}{z_H - z_L}$, i.e. the variables $\boldsymbol{\pi}$, $\mathbf{n}$ refer to histogram heights normalized to sum to unity and to unit area respectively. The variable $\mathbf{n}^{\rm B}$ and $\mathbf{n}^{\rm R}$ refer to the histogram heights of the base and reference samples. This likelihood assumes independence between neighboring redshift bins. We can, however, expect a degree of correlation especially for lower redshift bins due to magnification effects.

Eq. (28) is approximately independent of the modelling of the $\overline{w}_{\rm DM}$ term, assuming that we pick sufficiently thin redshift bins to 'divide-out' the redshift-dependence of the dark matter clustering term $\overline{w}_{\rm DM}$. Based on the aforementioned independence assumption

[11] In reality, we expect that the likelihood will deviate from the Gaussian assumption (see e.g. Hahn et al. 2019).

[12] While the covariance matrix will be dominated by the diagonal, we can expect, that the cross-correlation measurements in different bins will be correlated. Thus the assumption of a diagonal covariance matrix is an approximation.

between redshift bins, the joint likelihood for all bins i now reads:

$$p(\mathbf{\Gamma}|\mathbf{b^b}, \mathbf{b^R}, \mathbf{n^R}, \mathbf{n^B}) = \prod_{i=1}^{N_{\rm bins}} p(\Gamma_i|b_i^B, b_i^R, n_i^R, n_i^B) \tag{33}$$

The function that describes the set of ratios b_i^B/b_i^R will be denoted as $C(z, \Delta\chi_\perp)$ and depends both on redshift and the size of the annulus $\Delta\chi$. For a selected annulus size we will use the abbreviation $C(z)$.

6 THE COMPOSITE LIKELIHOOD

To formulate a joint likelihood for the data vector of both galaxy positions and photometry, we use the composite likelihood ansatz (e.g. Varin et al. 2011) that uses the product of marginal likelihoods for both the photometry $\hat{\mathbf{F}}$ and the vector of cross-correlation functions $\mathbf{\Gamma}$:

$$p(\hat{\mathbf{F}}, \mathbf{\Gamma}|\bar{\mathbf{n}}^{\rm B,sys}, \bar{\mathbf{n}}^{\rm B}, \bar{\mathbf{n}}^{\rm R}, \mathbf{C}(\mathbf{z})) = p(\hat{\mathbf{F}}|\bar{\mathbf{n}}^{B})^{\upsilon_1}\, p(\mathbf{\Gamma}|\bar{\mathbf{n}}^{\rm B,sys}, \mathbf{C}(\mathbf{z}))^{\upsilon_2}\,, \tag{34}$$

where $\boldsymbol{\upsilon}$ are weights that can be selected to improve the efficiency of the estimation (see e.g. Varin et al. 2011) by increasing the influence of one part of the composite likelihood over the other. Furthermore the composite likelihood can be conditioned on auxiliary parameters such as the field. For simplicity we consider here only the simple case of $\boldsymbol{\upsilon} = 1$ and refer for an additional discussion to § 10.

We note that measurements of LSS and weak lensing often use galaxy samples that are selected by increasing redshift, to form tomographic bins. This analysis methodology can be incorporated into Eq. (34) by replacing $\hat{\mathbf{F}}$ and $\mathbf{\Gamma}$ with the joint data vectors of the selected galaxy samples, which would include covariances between the $\mathbf{\Gamma}$ measurements for different tomographic bins. Furthermore the quality and number of available photometric bands can change for different spatial areas. Similarly we need to construct a joint data vector of $\hat{\mathbf{F}}$ and $\mathbf{\Gamma}$ that incorporates these covariances. This can be modelled either analytically (e.g. Stoyan & Stoyan 1994; Sánchez et al. 2020) or by using spatial resampling techniques. In this work we will concentrate on the composite likelihood as given in Eq. (34) and refer extensions of the method to future work.

7 MODEL EVALUATION AND PARAMETRIZATION OF SYSTEMATICS

Parameter inference is only a single step in a full statistical analysis and needs to be combined with additional analysis steps. We need to ensure that parameters can be uniquely inferred and the posterior does not exhibit flat regions or strong degeneracies, which can make the application of MCMC techniques difficult (see e.g. Rothenberg 1971; Raue et al. 2013). Furthermore, one needs to investigate the sensitivity of the results against changes in the prior and likelihood. Finally, one has to judge if the inferred posteriors are sensible in the context of the cosmological/astrophysical model and evaluate if the fitted model is a good representation of the observed data. The last step will be the topic of this section. § 7.1 describes posterior predictive checks as a means to evaluate the goodness of fit of the model and in § 7.2 we propose a method to parametrize systematics due to biased photometric likelihoods.

7.1 Model Evaluation: Posterior Predictive Checks

The idea of posterior predictive checks (PPC) is to simulate synthetic data from a fitted model, that is then compared with the original measurements to serve as an internal consistency check. For example there exist several approaches that allow us to estimate the quality of probability calibration based on the distribution of posterior predictive p-values (e.g. Gelman et al. 1996). This paper will only provide a short discussion of posterior predictive testing, which is still an area of active research. The basic idea of model checking is to investigate if data predicted by the fitted model is representative of the observed data. The classical approach, for example developed for linear regression, uses an analytical probability distribution for the test statistic (for example the χ^2 distribution) and evaluates the tail probability to test how much of an 'outlier' the observed data is, given the fitted model. This approach can be problematic if the model is complex[13], which makes it difficult to derive an analytic sampling distribution for a suitable statistic given a fitted model. Further problems arise due to significant influence of outliers, or if parameters are subject to boundary conditions. Posterior predictive checks are extensions to this classical approach in the Bayesian framework.

Starting from the composite likelihood defined in Eq. (34), the posterior predictive distribution reads

$$p(\mathcal{D}_{\rm rep}|\mathcal{D}) = \iint \mathrm{d}\mathbf{n^B}\, \mathrm{d}\mathbf{C}(\mathbf{z})\, p(\mathcal{D}^{\rm rep}|\mathbf{n^B}, \mathbf{C}(\mathbf{z}))p(\mathbf{n^B}, \mathbf{C}(\mathbf{z})|\mathcal{D})\,, \tag{35}$$

where $\mathcal{D} = \{\Gamma(w^{\rm BR}, w^{\rm RR}), \hat{\mathbf{F}}\}$ and $\mathcal{D}^{\rm rep}$ denote the replicated, or predicted, measurements sampled from the fitted model. We note that our model specifies only the ratio between $w^{\rm RB}$ and $w^{\rm RR}$. However, we can always sample replications for one of these quantities using the measurement of the other via the data transformation specified in Eq. (26).

Posterior predictive checking is particularly useful as it allows us to access model calibration quality and predictive accuracy without spectroscopic (or accurate multiband photometric) validation data. This is a decisive advantage of specifying the photometric likelihood over empirical approaches based on machine learning or, more specifically, conditional density estimation. By modelling the data generating process of SED evolution and redshifting, we can generate new photometry using a fitted physical model, giving us the opportunity to develop statistical tests based on the predictive distribution to probe the quality of the photometric likelihood calibration.

To avoid confusion, it is important to clarify that posterior predictive checking and model comparison, while often methodologically similar, have different goals. Posterior predictive checks aim to provide internal consistency tests for a given model and inference framework. Model comparison/combination has the goal to compare/combine multiple models based on some measure of fitting accuracy. However, model comparison and combination can also be based on posterior predictive accuracy (see e.g. Gelman et al. 2013). In the development of new models and the evaluation of directions for improvement, both of these concepts work together.

An important prerequisite for the fitting of complex statistical models is the evaluation of model parameter degeneracies. We have discussed the fact that the parameters that describe the

[13] An example are models that do not have the form of a generalized linear model (see, e.g., Gelman et al. 1996).

redshift-dependent galaxy-dark matter bias and the sample redshift distribution enter the clustering likelihood in a completely degenerate way. Accordingly, completely different $\mathbf{n^B}$-$\mathbf{C(z)}$ combinations will produce the same data distribution. This therefore limits the information that can be obtained on $\mathbf{n^B}$ depending on the prior information imposed on $\mathbf{C(z)}$. A combination with the photometric likelihood can help to break these degeneracies, as long as the SED modelling itself does not exhibit strong degeneracies, e.g. between the SED type and the redshift of galaxies. A dedicated study of model checking in color space is left for future work.

7.2 Parametrizing Systematics: Smoothing Kernel

In §8.3, we will discuss how miscalibrated likelihoods can lead to systematic biases and uncertainties in the deconvolution operation. Our goal in this section is to include a simple transformation into the model that parametrizes these systematics. A simple choice is the Gaussian convolution kernel, which modifies the sample redshift distribution as

$$p(z|\mathbf{n}^{\mathrm{B,sys}}) = \int \mathrm{d}\boldsymbol{\tau}\, p(\boldsymbol{\tau}) \int p(z|\overline{z},\boldsymbol{\tau}) p(\overline{z}|\mathbf{n^B})\, \mathrm{d}\overline{z}\,. \tag{36}$$

Here $\mathbf{n}^{\mathrm{B,sys}}$ denotes the parameters that describe the sample redshift distribution after the convolution is applied to the original sample redshift distribution $p(\overline{z}|\mathbf{n^B})$, where the convolution kernel is a gaussian with standard deviation and shift in the mean $\boldsymbol{\tau} = (\Delta\mu, \Delta\sigma)$:

$$p(z|\overline{z},\boldsymbol{\tau}) = \frac{1}{\sqrt{(2\pi\Delta\sigma)^2}} \exp\left[-\frac{1}{2}\left(\frac{z-\overline{z}+\Delta\mu}{\Delta\sigma}\right)^2\right] \tag{37}$$

Assuming the same histogram parametrization for $p(\overline{z}|\mathbf{n^B})$ as for $p(z|\mathbf{n^{B,sys}})$, we see that this implies an affine transformation $\mathbf{n}^{\mathrm{B}} \rightarrow \mathbf{A}(\tau)\cdot\mathbf{n}^{\mathrm{B}}$ where the matrix $\mathbf{A}$ is given as

$$\begin{aligned} A_{\beta j} &= \Phi\left(z_H^j \left| \frac{z_H^\beta - z_L^\beta}{2} + \Delta\mu, \Delta\sigma\right.\right) - \Phi\left(z_L^j \left| \frac{z_H^\beta - z_L^\beta}{2} + \Delta\mu, \Delta\sigma\right.\right) \\ &= \frac{1}{2}\left(\mathrm{erf}\left(\frac{z_H^j - \left(\frac{z_H^\beta - z_L^\beta}{2}\right) - \Delta\mu}{\Delta\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{z_L^j - \left(\frac{z_H^\beta - z_L^\beta}{2}\right) - \Delta\mu}{\Delta\sigma\sqrt{2}}\right)\right), \end{aligned} \tag{38}$$

where Φ denotes the cumulative distribution function of a normal distribution and erf the error function. We choose a Gaussian smoothing kernel since it has been shown that unbiased cosmological inference from measurements of weak lensing and LSS critically depends on accurate recovery on the mean and standard deviation of the photometric sample redshift distribution. Biases in both of these statistics can be parametrized using this kernel. In contrast to the normal distribution, which exhibits a closed form solution under an affine transformation, the logit-normal does not have this property. However we empirically find that we can approximate the shape of the distribution after an affine transformation as a logit-normal distribution to good accuracy. For a given set of $\boldsymbol{\tau}$, we can find the updated parameter values by assumed density filtering.

While marginalization using sampling techniques is possible, we choose a computationally efficient approximation and marginalize over a discrete model set that consists of different smoothing sizes $\Delta\sigma_i$ and $\Delta\mu = 0$

$$p(\mathbf{nz}^{\mathcal{B}}, \mathbf{C(z)}|\boldsymbol{\Gamma},\hat{\mathbf{F}}) = \sum_{\Delta\sigma_i} p(\Delta\sigma_i|\boldsymbol{\Gamma},\hat{\mathbf{F}})\, p(\mathbf{nz^B}, \mathbf{C(z)}|\boldsymbol{\Gamma},\hat{\mathbf{F}},\Delta\sigma_i)\,. \tag{39}$$

This assumes that there is no systematic shift in the sample redshift distribution, and deviations from the true underlying distribution are due to miscalibrated but on average unbiased individual galaxy likelihoods. Furthermore, we will assume that $p(\Delta\sigma_i|\boldsymbol{\Gamma},\hat{\mathbf{F}}) = p(\Delta\sigma_i)$, i.e., the smoothing size is a prior choice independent of the data that can be calibrated on simulations. For simplicity we will use a flat prior $p(\Delta\sigma)$ here.

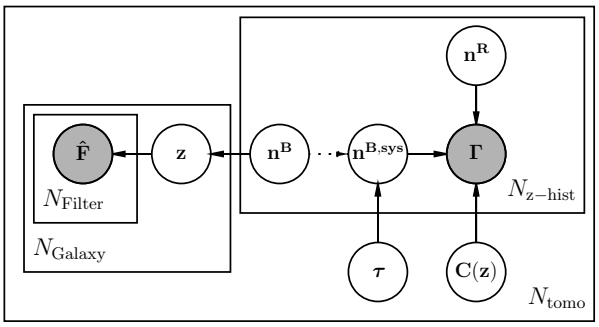

Figure 4. Directed graphical representation of the statistical model described in this paper. Empty/filled circles denote random variables that are latent/observed. $N_{(*)}$ denotes the dimensionality of the random variable. Boxes encapsulate random variables with the same dimensionality. Solid/dotted lines indicate random/deterministic relationships between random variables.

7.3 Complete model summary

Here we review and summarize our complete model. We review the structure of all components in § 7.3.1 and review the inference strategy in § 7.3.2.

7.3.1 Model Structure

Fig. 4 summarizes the joint inference strategy presented in the previous sections in a directed graphical model. Each random variable is denoted as a circle, probabilistic/deterministic relationships are denoted as solid/dotted lines. Boxes around random variables denote the dimensionality of the random variable. For example the color vector $\mathbf{f}_i$ is an N_{filter} dimensional random variable for N_{galaxies} in N_{tomo} tomographic bins. Filled circles denote observed random variables, in our case the photometry $\hat{\mathbf{F}}$ and the cross correlation ratios $\boldsymbol{\Gamma}$.

The graphical model is structured into three parts: the left part represents the photometric likelihood, the middle bullets describe our treatment of systematics, and the right part describes the clustering redshift likelihood, which depends on the spectroscopic redshift distribution and the redshift-dependent galaxy-dark matter bias ratio.

The structure of the graph illustrates the construction of the model via the composite likelihood ansatz discussed in § 6. It separates the two data sources $\hat{\mathbf{F}}$ and $\boldsymbol{\Gamma}$ in the left and right part of the graph. As we are mainly interested in performing inferences on

the $\mathbf{n}^{\rm B}$ variables, we marginalize over the z variables, which provides significant computational advantages. The mapping between $\mathbf{n}^{\rm B}$ and $\mathbf{n}^{\rm B,sys}$ takes the form of a deterministic transformation and is therefore indicated by a dotted line.

While $\mathbf{n}^{R}$ is here treated as a random variable, its 'shot noise' uncertainties are very small for the considered sample sizes of the spectroscopic sample. We therefore decided to fix its value to the maximum likelihood value, i.e., the histogram height.

7.3.2 *Model Inference*

The presented model consists of two likelihood terms and a deterministic transformation $\mathbf{n}^{\rm B} \rightarrow \mathbf{n}^{\rm B,sys}$. Starting with the photometric likelihood, we employ the inference scheme detailed in Appendices A and B that results in a posterior $p(\mathbf{n}^{\rm B}|\hat{\mathbf{F}})$ defined in Eq. (16). We then employ the transformation detailed in § 7.2 to parametrize systematics in the inferred posterior from biased photometric likelihoods. This yields a systematics corrected posterior $p(\mathbf{n}^{\rm B,sys}|\hat{\mathbf{F}})$ using the methodology described in § 7.2. The final combination with the clustering likelihood term $p(\mathbf{n}^{\rm B,sys}, \mathbf{C}(\mathbf{z})|\hat{\mathbf{F}}, \mathbf{\Gamma}) \propto p(\mathbf{\Gamma}|\mathbf{n}^{\rm B,sys}, \mathbf{C}(\mathbf{z}))\, p(\mathbf{n}^{\rm B,sys}|\hat{\mathbf{F}})$ is then performed using a Monte Carlo Markov chain (MCMC) sampling approach.

We update the parameters $(\mathbf{n}^{\rm B,sys}, \mathbf{C}(\mathbf{z}))$ in two sampling blocks: the set of parameters that describe the redshift distribution of the base sample $\mathbf{n}^{\rm B,sys}$ and the parameters $\mathbf{c}$ that describe the evolution of the redshift-dependent galaxy-dark matter bias ratio $\mathbf{C}(\mathbf{z})$.

Concretely, we iteratively sample from the conditional distributions $p(\mathbf{n}^{\mathbf{B,sys}}|\hat{\mathbf{F}}, \mathbf{\Gamma}, \mathbf{c})$ and then from $p(\mathbf{c}|\hat{\mathbf{F}}, \hat{\mathbf{\Gamma}}, \mathbf{n}^{\rm B,sys})$. This means that we iteratively sample each parameter block in turn, while holding the other parameter block fixed. The sampling method that can be employed to update each parameter blocks is flexible[14]. We use a Metropolis-Hastings sampling scheme to sample the $\mathbf{c}$ parameters. To sample from the conditional $p(\mathbf{n}^{\mathbf{B,sys}}|\hat{\mathbf{F}}, \mathbf{\Gamma}, \mathbf{c})$ we also employ a Metropolis scheme, however we perform the sampling not in terms of the $\mathbf{n}^{\rm B,sys}$ parameters, but in logit space, i.e. in terms of the $\mathbf{y}$ parameters that are connected with $\mathbf{n}^{\rm B,sys}$, or their normalized analog $\pi^{\rm sys}$, via Eq. (B1). In this way we can utilize proposal distributions that are defined in real space to sample a distribution defined on the simplex. We reiterate that the posterior $p(\mathbf{n}^{\mathbf{B}}|\hat{\mathbf{F}})$ has, in our framework, an analytical form and sampling is therefore very efficient. However if we include a treatment of systematics or a clustering redshift likelihood into the inference, we need to employ sampling approaches because the posterior has no longer a closed form solution.

8 FORECAST USING SIMULATION DATA

To demonstrate the effectiveness of our inference methodology, we consider an idealized setup which allows us to forecast the constraints on the sample redshift distribution that we can expect from a DESI-like spectroscopic survey overlapping with the LSST Y10 footprint. We assume in this section that the composite likelihood is well calibrated, both in the clustering and in the template fitting part. This assumption will likely not hold in practice and we therefore study the impact of likelihood mis-specification on the inference in § 8.3. In particular we will evaluate the performance of our methodology by comparing with science requirements of the first year ('LSST Y1') and the 10th year ('LSST Y10') of the LSST data release defined in The LSST Dark Energy Science Collaboration et al. (2018). We note that we will utilize a mock simulation of galaxy photometry likelihoods in § 8.2 instead of SED likelihoods constructed on the simulated photometry, because the simulated photometry of the DC2 simulations showed a discontinuous color-redshift mapping, which induced an unrealistically large error in our template fitting results. In § 8.3, which will discuss aspects of model checking and will not interpret results in the context of LSST science requirements, we will use both Machine Learning and Template Fitting methods described in § 2.1.

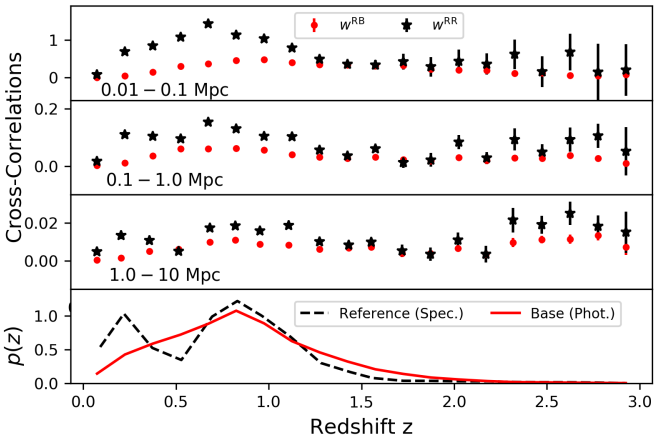

Figure 5. The top three panels show cross-correlation measurements between the photometric base sample and the spectroscopic reference sample $w^{\rm RB}$ and correlation function measurements of the spectroscopic reference sample $w^{\rm RR}$ for 3 different annuli (see § 5) as a function of redshift. The errorbars correspond to the $\pm 1\sigma$ measurement errors of the correlation function measurements. The lowest panel plots the true sample redshift probability density function of the spectroscopic reference sample ('Spec.') and the photometric base sample ('Phot.').

8.1 Measuring Cross-Correlations

We use the software package 'the-wizz'[15] (Morrison et al. 2017) to measure cross-correlations between the reference (spectroscopic) and base (photometric) samples $w^{\rm RB}$ and correlations of the reference sample $w^{\rm RR}$ in 20 equally spaced redshift bins from $z \in (0, 3)$, corresponding to 3042 Mpc comoving distance at the mean redshift of $\langle z \rangle = 0.88$. Fig. 5 shows these measurements in the three top panels for three annuli (see § 5) of $0.01 - 0.1$ Mpc, $0.1 - 1.0$ Mpc and $1.0 - 10$ Mpc. In the lowest panel we show the sample redshift probability density functions for the reference and base samples. The correlations $\mathbf{w}^{\rm RR}$ are larger than the cross-correlations $\mathbf{w}^{\rm RB}$ in all three panels, implying on average a lower than unity ratio $b^{\rm B}(z)/b^{\rm R}(z)$. The errorbars increase with redshift due to the decreasing number of galaxies, leading to a larger shot noise error. Consider the shape of $\mathbf{w}^{\rm RR}$ and $\mathbf{w}^{\rm RB}$ for the largest annuli $[1.0 - 10.]$ Mpc, in the low redshift range of $z < 1.0$. We see that the measurements of $\mathbf{w}^{\rm RB}$

[14] We tried several approaches like Hamiltonian MCMC or Elliptical Slice sampling. All approaches work well; we discuss here the structurally simplest scheme.

[15] https://github.com/morriscb/the-wizz

and $\mathbf{w}^{\mathrm{RR}}$ roughly resemble the shape of the reference and base sample redshift distributions (lowest panel), implying a roughly linear ratio $b^B(z)/b^R(z)$ in this redshift range (compare with Fig. 6). For smaller annuli, the change in slope around $z = 0.5$ is less pronounced, producing a step around $z = 0.5$ in the galaxy-dark matter bias ratio. At the high-redshift tail, where the redshift distribution of the reference sample flattens out, we see that $\mathbf{w}^{\mathrm{RB}}$ and $\mathbf{w}^{\mathrm{RR}}$ are approximately equal. Here, the sample redshift distribution of the base sample is larger than the one of the reference sample. Since the QSO sample will have a larger galaxy-dark matter bias than the base sample, we can expect $b^B(z)/b^R(z) < 1.0$ at high redshift. In order to represent a 5000 deg^2 overlap between DESI and LSST Y10, using measurements obtained on the the 300 deg^2 CosmoDC2 simulations, we scale the error on these measurements by a factor of 4 in the following analyses.

We would like to generate a cross-correlation likelihood that allows us full control over the imposed redshift-dependent galaxy-dark matter bias ratio model and that has roughly[16] the correct width of the full DESI area. Furthermore, since the mean of the ratio distribution depends on both the mean and the variance of $\mathbf{w}^{\mathbf{RB}}$ and $\mathbf{w}^{\mathbf{RR}}$, scaling the measurement error will induce biases in the mean of this ratio and therefore in the reconstructed sample photometric redshift distribution.

To correct for possible biases that would occur when the measurements errors are naively scaled and allow for better control over the redshift dependent galaxy-dark matter bias ratio, we first fit the galaxy-dark matter bias ratio $C(z)$ to the original data within these 20 bins. We show the results of this fit in Fig. 6 for different ranges in physical distance and indicate the redshift range of the different spectroscopic samples by vertical lines, where these limits are meant to guide the eye and do not constitute sharp breaks (compare with Fig. 1). We see that within these redshift ranges, $C(z)$ is a smooth function and can be fitted by a 3rd degree Chebychev polynomial $C(z) = \sum_{i=1}^{3} c_i T_i(z)$. Here, $T_i(z)$ denote Chebychev functions and c_i denotes the expansion coefficients. Within the three redshift ranges $\{[0.0, 0.5], [0.5, 0.8], [0.8, 3.0]\}$, we perform a regression fit to the median of the $C(z)$ posterior due its heavy tails. For the following analysis we select the median annuli of $0.1-1$ Mpc, which provides good signal-to-noise, while being less sensitive to small scale effects, than the $0.01 - 0.1$ Mpc bin, for which accurate modelling of galaxy-dark matter bias will be more difficult. However, we are still considering the non-linear regime, in which more work is needed to model the galaxy-dark matter bias.

We scale the correlation functions w^{RR} and w^{RB} defined in Eq. (26) and forecast a new data vector for w^{RB}, while holding the measurement of w^{RR} fixed. This amounts to multiplying the ratio $\mathbf{w}^{\mathbf{RB}}/\mathbf{w}^{\mathbf{RR}}$ by a constant for each redshift bin that compensates for the difference in the mean of the reconstructed sample photometric redshift distribution before and after we impose the fitted redshift dependent galaxy-dark matter bias model and perform the scaling. In this way we ensure that our ratio distribution is self-consistent with the photometric sample redshift distribution. We then use these adjusted measurements in the composite likelihood (Eq. 34). This correction is necessary because we would otherwise merely use noisy measurements with wrongly decreased errorbars, which will lead to biases in the probability calibration of any inference.

[16] The uncertainty in the correlation function measurements will likely differ from this factor of four scaling in the real data. Our treatment of the cross-correlation data vector here is approximate and will be complemented in future work.

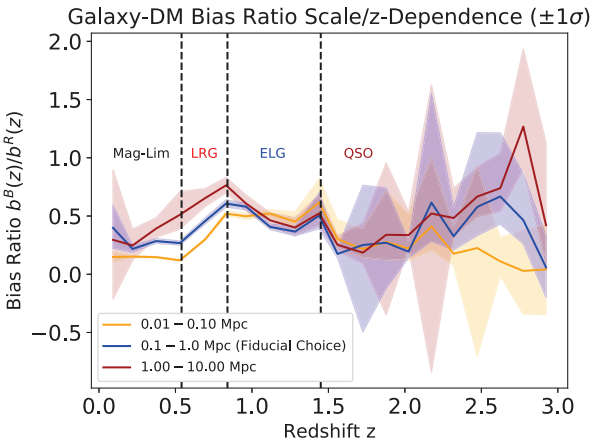

Figure 6. Galaxy-dark matter bias ratio as a function of redshift for different scales. We show here the [16, 84] percentiles, that correspond to $\pm 1\sigma$ for a Normal distribution. The redshift ranges of the different spectroscopic subsamples are plotted by vertical lines. Within these ranges, the bias ratio is a relatively smooth function of redshift, indicating a smooth redshift dependence of the galaxy-dark matter bias of the photometric sample. At the borders of these ranges, the bias ratio curves are discontinuous.

Furthermore we want to have control over the underlying redshift dependent galaxy-dark matter bias model to eliminate any additional specification error. It should therefore merely be seen as an approximate forecast of the constraining power that cross-correlations will add to the composite likelihood and a demonstration of the inference methodology. It is not an accurate treatment of galaxy-dark matter bias or the correlation function measurements expected in the final LSST measurement. For this we would require a more realistic simulation of the DESI-like spectroscopic sample, the final area and a much better understanding of the galaxy-dark matter bias of each galaxy population, all of which are subjects of active investigation in the field.

8.2 Applying the Model

For the photometric part of the composite likelihood we assume a redshift scaling of $\sigma(z) = 0.02\,(1+z)$, where z denotes the true, or spectroscopic, redshift. This scaling is a photometric redshift performance benchmark for LSST frequently adopted in the literature (e.g. Graham et al. 2020) and defined in the LSST science requirements document[17]. We then generate a mock catalog by sampling values from the true sample redshift distribution and generate a catalog of mock likelihoods by scattering these values within this redshift error model. We reiterate that we assume here that the redshift likelihoods constructed from the galaxies' photometry, mimicked here by the aforementioned redshift error model, is perfectly known. We note that this is an idealized assumption that we impose to demonstrate the methodology described in the previous sections.

Tab. 1 summarizes the different configurations we use in this work. In particular we investigate posteriors obtained using several different sample sizes and regularization techniques. In particular, the first and second columns show the abbreviation used in the text and the corresponding figure. The generated sample size of the mock catalog is shown in the third column. The columns 'Tikhonov

[17] https://docushare.lsst.org/docushare/dsweb/Get/LPM-17 (page 4)

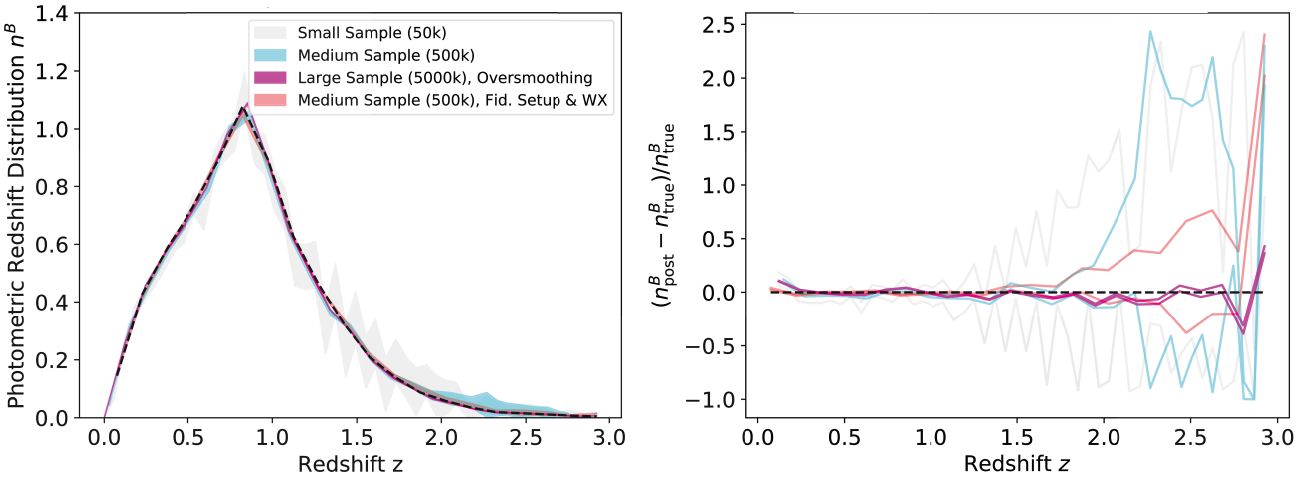

Figure 7. *Left panel:* Posteriors of the sample redshift probability density function $p(z)$ of the photometric sample (short: photometric redshift distribution) parametrized by the parameters $\mathbf{n^B}$ for different setups listed in Tab. 1. The x-axis shows the redshift value z, the y-axis the value of the $\mathbf{n^B}$ parameters. The errorbars are the $[16, 84]$ percentiles, which would correspond to 1σ intervals for a normal distribution. The black dashed curve shows the spectroscopic redshift distribution in the binning used by the cross-correlation measurements. We consider four cases, and refer to Tab. 1 and § 8 for details on the experimental setup. We highlight a variance-dominated posterior 'Small Sample (50k)', which shows a characteristic alternating, or 'zig-zag' pattern, as well as a comparison between the cyan 'Medium Sample (500k)' and 'Medium Sample (500k), Fid. Setup + WX' posteriors. Here, the latter includes a cross-correlation 'WX' data vector in its likelihood. This reduces the error especially in the high-redshift tail of the distributions. *Right panel:* The y-axis shows the relative difference between the posterior of the photometric redshift distribution parametrized by the $\mathbf{n^B_{post}}$ parameters and the spectroscopic redshift distribution $\mathbf{n^B_{true}}$ (black dashed curve in the left panel).

Abbreviation	Sample Size	Figure	Tikhonov Reg. α	Initial Binning	Effective Binning	WX
Small Sample (50k)	50k	Fig. 7	0.1	50	50	-
Medium Sample (500k)	500k	Fig. 7	0.08	50	31	-
Large Sample (5000k) Oversmoothing	5000k	Fig. 7	0.0001	25	25	-
Medium Sample (500k), Fid. Setup & WX	500k	Fig. 7	0.08	50	31	✓
Tik. Regul. Low	5000k	Fig. 8	0.0001	50	25	-
Tik. Regul. Medium	5000k	Fig. 8	1	50	25	-
Tik. Regul. High	5000k	Fig. 8	10	50	25	-
Oversmoothing	5000k	Fig. 8	0.0001	25	25	-
Tik. Regul. Low + WX	5000k	Fig. 8	0.0001	50	25	✓

Table 1. Summary of the different configurations that we test in this work. The first column lists the abbreviations, the second refers to the Figure where the setup is analysed. The next columns list the value of the Tikhonov regularization parameter α (see § 4.2.1), the number of initial bins, the effective bin number after (potentially) applying merging bin regularization (see § 4.2.1) and an indicator if the composite likelihood includes the cross-correlation data (see § 5).

Reg. α', 'Initial Binning' and 'Effective Binning' list the value of the Tikhonov regularization parameter α (see § 4.2.1), as well as the used initial and effective bin number[18], i.e., the histogram bin number after the merging bin regularization scheme. The final column indicates whether the cross-correlation data vector is included in the composite likelihood. Fig. 7 shows a selection of posteriors using setups from Tab. 1. The left hand panel shows the obtained probability density function, the right panel the relative difference between these posteriors and the spectroscopic redshift distribution that is shown as the black dashed line in both panels. The error bars are the $[16, 84]$ percentiles, corresponding to a Gaussian $\pm 1\sigma$ interval.

The 'Small Sample (50k)' setup highlights the noisy, i.e., variance-dominated deconvolution, where we clearly see the comparatively large and fluctuating errorbars in Fig. 7. We note that imposing a smoothing method will reduce these features. However this can come at the expense of additional biases as discussed later, and the characteristic covariance structure in the posterior is not *a priori* problematic, as long as draws from the posteriors are bounded and well-defined. Merging bin regularization exploits this

[18] As an approximate rule, one can expect a noisy deconvolution if no prior is applied, if the size of the bins is smaller than $\pm 1\sigma$ range of the individual galaxy redshift likelihoods for moderate sample sizes of the order 10^5 galaxies. For our redshift range and photometric redshift scatter this would imply an effective number of bins of 30 – 40.

anti-correlation structure to provide an 'objective' regularization without the need to carefully motivate an external prior or smoothing model choice.

An alternative that provides additional physical motivation is the inclusion of clustering redshift measurements into the composite likelihood. This can be seen by comparing the posteriors from the 'Medium Sample (500k)' and the 'Medium Sample (500k), Fid. Setup & WX' cases. The corresponding results in Fig. 7 show that for the same regularization, the inclusion of the cross-correlation data into the likelihood decreases the uncertainties, which is especially visible in the high-redshift tail. We note that these results are dependent on the chosen galaxy-dark matter bias model. As discussed in the beginning of this section, the parametrization used here is very flexible and the effective number of parameters can likely be reduced, if a more physical model is chosen. In this regard, we can view the presented reduction in the statistical error due to clustering redshifts as conservative. As mentioned previously, biases due to ill-motivated regularization choices play an important role, especially for large sample sizes, where the statistical error is small. We illustrate this here in the 'Large Sample (5000k) Oversmoothing' case, by deliberately choosing a coarser binning without merging bin regularization. We clearly see that the statistical error is quite small with the bias dominating.

In order to investigate the quality of probability calibration, we consider the posterior distribution over the mean values of photometric sample redshift distributions drawn from the posterior of $\mathbf{n^B}$.[19] This is a reasonable choice, as it has been shown that accurate modelling of weak lensing and LSS critically depends on accurate recovery of the posterior mean.

Fig. 8 shows five boxplots that each visualize the distribution of the posterior mean that corresponds to a different setup under consideration. The box edges denote the $[16, 84]$ percentiles, and the definition of the whiskers, i.e., the thin vertical lines with short horizontal edges represent the $[2.5, 97.5]$ percentiles. The horizontal line within the box represents the median[20]. The x-axis shows several different scenarios, as listed in Tab. 1; the y-axis shows the value of the posterior mean. The middle solid black line corresponds to the mean of the true redshift distribution, shown as the dashed black line in the left panel of Fig. 7. We reiterate that all results have been obtained using a mock catalog containing 5000k galaxies. The (dashed/dotted), (grey/magenta) horizontal lines represent the requirement values for (Y1/Y10), (LSS/WL) measurements as given in the LSST DESC Science Requirements Document (DESC SRD The LSST Dark Energy Science Collaboration et al. 2018). We note that the LSST DESC Science Requirements Documents considers a tomographic analysis and not a single bin, as we do here. We therefore restrict ourselves to a qualitative comparison. Furthermore it should be noted that higher order moments of the photometric sample redshift distribution will also correlate with cosmological parameters, especially for a clustering likelihood (see e.g. Nicola et al. 2020; Hadzhiyska et al. 2020), and our metric is therefore bound to be incomplete. Redefining these metrics and requirements is the subject of ongoing work.

We consider three scenarios: 'Tik Regul. Low', 'Tik. Regul. Medium' and Tik. Regul. High'. As can be seen from Tab. 1 these scenarios differ by their value of the Tikhonov regularization parameter α. With increasing α, the error bars decrease and the bias in the results increases. When comparing this with the 'oversmoothing' results, we see the same pattern. This similarity in behaviors arises because both a large α and choosing large bins reduces the variance of each bin.

Finally we show the impact of including the cross-correlation measurements into the data vector in the 'Tik. Regul. Low + WX' scenario, which adds clustering information to the 'Tik. Regul. Low' scenario. When comparing these two cases, we see that the distribution of the posterior mean is now symmetric and reasonably centered within the science requirements. In particular we note that the uncertainties are still much larger when compared with the previously considered, strongly regularized cases. This shows that while the effect of reducing the variance of the posteriors is similar when using regularization or including cross-correlation data into the composite likelihood, the posteriors can be much better calibrated in the latter case. Using a smoothing, or regularization, method essentially makes assumptions about the true shape of the distribution without strict data evidence. In contrast, adding cross-correlations to the composite likelihood adds this information in a physical, data-driven way.

Another effect that needs consideration is the increase in the intrinsic estimator bias due to the 'downsampling' of the probability density function to a lower resolution, e.g., by picking larger bin width or by imposing a different regularization or smoothing scheme. This loss in resolution implies that we inadvertently limit the accuracy with which small scale structure can be reconstructed in the density field along the line-of-sight. As demonstrated and studied in detail in Rau et al. (2017), this effect can lead to biases in the cosmological parameter inference that are often small, but that would need scrutiny for upcoming data analyses. Since we gave a detailed description of this effect in Rau et al. (2017) including schemes to detect and mitigate these effects, we do not focus on it in detail here. However, this effect can be illustrated for the current setup, since the merging bin regularization downsamples the resolution to a relatively coarse grid of 31 bins. Furthermore consider the redshift distribution of true redshifts discretized using the 20 bin grid used to obtain the cross-correlation measurements. Since we use this distribution, i.e. the black curve in Fig. 7, as a reference, we also have to consider its intrinsic discretization error. Concretely, when comparing the mean estimated from this curve with the sample mean, we obtain a difference in these values of 0.0079. While this is of the same order as the Y1 science requirements in Fig. 8, Y10 requirements will necessitate an increase in sample size or the inclusion of cross-correlation constraints that will allow us to perform inference at a higher resolution. Due to the slow expected convergence of deconvolution estimators with sample size (see e.g. Carroll & Hall 1988), it is likely that several orders of magnitude increase in sample size will be necessary. This is attainable for the large numbers of observed galaxies in LSST Y10, and our methodology can scale to large sample sizes. However, in order to reach the sample sizes that are expected for LSST observations, we need to develop an implementation that optimizes storage space and uses an efficient parallelization strategy, which is beyond the scope of this work.

Alternatively, we could use a different scheme that employs a continuous model like logistic Gaussian processes (Rau et al. 2020) or Dirichlet processes. The convergence of these density estimators will likely be better, however they will also require additional computational overhead in the inference. A detailed study of estimator convergence is needed to settle on a recommendation and prove sig-

[19] Concretely, we draw a number of n^B realizations that each parametrize a photometric sample redshift distribution and evaluate the mean on each of these distributions.

[20] We note that the original definition of the boxplot uses a different definition of the box size and the whiskers. We refer to Wickham & Stryjewski (2012) for more details.

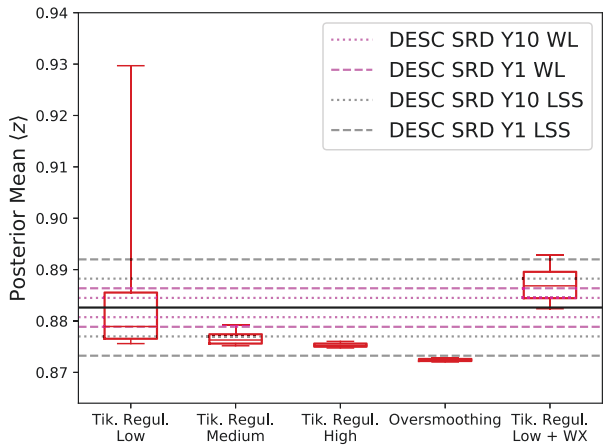

Figure 8. Boxplot illustration of the mean of the posterior sample redshift probability density function of the photometric sample (short: posterior mean) for different experimental setups listed in Tab. 1 and detailed in § 8. The x-axis lists the different scenarios, the y-axis the value of the posterior mean. The box shows the $[16, 84]$ percentiles, the vertical lines with horizontal edges (whiskers) show the $[2.5, 97.5]$ percentiles, corresponding to the 1σ and 2σ intervals for the normal distribution. The horizontal line in the box is the median. The (dashed/dotted), (grey/magenta) lines correspond to the requirement on the uncertainty of the posterior mean as quoted in the LSST DESC Science Requirements Document (DESC SRD, The LSST Dark Energy Science Collaboration et al. 2018) for (Y1/Y10) (LSS/WL) measurements. The central, solid grey line is the mean of the true redshift distribution, shown as the dashed black line in the left panel of Fig. 7. All results have been obtained using a mock catalog of 5000k galaxies with photometric redshift scatter that is perfectly calibrated. We highlight the decrease in the statistical error and potential increase in systematic bias for larger regularization, going from the leftmost to the fourth case. The rightmost boxplot shows the impact of including clustering redshift information into the likelihood.

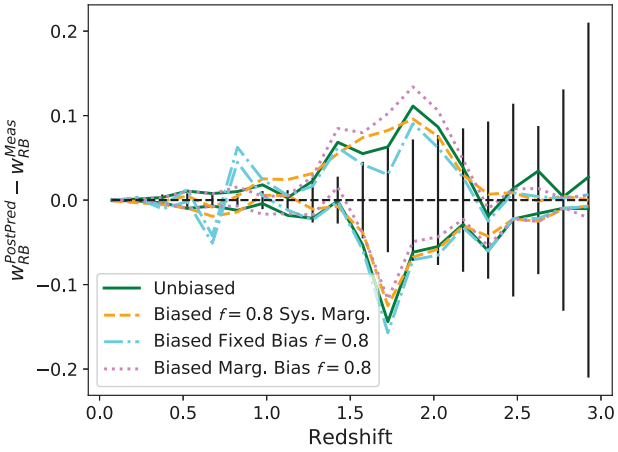

Figure 9. We plot the residual between replicated and true scale-averaged cross-correlation measurement $\mathbf{w}^{\mathrm{RB}}$ between a DESI-like spectroscopic reference ('R') and photometric base ('B') sample as a function of redshift. The horizontal line with errorbars shows the uncertainties in the original measurement. The green/magenta contours show the scenario where we marginalize over all parameters $\mathbf{c}$ that parametrize the redshift dependent galaxy-dark matter bias ratio function $C(z)$ (see § 8.1) without the systematics kernel for the unbiased/biased ($f = 0.8$) cases. The yellow/blue contours consider the biased scenario ($f = 0.8$), but fix $C(z)$ and do/do not marginalize over the systematics kernel. The error bars and contours show the [5, 95] percentiles. We see that the replicated measurements do not show significant tension with the original measurements, if we either marginalize over the systematic ('Biased f=0.8 Sys. Marg.') or if we use a flexible redshift-dependent galaxy-dark matter bias model ('Biased Marg. Bias f=0.8'). Only if the form of the galaxy-dark matter bias is known to good precision – in this case we hold its values fixed – are PPCs using cross-correlations sensitive in detecting tensions.

nificant improvement over the simple histogram scheme employed here; we will leave this for future work.

Most importantly, however, it is likely that systematic errors due to the miscalibration and mis-specification of the composite likelihood, either by a suboptimal galaxy-dark matter bias model or due to miscalibrated SED likelihoods, will lead to an error budget that will dominate the aforementioned errors. If the misspecification can be parametrized and marginalized over, the variance of the parameter posteriors will be increased, otherwise they will lead to biases in the resulting parameter posteriors. In the following section we will discuss these sources of error. We will showcase the usage of posterior predictive checks as a way to detect miscalibration and suggest procedures for consistent model checking and refinement.

8.3 Testing the Model

We discussed and showcased our inference methodology in the previous section using idealized data. To complement the discussion, this section highlights how miscalibrated likelihoods can lead to biases in the inferred sample redshift distribution and how posterior predictive checks can be used to detect these issues.

To mimic well-calibrated photometric likelihoods we use conditional density estimates from the FlexZboost package (Dalmasso et al. 2020). We note that these conditional densities are not photometric likelihoods in the sense of Eq. (34). The free parameters in the photometric likelihood are the redshifts of the galaxies and parameters that describe properties of the Spectral Energy Distribution (SED). In conditional density estimation, non-physical parameters describe a flexible model that provides a mapping between photometry and redshift. This flexible model is then fitted to known calibration data. Thus while in SED fitting the distribution of redshift constitutes a posterior distribution, conditional density estimation treats it as a predictive distribution (often without marginalizing over the modelling uncertainty). However since the goal of this subsection is to demonstrate potential systematic biases and uncertainties in the deconvolution operation, this difference is not of great importance here.

In order to simulate the impact that a population of galaxies with inaccurately calibrated photometric redshift likelihoods has on the deconvolved redshift distribution, we consider a redshift range of $0.2 - 0.8$ and a total of 500k galaxies. We randomly substitute 80% of the FlexZboost conditional density predictions with photometric likelihoods obtained using a template fitting run from the BPZ code by employing a k-nearest neighbor substitution in redshift. The result is a dataset in which a fraction of 80% ($f = 0.8$) of galaxies have a likelihood from the BPZ code, and only 20% retain their conditional density predictions from the FlexZboost code. We picked this setup because the BPZ predictions within this redshift range, while being inferior to the FlexZboost predictions, still have an acceptable quality. We perform this experiment by selecting a range in redshift because we will perform posterior predictive checks (PPC) using cross-correlations and need to control where we would expect systematics. This will allow us to disentangle model misspecification

issues from the systematics, e.g. from the 'noisy' deconvolution, described previously. We note that the quality of photometric redshift likelihoods does not sharply change with redshift in this way, in real photometric samples. Instead photometric redshift quality is a complex function in color space that strongly depends e.g. on the quality of the photometry, the number of available bands, the amount of calibration data and the template set. Modelling this accurately is beyond the scope of this work and would require the measurement of cross-correlations in color cells and the extension of PPC to the full composite likelihood, that includes sampling the photometry of galaxies in these color cells. Using the mean of the FlexZboost conditional densities for the selection instead of the true redshifts would 'smooth-out' the quality of the likelihoods as a function of redshift at the boundaries of the $0.2 - 0.8$ redshift interval. However this will also not be representative of the aforementioned difficulties. We therefore choose an unrealistically simple case that nonetheless illustrates the usefulness of PPC. Furthermore it allows us to highlight difficulties in their application in a controlled manner, by picking a fixed redshift range in which individual galaxy likelihoods are biased.

In the spirit of PPC, we generate new cross-correlation measurements using the joint posterior of the sample redshift distribution parameters $\mathbf{n^B}$ and the parameters that govern the galaxy-dark matter bias ratio evolution $\mathbf{c}$, following the Chebychev basis expansion described in § 8.1. For simplicity we will deconvolve the redshift distribution on the same 20 bin redshift grid used in the cross-correlation data vector. For 500k galaxies, this leads to very small statistical errorbars in the deconvolution. As a simplification we can then fix the $\mathbf{n^B}$ posterior to its maximum likelihood value. As mentioned in the previous section, this oversmoothing will lead to biases in the $\mathbf{n}^\mathrm{B}$ posteriors. However, since we will only perform a posterior predictive analysis with respect to the clustering likelihood, that is less constraining than the photometric likelihood, the systematics incurred by these simplifications and the underestimation of statistical error, are sub-dominant compared with the overall statistical error budget from the correlation function measurements.

Fig. 9 shows the sampled cross correlation measurements from the fitted joint model, in residual to the original measurements. We showcase four scenarios. In the unbiased case we use FlexZBoost PDFs and marginalize over all $\mathbf{c}$ parameters. Due to the good calibration of these Machine Learning-produced conditional distributions, we obtain very similar results compared with the previous section. The reason for this success is, of course, the representative training set that would not be available in a practical application. Furthermore we consider three scenarios with $f = 0.8$. The scenario shown in yellow fixes the galaxy-dark matter bias parameters ($\mathbf{c}$), but marginalizes over the parameters of the systematics kernel. The blue/magenta lines fix/marginalize over the $\mathbf{c}$ parameters, but do not include the systematics kernel correction.

As can be seen in Fig. 9, the treatment of galaxy-dark matter bias has a profound impact on the consistency between the replicated and original cross-correlation measurements. Within the errors, the results are consistent for all scenarios except the one without systematics kernel correction that uses a fixed $C(z)$ model. As shown in the yellow line, these biases can be corrected by the systematics kernel marginalization. This illustrates that if sufficient information about $C(z)$ is available, the clustering likelihood alone can allow for powerful posterior predictive checks. If this is not the case, consistency tests of redshift distributions with respect to clustering redshift measurements can be misleading. Provided sufficient information on the galaxy-dark matter bias, we can parametrize the biases in the deconvolved density estimate using, e.g., a convolution with a kernel function as described in § 7.2. We show these results as the yellow lines 'Biased $f = 0.8$ Sys. Marg'. Here we perform a discretized marginalization as described in § 7.2, by convolving the $\mathbf{n^B}$ vector with a Gaussian kernel function of width $\Delta\sigma \in [0.001, 0.2]$ in 40 steps. We see that this correction can compensate for the misspecified likelihoods even in the case of a fixed $C(z)$ model. The degeneracy between the redshift-dependent galaxy-dark matter bias ratio model and the $\mathbf{n^B}$ parameters highlights the importance of performing posterior predictive checks in color space to provide additional information on the redshift distribution. However, this requires careful modelling of SEDs and the development of a transparent, reproducible analysis framework that additionally includes tests for parameter degeneracies and a model comparison framework. This is beyond the scope of this work, but will be addressed in a future paper.

9 SUMMARY AND CONCLUSIONS

Accurate photometric redshift inference is one of the most important challenges in large area photometric surveys like LSST, DES, HSC, or KiDS. As discussed in detail in § 3, photometric redshift inference is, from a statistical point of view, a deconvolution problem, where an underlying true redshift distribution is convolved with an SED model-dependent error distribution given by the photometric likelihood. The deconvolution inference of sample redshift distribution is not new (e.g. Padmanabhan et al. 2005; Leistedt et al. 2016; Malz & Hogg 2020), and spatial information has also been incorporated into the inference (e.g. Alarcon et al. 2020b; Sánchez & Bernstein 2019; Jones & Heavens 2019; Rau et al. 2020). We extended these prior works by developing a fast approximate inference scheme for deconvolution, that combines redshift information from both the photometry and the spatial distribution of galaxies in terms of a composite likelihood ansatz. We particularly provided a discussion on regularization techniques and the tradeoff between bias and variance in the Bayesian context for medium to large sample sizes.

In particular, our goal is to include the treatment of photometric redshift via the likelihood of the galaxies' photometry into the current cosmological inference framework, which is based on correlation functions. The main reason for our likelihood choice is to allow the easy integration into the likelihood inference framework based on two-point statistics of galaxy density and shear fields. This is more difficult for other approaches presented in Alarcon et al. (2020b) and Sánchez & Bernstein (2019) since, in the currently demonstrated form, the redshift information from cross-correlating the overlapping spectroscopic sample is included via an estimator and not using a likelihood (that would depend on cosmological parameters). While the works of Padmanabhan et al. (2005); Leistedt et al. (2016); Malz & Hogg (2020) are structurally similar in terms of the treatment of the photometric likelihood, they do not discuss the effect of including clustering information. It is noteworthy that the early work by Padmanabhan et al. (2005) provides an excellent, explicit discussion of regularization, which is the main difficulty in the photometric redshift problem, although not in the context of probability calibration and the aforementioned joint inference framework. We summarized our approach to the challenges that arise in the estimation of redshift distributions for samples of galaxies in the context of photometric surveys. Concretely, we considered the combination of photometric information with two-point statistics, the scalability and regularization of the deconvolution inference in the large sample scenario, and investigate the impact of

systematics from misspecified individual galaxy photometric likelihoods, proposing parametrizations for these systematics. These achievements lay the foundations for future extensions that we will discuss in the next section.

In § 4.1 we described our inference methodology that is designed to facilitate inference on large galaxy catalogs to be expected in LSST. The scheme uses a Laplace Approximation in logit space and facilitates inference using an iterative scheme of expectation maximization update equations. This provides computational advantages over sampling approaches. Additionally this methodology facilitates fast joint inference with a cross-correlation data vector (see § 5) that we included in a composite likelihood ansatz. As highlighted in § 6, this provides the possibility of additional extensions that include two-point statistics from cosmological weak lensing and galaxy-galaxy lensing measurements. As we discussed in § 3, ensemble redshift distribution inference based on a photometric likelihood is a deconvolution problem, which requires regularization to yield bounded and well-defined results. In this context, we discussed a regularization scheme that consists of a combination of Tikhonov regularization with (more importantly) a scheme that merges neighboring bins to exploit the characteristic covariance structure in the deconvolved densities. In agreement with the findings of the original paper by Kuusela (2016) that proposed and applied this scheme to the Poisson inverse problem, we find that the 'Merging Bin' scheme leads to better calibrated results as compared with Tikhonov regularization and with an oversmoothing scheme that selects a coarser redshift binning for the sample redshift histograms.

In order to test and discuss the quality of our posterior inference, we used data from the CosmoDC2 simulations to generate a spectroscopic DESI-like sample and a photometric mock catalog, that uses an LSST-like photometric error model. This allowed us to test the impact of a spectroscopic calibration sample with an inhomogeneous galaxy population as a function of redshift. We found that the ratio between the redshift-dependent galaxy-dark matter bias of the photometric and the spectroscopic sample is a smooth function of redshift, if the spectroscopic calibration sample consists of a single galaxy population, and is discontinuous if the galaxy population strongly changes. We therefore employed a step-wise smooth function based on a Chebychev polynomial expansion to parametrize this ratio.

In § 8 we performed a forecast of redshift inference performance on ideal data, assuming perfectly calibrated individual galaxy redshift likelihoods. We found that using the aforementioned merging bin regularization, we were able to produce accurate posteriors of ensemble redshift distributions. We reiterate that using other regularization schemes, like an overly large Tikhonov regularization parameter, or an oversmoothing approach that picks overly wide histogram bins, can lead to significant biases in the recovered posterior mean.

When compared with the DESC science requirements for WL and large scale structure measurements in terms of the mean of the photometric sample redshift distribution, we found that we can meet the DESC SRD Y1 goals and remain consistent with the DESC SRD Y10 goals with 5000k galaxies, if cross-correlations are included in the joint composite likelihood. In practical applications, however, Spectral Energy Distribution (SED) templates for galaxies will be subject to modelling biases that cannot be well calibrated using spectroscopic data (see e.g. Hartley et al. 2020). We therefore proposed to use posterior predictive checks (PPC) as a means to evaluate the quality of our inference. Here, we compared replications of the data sampled from the fitted model with the original measurement to evaluate model goodness-of-fit. Specifically cross-correlation redshift inference is often used to calibrate photometric redshifts obtained using photometry (Newman 2008; Johnson et al. 2016; Davis et al. 2017). In § 8.3 we demonstrated that PPC of cross-correlation measurements can detect systematic biases in the recovered sample redshift distribution if the galaxy-dark matter bias of the photometric and spectroscopic samples is known to sufficient accuracy.

In order to parametrize potential biases in the sample redshift distribution posteriors caused by misspecified photometric likelihoods, particularly over-deconvolution effects that lead to overly narrow redshift distributions, we proposed a simple Gaussian filter that, as demonstrated in § 8.3, was able to correct these biases.

10 FUTURE WORK

In future work, it will be important to extend the inference scheme developed in this paper. We plan to consider a range of extensions, e.g., iterated nested Laplace approximations (Bornkamp 2011) in logit space, the usage of more flexible distributions that can be fitted using variational inference schemes, as well as the development of specialized subsampling MCMC schemes. The different techniques will be evaluated in combination with regularization approaches based on the quality of their probability coverage. Another extension, particularly to reduce the bias in the density estimation, is to consider other parametrizations for the deconvolved density either by employing density estimators with better mean squared error scaling like Kernel Density estimators, basis function expansions or using methods such as logistic Gaussian Processes (e.g. Rau et al. 2020). The combination of photometric and clustering information can be extended by connecting the modelling of SEDs and redshift-dependent galaxy-dark matter bias modelling via the luminosity function as shown in van Daalen & White (2018). This also has the potential to reduce the degeneracy between SED and redshift-dependent galaxy-dark matter bias systematics. Finally, we note that the data quality of the photometry will not be the same in all areas on the sky. In order to include these field-to-field variations into the composite likelihood framework, we can, for example, condition the likelihood on the field and include a corresponding data covariance into the likelihood by employing either resampling techniques (see e.g. Davison & Hinkley 2013) or using theoretical modelling (e.g. Stoyan & Stoyan 1994; Sánchez et al. 2020).

To conclude, we have presented an efficient photometric redshift inference framework that combines information from both the photometry and the spatial distribution of galaxies. The methodology is designed to scale well to large samples. We complement this framework with methods for regularization, model checking and redshift systematics parametrization. The forecasts we performed using CosmoDC2 data give us confidence that, with the additional improvements described here, the methodology presented will enable accurate and well-calibrated redshift inference for LSST and other ongoing and future large area photometric surveys.

SOFTWARE

Besides software referenced directly in the text, we performed the analyses in this work using the following software packages: the python language (van Rossum 1995), scipy (Virtanen et al. 2020), numpy (Harris et al. 2020), jupyter notebook (Kluyver et al. 2016),

ipython (Perez & Granger 2007), matplotlib (Hunter 2007) and pandas (Wes McKinney 2010).

DATA AVAILABILITY STATEMENT

The cosmoDC2 extragalactic catalog is publicly available at `https://portal.nersc.gov/project/lsst/cosmoDC2/_README.html`. The ancillary catalogs (photo-z and DESI-like selection for cosmoDC2) and other derived data underlying this article will be shared on reasonable request to the corresponding author. The source code that implements the algorithms presented in this article will be made available via Zenodo.

ACKNOWLEDGEMENTS

This paper has undergone internal review in the LSST Dark Energy Science Collaboration. We would like to thank the internal reviewers David Alonso, Will Hartley and David Kirkby for their insightful comments.

MMR led and planned the project from the initial idea and motivation to the experimental design. He performed the analyses in interaction with the coauthors, and prepared the paper. CBM contributed by: development of clustering redshifts software, creation of clustering redshifts photometric and spectroscopic samples, creation of clustering redshift data products. SJS constructed the photometric redshift catalogs and contributed to writing the corresponding portions of the paper. SW provided feedback on the method development and manuscript. RM advised on motivation, scope, experimental design, and analysis, and contributed to the editing of the paper draft. YYM developed the access tools for cosmoDC2 and photo-z data products.

MMR and RM are supported by DOE grant DE-SC0010118 and NSF grant IIS-1563887. SJS acknowledges support from DOE grant DESC0009999 and NSF/AURA grant N56981C. MMR was supported in part by the Simons Foundation (Simons Investigator Award, PI: Mandelbaum). This material is based upon work supported in part by the National Science Foundation through Cooperative Agreement 1258333 managed by the Association of Universities for Research in Astronomy (AURA), and the Department of Energy under Contract No. DE-AC02-76SF00515 with the SLAC National Accelerator Laboratory. Additional LSST funding comes from private donations, grants to universities, and in-kind support from LSSTC Institutional Members. SJS acknowledges support from DOE grant DE-SC0009999 and NSF/AURA grant N56981C. Support for YYM was provided by NASA through the NASA Hubble Fellowship grant no. HST-HF2-51441.001 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555.

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science & Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515.

REFERENCES

Abbott T. M. C., et al., 2018a, Phys. Rev. D, 98, 043526
Abbott T. M. C., et al., 2018b, ApJS, 239, 18
Aihara H., et al., 2018, PASJ, 70, S4
Alarcon A., et al., 2020a, arXiv e-prints, p. arXiv:2007.11132
Alarcon A., Sánchez C., Bernstein G. M., Gaztañaga E., 2020b, MNRAS, 498, 2614
Albrecht A., et al., 2006, arXiv e-prints, pp astro–ph/0609591
Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
Atchison J., Shen S., 1980, Biometrika, 67, 261
Benítez N., 2000, ApJ, 536, 571
Benjamin J., et al., 2013, MNRAS, 431, 1547
Benson A. J., 2012, New Astron., 17, 175
Bernstein G., Huterer D., 2010, MNRAS, 401, 1399
Bishop C. M., 2006, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg
Bonnett C., 2015, MNRAS, 449, 1043
Bornkamp B., 2011, Journal of Computational and Graphical Statistics, 20, 656
Box G. E. P., Muller M. E., 1958, Ann. Math. Statist., 29, 610
Brown M. J. I., et al., 2014, ApJS, 212, 18
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Carroll R. J., Hall P., 1988, Journal of the American Statistical Association, 83, 1184
Cawthon R., et al., 2020, arXiv e-prints, p. arXiv:2012.12826
Chang C., et al., 2016, MNRAS, 459, 3203
Chen T., Guestrin C., 2016, in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowled ge Discovery and Data Mining. KDD '16. ACM, New York, NY, USA, pp 785–794, doi:10.1145/2939672.2939785, `http://doi.acm.org/10.1145/2939672.2939785`
Clerkin L., Kirk D., Lahav O., Abdalla F. B., Gaztañaga E., 2015, MNRAS, 448, 1389
Collister A. A., Lahav O., 2004, Publications of the Astronomical Society of the Pacific, 116, 345
Craig I. J. D., Brown J. C., 1986, Inverse problems in astronomy. A guide to inversion strategies for remotely sensed data
DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036
Dalmasso N., Pospisil T., Lee A. B., Izbicki R., Freeman P. E., Malz A. I., 2020, Astronomy and Computing, 30, 100362
Davis C., et al., 2017, arXiv e-prints, p. arXiv:1710.02517
Davison A. C., Hinkley D. V., 2013, Bootstrap Methods and Their Application. Cambridge University Press, USA
Feldmann R., et al., 2006, MNRAS, 372, 565
Gatti M., et al., 2018, MNRAS, 477, 1664
Gatti M., et al., 2020, arXiv e-prints, p. arXiv:2012.08569
Gelman A., li Meng X., Stern H., 1996, Statistica Sinica, pp 733–807
Gelman A., Hwang J., Vehtari A., 2013, arXiv e-prints, p. arXiv:1307.5928
Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823
Graham M. L., et al., 2020, AJ, 159, 258
Greisel N., Seitz S., Drory N., Bender R., Saglia R. P., Snigula J., 2015, MNRAS, 451, 1848
Hadzhiyska B., Alonso D., Nicola A., Slosar A., 2020, arXiv e-prints, p. arXiv:2007.14989
Hahn C., Beutler F., Sinha M., Berlind A., Ho S., Hogg D. W., 2019, MNRAS, 485, 2956
Harris C. R., et al., 2020, Nature, 585, 357–362

Hartley W. G., et al., 2020, MNRAS, 496, 4769
Hearin A., Korytov D., Kovacs E., Benson A., Aung H., Bradshaw C., Campbell D., LSST Dark Energy Science Collaboration 2020, MNRAS, 495, 5040
Heitmann K., et al., 2019, ApJS, 245, 16
Heymans C., et al., 2020, arXiv e-prints, p. arXiv:2007.15632
Hikage C., et al., 2019, PASJ, 71, 43
Hildebrandt H., et al., 2017, MNRAS, 465, 1454
Hildebrandt H., et al., 2020, arXiv e-prints, p. arXiv:2007.15635
Hoyle B., 2016, Astronomy and Computing, 16, 34
Hoyle B., Rau M. M., 2019, MNRAS, 485, 3642
Hoyle B., Rau M. M., Bonnett C., Seitz S., Weller J., 2015, MNRAS, 450, 305
Hoyle B., et al., 2018, MNRAS, 478, 592–610
Hunter J. D., 2007, Computing in Science Engineering, 9, 90
Huterer D., Takada M., Bernstein G., Jain B., 2006, MNRAS, 366, 101
Huterer D., Lin H., Busha M. T., Wechsler R. H., Cunha C. E., 2014, MNRAS, 444, 129
Ilbert O., et al., 2006, A&A, 457, 841
Ivezić Ž., et al., 2019, ApJ, 873, 111
Izbicki R., Lee A. B., 2017, Electron. J. Statist., 11, 2800
Johnson A., et al., 2016, MNRAS, 465, 4118
Jones D. M., Heavens A. F., 2019, MNRAS, 483, 2487
Joudaki S., et al., 2018, MNRAS, 474, 4894
Joudaki S., et al., 2020, A&A, 638, L1
Kalmbach J. B., Connolly A. J., 2017, AJ, 154, 277
Kluyver T., et al., 2016, in Loizides F., Schmidt B., eds, Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp 87 – 90
Korytov D., et al., 2019, ApJS, 245, 26
Kress R., 1998, Numerical Analysis. Graduate Texts in Mathematics, Springer New York, https://books.google.com.na/books?id=R6182rh0tKEC
Kuusela M. J., 2016, PhD thesis, Lausanne, EPFL, doi:10.5075/epfl-thesis-7118, http://infoscience.epfl.ch/record/220015
Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193
Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258
Ma Z., Hu W., Huterer D., 2006, ApJ, 636, 21
Malz A. I., Hogg D. W., 2020, arXiv e-prints, p. arXiv:2007.12178
Mandelbaum R., 2018, ARA&A, 56, 393
Matarrese S., Coles P., Lucchin F., Moscardini L., 1997, MNRAS, 286, 115
McLeod M., Balan S. T., Abdalla F. B., 2017, MNRAS, 466, 3558
McQuinn M., White M., 2013, MNRAS, 433, 2857
Meister A., 2009, Deconvolution Problems in Nonparametric Statistics. Lecture Notes in Statistics, Springer Berlin Heidelberg, https://books.google.de/books?id=ItGkJPZQj-MC
Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, arXiv e-prints, p. arXiv:1303.4722
Morrison C. B., Hildebrandt H., Schmidt S. J., Baldry I. K., Bilicki M., Choi A., Erben T., Schneider P., 2017, MNRAS, 467, 3576
Myles J., et al., 2020, arXiv e-prints, p. arXiv:2012.08566
Neal R. M., Hinton G. E., 1993, in Learning in Graphical Models. Kluwer Academic Publishers, pp 355–368
Newman J. A., 2008, ApJ, 684, 88
Newman J. A., et al., 2015, Astroparticle Physics, 63, 81
Nicola A., et al., 2020, J. Cosmology Astropart. Phys., 2020, 044
Padmanabhan N., et al., 2005, MNRAS, 359, 237
Pawitan Y., 2001, In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford science publications, OUP Oxford, https://books.google.com/books?id=M-3pSCVxV5oC
Perez F., Granger B. E., 2007, Computing in Science Engineering, 9, 21
Prat J., et al., 2018, MNRAS, 473, 1667
Quiroz M., Villani M., Kohn R., Tran M.-N., Dang K.-D., 2018, arXiv e-prints, p. arXiv:1807.08409
Raccanelli A., Rahman M., Kovetz E. D., 2017, MNRAS, 468, 3650
Ranganathan A., 2004, Assumed Density Filtering, http://www.ananth.in/Notes_files/adf.pdf
Rau M. M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, MNRAS, 452, 3710
Rau M. M., Hoyle B., Paech K., Seitz S., 2017, MNRAS, 466, 2927
Rau M. M., Wilson S., Mandelbaum R., 2020, MNRAS, 491, 4768
Raue A., Kreutz C., Theis F., Timmer J., 2013, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 371, 20110544
Rothenberg T. J., 1971, Econometrica, 39, 577
Salvato M., Ilbert O., Hoyle B., 2019, Nature Astronomy, 3, 212
Sánchez C., Bernstein G. M., 2019, MNRAS, 483, 2801
Sánchez C., Raveri M., Alarcon A., Bernstein G. M., 2020, MNRAS,
Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, MNRAS, 431, 3307
Scottez V., et al., 2016, MNRAS, 462, 1683
Scranton R., et al., 2005, ApJ, 633, 589
Simon P., Hilbert S., 2018, A&A, 613, A15
Speagle J. S., Eisenstein D. J., 2015, arXiv e-prints, p. arXiv:1510.08073
Spergel D., et al., 2015, arXiv e-prints, p. arXiv:1503.03757
Stölzner B., Joachimi B., Korn A., Hildebrandt H., Wright A. H., 2020, arXiv e-prints, p. arXiv:2012.07707
Stoyan D., Stoyan H., 1994, Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics. Wiley Series in Probability and Statistics, Wiley, https://books.google.com/books?id=Dw3vAAAAMAAJ
Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, Neural Networks for Photometric Redshifts Evaluation. pp 226–234, doi:10.1007/978-3-540-45216-4_26
Tanaka M., et al., 2018, PASJ, 70, S9
The LSST Dark Energy Science Collaboration et al., 2018, arXiv e-prints, p. arXiv:1809.01669
Uitert E., et al., 2017, MNRAS, 476
Varin C., Reid N., Firth D., 2011, Statist. Sinica, pp 5–42
Virtanen P., et al., 2020, Nature Methods, 17, 261
Wes McKinney 2010, in Stéfan van der Walt Jarrod Millman eds, Proceedings of the 9th Python in Science Conference. pp 56 – 61, doi:10.25080/Majora-92bf1922-00a
Wickham H., Stryjewski L., 2012, Technical report, 40 years of boxplots. had.co.nz
Zhou R., et al., 2020a, arXiv e-prints, p. arXiv:2001.06018
Zhou R., et al., 2020b, Research Notes of the American Astronomical Society, 4, 181
van Daalen M. P., White M., 2018, MNRAS, 476, 4649
van Rossum G., 1995, Technical Report CS-R9526, Python tutorial. Centrum voor Wiskunde en Informatica (CWI), Amsterdam

APPENDIX A: DERIVING THE E-M UPDATE EQUATIONS

We start the discussion with an intuitive motivation for the theoretical foundation of the E-M algorithm. Assume a 'system' that consists of hidden variables Y and observed variables Z. We wish to find a set of parameters θ that maximize the joint distribution of both variables given θ. We know from statistical physics that the free energy of this system $F(p, \theta)$ should be minimized and depends on the distribution of hidden variables, or states, $p(y)$ and the parameters of the conditional $p(y|z, \theta)$. The E-M algorithm performs this minimization iteratively, where we assume an initial choice for θ. In the E-step, we choose a distribution $p(y)$, while holding θ fixed, so that $F(p, \theta^{\rm old})$ is minimized. In the subsequent M-step we hold p fixed, but choose θ in a way that $F(p^{\rm old}, \theta)$ is minimized. This procedure is iterated until the free energy does not change much with additional iterations, i.e., the scheme converges. In practical calculations, the connection with the variational free energy is not often used, but it is a useful concept to build up an intuitive understanding of the method. We refer the interested reader to Neal &

Hinton (1993) for a more detailed explanation. In the following we will describe the derivation of the E-M algorithm in the concrete context of finding the maximum likelihood solution of our photometric likelihood. For that we will use a different notation, however the intuition remains unchanged, if we associate the free energy (up to a sign) with the term $\mathcal{L}(q, \boldsymbol{\pi})$ in Eq. (A3).

To derive the Expectation-Maximization algorithm[21], we first introduce the parameter vector $\boldsymbol{\zeta}_i$ for each galaxy i that is a $N_{\rm tot}$ dimensional vector to indicate bin membership[22] in the '1-hot encoding' scheme. This means that for each galaxy, we have a $N_{\rm tot}$ dimensional binary vector, where ('0'/'1') indicates that the galaxy (resides/does not reside) in the respective redshift bin. For the remainder of this section we will work with the normed histogram bin heights $\boldsymbol{\pi} = \mathbf{n}^B \Delta z$ that parametrize the prior distribution over $z - \alpha$ as defined in Eq. 13. The prior distribution over $\overline{\boldsymbol{\zeta}}$ [23] is then given as

$$p(\overline{\boldsymbol{\zeta}}|\boldsymbol{\pi}) \propto \prod_{k=1}^{N_{\rm tot}} \prod_{n=1}^{N_{\rm gal}} \pi_k^{\zeta_{n,k}} \,, \tag{A1}$$

where $\sum_{k=1}^{N_{\rm tot}} \pi_k = 1$. Using $\overline{\boldsymbol{\zeta}}$ we can write the conditional distribution of the measured photometry given $\overline{\boldsymbol{\zeta}}$ as

$$p(\hat{\mathbf{F}}|\overline{\boldsymbol{\zeta}}, \boldsymbol{\pi}) \propto \prod_{\beta=1}^{N_{\rm gal}} \prod_{k=1}^{N_{\rm tot}} \left(\int_{z_L^k}^{z_R^k} {\rm d}z_\beta \int_{\alpha_L^k}^{\alpha_R^k} {\rm d}\alpha_\beta \, p(\mathbf{f}_\beta | \mathcal{T}(z_\beta, \alpha_\beta), \Sigma_\beta) \right)^{\zeta_{\beta,{\rm k}}} \,. \tag{A2}$$

It is important at this point to note that a marginalization over the parameter vectors ζ_i for all galaxies will yield the second, i.e. the likelihood, term in Eq. (15).

To derive the iterative optimization scheme we first consider the decomposition of the posterior as

$$\log p(\boldsymbol{\pi}|\hat{\mathbf{F}}) = \mathcal{L}(q, \boldsymbol{\pi}) + KL(q||p) + \log p(\boldsymbol{\pi}) - \log p(\hat{\mathbf{F}}) \,, \tag{A3}$$

where

$$\mathcal{L}(q, \mathbf{n^B}) = \sum_{\overline{\boldsymbol{\zeta}}} q(\overline{\boldsymbol{\zeta}}) \log \left(\frac{p(\hat{\mathbf{F}}, \overline{\boldsymbol{\zeta}})}{q(\overline{\boldsymbol{\zeta}})} \right) \tag{A4}$$

$$KL(q||p) = -\sum_{\overline{\boldsymbol{\zeta}}} q(\overline{\boldsymbol{\zeta}}) \log \left(\frac{p(\overline{\boldsymbol{\zeta}}|\hat{\mathbf{F}}, \boldsymbol{\pi})}{q(\overline{\boldsymbol{\zeta}})} \right) \tag{A5}$$

This decomposition implies an iterative scheme to maximize $\log p(\boldsymbol{\pi}|\hat{\mathbf{F}})$. Given an initial parameter vector $\boldsymbol{\pi}^{\rm old}$, we first minimize $KL(q||p)$ in the 'E-step' which directly implies $q(\overline{\boldsymbol{\zeta}}) = p(\overline{\boldsymbol{\zeta}}|\hat{\mathbf{F}}, \boldsymbol{\pi})$. In the 'M'-step we fix the distribution $q(\overline{\boldsymbol{\zeta}})$ and maximize $\mathcal{L}(q, \boldsymbol{\pi})$. This maximization directive is then given as

$$\mathcal{L}(q, \boldsymbol{\pi}) = S\left(\boldsymbol{\pi}, \boldsymbol{\pi}^{\mathbf{old}}\right) = \sum_{\overline{\boldsymbol{\zeta}}} p(\overline{\boldsymbol{\zeta}}|\hat{\mathbf{F}}, \boldsymbol{\pi}^{\mathbf{old}}) \log p(\hat{\mathbf{F}}, \overline{\boldsymbol{\zeta}}|\boldsymbol{\pi}) + \text{const.} \,, \tag{A6}$$

which is the expectation of the data log-likelihood with respect to $\overline{\boldsymbol{\zeta}}$. After a new parameter vector $\boldsymbol{\pi}^{\mathbf{new}}$ is obtained, we continue with the 'E'-step holding $\boldsymbol{\pi}^{\mathbf{new}}$ fixed. This process is continued until convergence. In the following we will derive the corresponding update equations.

[21] The interested reader will find the following derivation in analogy to the derivation to the E-M update equations for the Gaussian Mixture model (see Bishop 2006).

[22] We are referring to bins as defined in Eq. (13).

[23] Here, $\overline{\boldsymbol{\zeta}}$ denotes the collection of $\boldsymbol{\zeta}$ vectors of all galaxies.

E-step: Given an old parameter vector $\pi_{\rm old}$ we evaluate

$$p(\overline{\boldsymbol{\zeta}}|\hat{\mathbf{F}}, \boldsymbol{\pi}_{\rm old}) \propto \tag{A7}$$

$$\prod_{\beta=1}^{N_{\rm gal}} \prod_{j=1}^{N_{\rm tot}} \left(\pi_{\rm j,old} \int_{z_L^k}^{z_R^k} {\rm d}z_\beta \int_{\alpha_L^k}^{\alpha_R^k} {\rm d}\alpha_\beta \, p(\mathbf{f}_\beta | \mathcal{T}(z_\beta, \alpha_\beta), \Sigma_\beta) \right)^{\zeta_{\beta{\rm j}}} \,. \tag{A8}$$

M-step: In the maximization step of the algorithm we want to maximize the expected data log-likelihood with respect to the parameter $\overline{\boldsymbol{\zeta}}$. Given the updated posterior $p(\overline{\boldsymbol{\zeta}}|\hat{\mathbf{F}}, \boldsymbol{\pi}_{\rm old})$ this expectation is given as:

$$S\left(\boldsymbol{\pi}_{\rm new}, \boldsymbol{\pi}_{\rm old}\right) = \sum_{\beta=1}^{N_{\rm gal}} \sum_{j=1}^{N_{\rm tot}} E\left[\zeta_{\beta{\rm j}}\right] \times$$

$$\left(\log\left(\pi_{\beta{\rm j}}\right) + \log\left(\int_{z_L^j}^{z_R^j} {\rm d}z_\beta \int_{\alpha_L^j}^{\alpha_R^j} {\rm d}\alpha_\beta \, p(\mathbf{f}_\beta | \mathcal{T}(z_\beta, \alpha_\beta), \Sigma_\beta) \right) \right) \,, \tag{A9}$$

where

$$E\left[\zeta_{\beta{\rm j}}\right] = \frac{\sum_{\zeta_{n,k}} \zeta^{nk} p(\zeta_{nk}|\hat{\mathbf{F}}, \boldsymbol{\pi}_{\rm old})}{\sum_{\zeta_{nk}} p(\zeta_{n,k}|\hat{\mathbf{F}}, \boldsymbol{\pi}_{\rm old})}$$

$$= \frac{\pi_{\rm i,old} \int_{z_L^i}^{z_R^i} {\rm d}z^\beta \int_{\alpha_L^i}^{\alpha_R^i} {\rm d}\alpha^\beta \, p(\mathbf{f}_\beta | \mathcal{T}(z_\beta, \alpha_\beta), \Sigma_\beta)}{\sum_j \pi_{\rm j,old} \int_{z_L^j}^{z_R^j} {\rm d}z^\beta \int_{\alpha_L^j}^{\alpha_R^j} {\rm d}\alpha^\beta \, p(\mathbf{f}_\beta | \mathcal{T}(z_\beta, \alpha_\beta), \Sigma_\beta)} \tag{A10}$$

We optimize $S\left(\mathbf{nz}_{\rm z,t,new}, \mathbf{nz}_{\rm z,t,old}\right)$ under the constraint $\sum_k \pi_k = 1$ using the Lagrange multiplier formalism:

$$\overline{S}\left(\boldsymbol{\pi}_{\mathbf{new}}, \boldsymbol{\pi}_{\mathbf{old}}\right) = S\left(\boldsymbol{\pi}_{\mathbf{new}}, \boldsymbol{\pi}_{\mathbf{old}}\right) + \lambda \left(\sum_k \pi_k - 1 \right) \tag{A11}$$

Equating $\nabla_{\boldsymbol{\pi}} \overline{S}\left(\boldsymbol{\pi}_{\mathbf{new}}, \boldsymbol{\pi}_{\mathbf{old}}\right) == \mathbf{0}$, performing a summation over k, and using the summation constraint of $\boldsymbol{\pi}$, we obtain

$$-\lambda = N_{\rm gal} \,. \tag{A12}$$

This leads to the update equations for the E-M scheme that are iterated until we reach convergence in $\boldsymbol{\pi}$[24]:

$$N_k^t = \pi_k^{t-1} \sum_{i=1}^{N_{\rm gal}} \left(\frac{\int_{z_L^k}^{z_R^k} {\rm d}z_i \int_{\alpha_L^k}^{\alpha_R^k} {\rm d}\alpha_i \, p(\hat{\mathbf{f}}_\mathbf{i} | \mathcal{T}(z_i, \alpha_i), \Sigma_i)}{\sum_{j=1}^{N_{\rm tot}} \pi_j^{t-1} \int_{z_L^j}^{z_R^j} {\rm d}z_i \int_{\alpha_L^j}^{\alpha_R^j} {\rm d}\alpha_i \, p(\hat{\mathbf{f}}_\mathbf{i} | \mathcal{T}(z_i, \alpha_i), \Sigma_i)} \right) \tag{A13}$$

$$\pi_k^t = \frac{N_k^{t-1}}{\sum_k N_k^{t-1}} \,. \tag{A14}$$

In appendix B we will derive a Laplace approximation to the posterior based on this optimization scheme. We apply and discuss this scheme in § 4 and § 8.

[24] In practice we would iterate until the log-likelihood changes only by an extremely small amount.

APPENDIX B: DERIVING THE LAPLACE APPROXIMATION

In the previous appendix we derived an iterative scheme to obtain maximum likelihood estimates of the vector of normed histogram heights $\boldsymbol{\pi}_{\rm ML}$ based on the E-M algorithm. We note that the E-M algorithm is guaranteed to produce a maximum likelihood estimate $\boldsymbol{\pi}_{\rm ML}$ that lies on the simplex. The direct application of the Laplace approximation will effectively estimate Gaussian errors on the values. Applying this approximation around $\pi_{\rm ML}$ will lead to posteriors that reach to negative values, i.e. the posterior draws are not guaranteed to lie on the simplex. To extend the Laplace approximation to random variables that lie on the simplex, we first consider a mapping from simplex space to $\mathbb{R}^{N_{\rm bins}-1}$. This mapping is realized by the additive logistic transformation. Assume $\mathbf{y} \in \mathbb{R}^{N_{\rm bins}-1}$, we define the function

$$\boldsymbol{\pi}(\mathbf{y}) = \left(\frac{e^{y_1}}{1+\sum_{i=1}^{N_{\rm bins}-1} e^{y_i}}, \ldots, \frac{e^{y_{N_{\rm bins}-1}}}{1+\sum_{i=1}^{N_{\rm bins}-1} e^{y_i}}, \frac{1}{1+\sum_{i=1}^{N_{\rm bins}-1} e^{y_i}} \right)^T , \tag{B1}$$

with its inverse

$$\mathbf{y}(\boldsymbol{\pi}) = \left[\log\left(\pi_1/\pi_{N_{\rm bins}}\right), \ldots, \log\left(\pi_{N_{\rm bins}-1}/\pi_{N_{\rm bins}}\right) \right] . \tag{B2}$$

We see that the transformed variables $\mathbf{y}$ are now defined in real space and we perform the Laplace approximation as usual. Assuming a flat prior in logistic space, we can directly utilize the invariance of the Maximum Likelihood estimate under variable transformations (see e.g. Pawitan 2001) and approximate the posterior

$$p(\mathbf{y}|\hat{\mathbf{F}}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\rm ML}, \boldsymbol{\Sigma}) , \tag{B3}$$

where

$$\boldsymbol{\mu}_{\rm y\,ML} = \mathbf{y}(\boldsymbol{\pi}_{\rm ML}) , \tag{B4}$$

and

$$\boldsymbol{\Sigma}_{\mathbf{y}} = - \mathbf{H}^{-1}\Big|_{\mathbf{y}=\mathbf{y}_{\rm ML}} . \tag{B5}$$

Here $\mathbf{H}$ is the hessian of the log-likelihood (the second term in Eq. (15)) as a function of $\mathbf{y}$ evaluated at $\mathbf{y}(\boldsymbol{\pi}_{\rm ML})$.

The components of the hessian are given as

$$H_{az} = - \sum_{i=1}^{N_{\rm gal}} \frac{\left(\sum_j^{N_{\rm tot}} \left(\frac{\partial \pi_j}{\partial y_z}\right) I_{ij}\right)\left(\sum_{j=1}^{N_{\rm tot}} \left(\frac{\partial \pi_j}{\partial y_a}\right) I_{ij}\right)}{\left(\sum_{j=1}^{N_{\rm tot}} \pi_j I_{ij}\right)^2} \tag{B6}$$

$$+ \sum_{i=1}^{N_{\rm gal}} \frac{1}{\left(\sum_{j=1}^{N_{\rm tot}} \pi_j I_{ij}\right)} \sum_{j=1}^{N_{\rm tot}} \left(\frac{\partial^2 \pi_j}{\partial y_a \partial y_z}\right) I_{ij} , \tag{B7}$$

where

$$I_{ij} = \int_{z_L^j}^{z_R^j} \mathrm{d}z_i \int_{\alpha_L^j}^{\alpha_R^j} \mathrm{d}\alpha_i \; p(\hat{\mathbf{f}}_{\mathbf{i}}|\mathcal{T}(z_i, \alpha_i), \Sigma_i) . \tag{B8}$$

The first and second order derivatives are then evaluated to

$$\frac{\partial \pi^i}{\partial y_j} = \begin{cases} \pi_i(1-\pi_i), & i = j \wedge i < N_D \\ -\pi_i \pi_j, & i \neq j \wedge i < N_D \\ -\frac{\pi_j}{1+\sum_{z=1}^{D-1} \exp y_z}, & i = N_D \end{cases} \tag{B9}$$

and

$$\frac{\partial^2 \pi^i}{\partial y_\alpha \partial y_j} = \begin{cases} \frac{\partial \pi_i}{\partial y_\alpha} - 2\pi_i \frac{\partial \pi_i}{\partial y_\alpha}, & i = j \wedge i < N_D \\ -\frac{\partial \pi_i}{\partial y_\alpha}\pi_j - \pi_i \frac{\partial \pi_j}{\partial y_\alpha}, & i \neq j \wedge i < N_D \\ \frac{\pi_j \pi_\alpha - \frac{\partial \pi_j}{\partial y_\alpha}}{1+\sum_{z=1}^{D-1} \exp y_z}, & i = N_D \end{cases} \tag{B10}$$

Transformed into probability, or simplex, space, this posterior is then identified as a logit-normal distribution

$$p(\boldsymbol{\pi}|\hat{\mathbf{F}}) \approx \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_{\mathbf{y}}|}} \frac{1}{\prod_{i=1}^{N_{\rm bins}} \pi_i} \tag{B11}$$

$$\exp\left(-0.5 \left(\log\left(\frac{\boldsymbol{\pi}_{-\mathrm{N_{bins}}}}{\pi_{N_{\rm bins}}}\right) - \boldsymbol{\mu}_{\rm y,ML}\right) \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \left(\log\left(\frac{\boldsymbol{\pi}_{-\mathrm{N_{bins}}}}{\pi_{N_{\rm bins}}}\right) - \boldsymbol{\mu}_{\rm y,ML}\right)\right) . \tag{B12}$$

We note that the logit-normal is a probability distribution on the simplex, just as the Dirichlet. In fact, the Dirichlet can be approximated well by a logit-normal (Atchison & Shen 1980). However the logit-normal allows for a more complex covariance structure. The scheme developed in this appendix is applied and analysed in § 4 and § 8.

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