Draft version July 15, 2020
Preprint typeset using LaTeX style emulateapj v. 01/23/15

HOW TO OBTAIN THE REDSHIFT DISTRIBUTION FROM PROBABILISTIC REDSHIFT ESTIMATES

Alex I. Malz[1,2] and David W. Hogg[2,3,4,5]

Draft version July 15, 2020

ABSTRACT

A trustworthy estimate of the redshift distribution $n(z)$ is crucial for using weak gravitational lensing and large-scale structure of galaxy catalogs to study cosmology. Spectroscopic redshifts for the dim and numerous galaxies of next-generation weak-lensing surveys are expected to be unavailable, making photometric redshift (photo-z) probability density functions (PDFs) the next-best alternative for comprehensively encapsulating the nontrivial systematics affecting photo-z point estimation. The established stacked estimator of $n(z)$ avoids reducing photo-z PDFs to point estimates but yields a systematically biased estimate of $n(z)$ that worsens with decreasing signal-to-noise, the very regime where photo-z PDFs are most necessary. We introduce Cosmological Hierarchical Inference with Probabilistic Photometric Redshifts (CHIPPR), a statistically rigorous probabilistic graphical model of redshift-dependent photometry, which correctly propagates the redshift uncertainty information beyond the best-fit estimator of $n(z)$ produced by traditional procedures and is provably the only self-consistent way to recover $n(z)$ from photo-z PDFs. We present the `chippr` prototype code , noting that the mathematically justifiable approach incurs computational expense. The CHIPPR approach is applicable to any one-point statistic of any random variable, provided the prior probability density used to produce the posteriors is explicitly known; if the prior is implicit, as may be the case for popular photo-z techniques, then the resulting posterior PDFs cannot be used for scientific inference. We therefore recommend that the photo-z community focus on developing methodologies that enable the recovery of photo-z likelihoods with support over all redshifts, either directly or via a known prior probability density.

Keywords: cosmology: cosmological parameters — galaxies: statistics — gravitational lensing: weak — methods: data analysis — methods: statistical

1. INTRODUCTION

Photometric redshift (photo-z) estimation has been a staple of studies of galaxy evolution, large-scale structure, and cosmology since its conception half a century ago (Baum 1962). An extremely coarse spectrum in the form of photometry in a handful of broadband filters can be an effective substitute for the time- and photon-intensive process of obtaining a spectroscopic redshift (spec-z), a procedure that may only be applied to relatively bright galaxies. Once the photometric colors are calibrated against either a library of spectral energy distribution (SED) templates or a data set of spectra for galaxies with known redshifts, a correspondence between photometric colors and redshifts may be constructed, forming a trustworthy basis for photo-z estimation or testing.

Calculations of correlation functions of cosmic shears and galaxy positions that constrain the cosmological parameters require large numbers of high-confidence redshifts of surveyed galaxies. Many more photo-zs may be obtained in the time it would take to observe a smaller number of spec-zs, and photo-zs may be measured for galaxies too dim for accurate spec-z confirmation, permitting the compilation of large catalogs of galaxies spanning a broad range of redshifts and luminosities. Photo-zs have thus enabled the era of precision cosmology, heralded by weak gravitational lensing tomography and baryon acoustic oscillation peak measurements.

However, photo-zs are susceptible to inaccuracy and imprecision in the form of their inherent noisiness resulting from the coarseness of photometric filters, catastrophic errors in which galaxies of one SED at one redshift are mistaken for galaxies of another SED at a different redshift, and systematics introduced by observational techniques, data reduction processes, and training or template set limitations. Figure 1 is an adaptation of the ubiquitous plots of photo-z vs. spec-z illustrating the assumptions underlying photo-z estimation in general, that spec-zs are a good approximation to true redshifts and photo-zs represent special non-linear projections of observed photometry to a scalar variable that approximates the true redshift.

There are several varieties of generally non-Gaussian deviation from a trivial relationship between redshift and data in Figure 1, represented by a $y = x$ diagonal line. The coarseness of the photometric filters causes scatter about the diagonal, with larger scatter perpendicular to the diagonal at redshifts where highly identifiable spectral features pass between the filters, as well as higher scatter at high redshifts where faint galaxies with large photometric errors are more abundant. There are pop-

aimalz@nyu.edu
[1] German Centre of Cosmological Lensing, Ruhr-Universität, Universitätsstraße 150, 44801 Bochum, Germany
[2] Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, 9th floor, New York, NY 10003, USA
[3] Simons Center for Computational Astrophysics, 162 Fifth Avenue, 7th floor, New York, NY 10010, USA
[4] Center for Data Science, New York University, 60 Fifth Avenue, 7th floor, New York, NY 10003, USA
[5] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

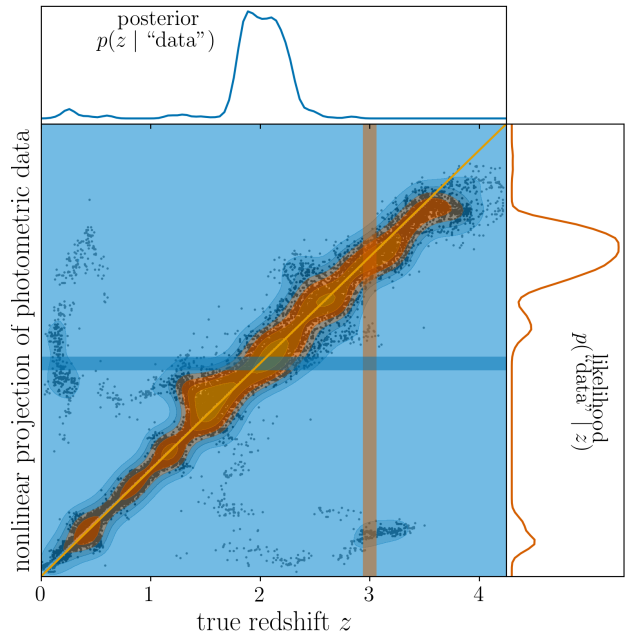

Figure 1. A generic probability space (darker in areas of higher probability density) of true redshift (x-axis) and a nonlinear projection of photometric data (y-axis), with vertical cuts and marginals (orange) indicating the construction of likelihoods and horizontal cuts and marginals (blue) indicating the construction of posteriors, with a theoretically perfect photo-z estimate on the diagonal (yellow) for reference. The data points were extracted using WebPlotDigitizer (Rohatgi 2019) from a spec-z vs. photo-z plot in Jain et al. (2015).

ulations of outliers, far from the diagonal, comprised of galaxies for which the redshift estimate is catastrophically distinct from the true redshift, showing that outliers are not uniformly distributed nor restricted to long tails away from a Gaussian scatter. And, though hardly perceptible in the plot, there is a systematic bias, wherein the average of the points would not lie on the diagonal but would be offset by a small bias, suggested by the trend of high-redshift points to lie below the diagonal.

Once propagated through the calculations of correlation functions of cosmic shear and galaxy positions, photo-z errors are a dominant contributor to the total uncertainties reported on cosmological parameters (Abruzzo & Haiman 2019). As progress has been made on the influence of other sources of systematic error, the uncertainties associated with photo-zs have come to dominate the error budget of cosmological parameter estimates made by current surveys such as DES (Hoyle et al. 2018), HSC (Tanaka et al. 2018), and KiDS (Hildebrandt et al. 2017). Based on the goals of a photometric galaxy survey, limits can be placed on the tolerance to these effects. For example, the Science Requirements Document (Mandelbaum 2017) states LSST's requirements for the main cosmological sample, reproduced in Table 1.

Much effort has been dedicated to improving photo-zs, though they are still most commonly obtained by a maximum likelihood estimator (MLE) based on libraries of galaxy SED templates, with conservative approaches to error estimation. The presence of galaxies whose SEDs are not represented by the template library tends to lead to catastrophic outliers distributed like the

Table 1
Photo-z requirements for LSST cosmology
(Mandelbaum 2017).

Number of galaxies	$\approx 10^7$
Root-mean-square error	$< 0.02(1+z)$
3σ catastrophic outlier rate	$< 10\%$
Canonical bias	$< 0.003(1+z)$

horizontally oriented population of Figure 1. For data-driven approaches, training sets that are incomplete in redshift coverage tend to result in catastrophic outliers like the vertically oriented population of Figure 1. The approaches of using a training set versus a template library are related to one another by Budavári (2009). Sophisticated Bayesian techniques and machine learning methods have been employed to improve precision (Carliles et al. 2010) and accuracy (Sadeh et al. 2016), while other advances have focused on identifying and removing catastrophic outliers when using photo-zs for inference (Gorecki et al. 2014).

The probability density function (PDF) in redshift space for each galaxy, commonly written as $\mathrm{p}(z)$, is an alternative to the MLE (with or without presumed Gaussian error bars) (Koo 1999). This option is favorable because it contains more potentially useful information about the uncertainty on each galaxy's redshift, incorporating our understanding of precision, accuracy, and systematic error. However, denoting photo-z PDFs as "$\mathrm{p}(z)$" is an abuse of notation, as it does not adequately convey what information is being used to constrain the redshift z; photo-z PDFs are *posterior* PDFs, conditioned on the photometric data and prior knowledge. In terms of Figure 1, photo-z PDFs are horizontal cuts, probabilities of redshift conditioned on a specific value of data, i.e. posteriors $\mathrm{p}(z \mid \vec{d})$, which constrain redshifts, whereas vertical cuts through this space are probabilities of data conditioned on a specific redshift, i.e. likelihoods $\mathrm{p}(\vec{d} \mid z)$, from which photometric data is actually drawn.

Photo-z posterior PDFs have been produced by completed surveys (Hildebrandt et al. 2012; Sheldon et al. 2012) and will be produced by ongoing and upcoming surveys (Abell et al. 2009; Carrasco Kind & Brunner 2014a; Bonnett et al. 2016; Masters et al. 2015). Photo-z posterior PDFs are not without their own shortcomings, however, including the resources necessary to calculate and record them for large galaxy surveys (Carrasco Kind & Brunner 2014b; Malz et al. 2018) and the divergent results of each method used to derive them (Hildebrandt et al. 2010; Dahlen et al. 2013; Sanchez et al. 2013; Bonnett et al. 2016; Tanaka et al. 2018). Though the matter is outside the scope of this paper, reviews of various methods have been presented in the literature (Sheldon et al. 2012; Ball et al. 2008; Carrasco Kind & Brunner 2013, 2014a; Schmidt et al. 2020). The most concerning weakness of photo-z posterior PDFs, however, is their usage in the literature, which is at best inconsistent and at worst incorrect.

Though their potential to improve estimates of physical parameters is tremendous, photo-z posterior PDFs have been applied only to a limited extent, most often by reduction to familiar point estimates. If the true redshifts $\{z_j^\dagger\}$ of galaxies j are known, then their redshift PDFs

are well-approximated by delta functions $\{\delta(z, z_j^\dagger)\}$ centered at the true redshift[6], and the redshift distribution is effectively approximated by a histogram or other interpolation of the delta functions $\{\delta(z, z_j^\dagger)\}$. When photo-$z$ posterior PDFs are available instead of true redshifts, the simplest approach reduces them to point estimates $\{\hat{z}_i\}$ of redshift by using $\delta(z, \hat{z}_j)$ in place of $\delta(z, z_j^\dagger)$. Though it is most common for $\hat{z}_j$ to be the maximum or *mode* of the photo-z posterior PDF, there are other, more principled point estimate reduction procedures (Tanaka et al. 2018).

Regardless of how it is done, any procedure that reduces photo-z posterior PDFs to point estimates discards valuable information about the uncertainty on redshift. Photo-z posterior PDFs have also been used to form selection criteria of samples from galaxy surveys without propagation through the calculations of physical parameters (van Breukelen & Clewley 2009; Viironen et al. 2015). Probability cuts on Bayesian quantities are not uncommon (Leung et al. 2017; DiPompeo et al. 2015), but that procedure does not fully take advantage of all information contained in a probability distribution for parameter inference.

The most prevalent application of photo-z posterior PDFs that preserves their information content is the estimation of the *redshift distribution function* $N(z)$, or, interchangably, its normalized cousin the *redshift density function* $n(z)$. $n(z)$ is used to calculate the redshift calibration bias b_z between the true and observed critical surface densities in galaxy-galaxy lensing (Mandelbaum et al. 2008) and the geometric lens efficiency $g_k(\chi)$ in tomographic weak lensing by large-scale structure (Benjamin et al. 2013). $N(z)$ may be used to validate survey selection functions used in generation of realistic, multipurpose mock catalogs (Norberg et al. 2002). As a key input to the traditional calculation of the power spectra of weak gravitational lensing and large-scale structure, the accuracy and precision to which $N(z)$ is estimated can strongly impact our constraints on the parameters of cosmological models (Bonnett 2015; Masters et al. 2015; Viironen et al. 2015; Asorey et al. 2016; Bonnett et al. 2016; Yang & Pullen 2018), so it is unsurprising that this last application dominates the canonical bias requirement of Table 1. Even with photo-zs adhering to the LSST requirements of Table 1, the degree to which constraints on the cosmological parameters can advance is limited by the accuracy and precision to which $n(z)$ is known (Abruzzo & Haiman 2019).

Though it is traditional to estimate $n(z)$ from photo-z point estimates (Abruzzo & Haiman 2019), it has become more common to use photo-z posterior PDFs directly to calculate the conceptually simple but mathematically inconsistent (Hogg 2012) *stacked estimator* $\hat{n}(z)$ of the redshift density function (Lima et al. 2008)

$$\hat{n}(z) = \frac{1}{J}\sum_{j=0}^{J} \mathrm{p}(z)_j \tag{1}$$

for a sample of J galaxies j, or, equivalently, the redshift distribution function $\hat{N}(z) = J\hat{n}(z)$, by effectively averaging the photo-z posterior PDFs. This summation procedure has been used extensively in cosmological analyses with photometric galaxy samples (Mandelbaum et al. 2008; Benjamin et al. 2013; Kelly et al. 2014).

Despite the growing prevalence of photo-z posterior PDF production, no implementation of inference using photo-z posterior PDFs has yet been presented with a mathematically consistent methodology. This paper challenges the logically invalid yet pervasive analysis procedure of stacking photo-z posterior PDFs by presenting and validating a hierarchical Bayesian technique for the use of photo-z posterior PDFs in the inference of $n(z)$, yielding a method applicable to arbitrary one-point statistics relevant to cosmology, large-scale structure, and galaxy evolution; future work will extend this methodology to higher-order statistics. We aim to develop a clear methodology guiding the use of photo-z posterior PDFs in inference so they may be utilized effectively by the cosmology community. Though others have approached the problem before (Leistedt et al. 2016; **?**), the method presented here differs in that it makes use of any existing catalog of photo-z posterior PDFs, rather than requiring a simultaneous derivation of the photo-z posterior PDFs and the redshift distribution, making it preferable to ongoing surveys for which there may be inertia preventing a complete restructuring of the analysis pipeline.

In Section 2, we present the CHIPPR model for characterizing the full posterior probability landscape of $N(z)$ using photo-z posterior PDFs. In Section 3, we present the `chippr` implementation of the CHIPPR model and the experimental set up by which we validate it, including the forward modeling of mock photo-z posterior PDFs. In Section 4, we present a number of informative test cases and compare the results of `chippr` with alternative approaches. In Section 5, we stress-test the CHIPPR model under nontraditional conditions. Finally, in Section 6, we make recommendations for future research involving $n(z)$ estimation.

2. MODEL

Consider a survey of J galaxies j, each with photometric data $\vec{d}_j$; thus the entire survey over some solid angle produces the ensemble of photometric magnitudes (or colors) and their associated observational errors $\{\vec{d}_j\}$. Each galaxy j has a redshift parameter z_j that we would like to learn. The distribution of the ensemble of redshift parameters $\{z_j\}$ may be described by the hyperparameters defining the redshift distribution function $n(z)$ that we would like to quantify. The redshift distribution function $n(z)$ is the number of galaxies per unit redshift, effectively defining the evolution in the number of galaxies convolved with the selection function of the sample (Ménard et al. 2013).

In Section 2.1, we establish a forward model encapsulating the causal relationship between $n(z)$ and photometry $\vec{d}$. In Section 2.2, we present the directed acyclic graph of this probabilistic generative model and interpret the corresponding mathematical expression, whose full derivation may be found in the Appendix. In Section 2.3, we summarize the necessary assumptions of the model.

[6] Note that spec-zs are not the same as known true redshifts; the PDFs of spec-zs would be narrow and almost always unimodal, but they would not be delta functions due to observational errors.

2.1. *Forward Model*

We begin by reframing the redshift distribution $n(z)$ from a probabilistic perspective. Here we define a redshift density $n(z)$ as the normalized probability density

$$\int_{-\infty}^{\infty} n(z)\, dz \;\equiv\; \frac{1}{J} \int_{-\infty}^{\infty} \sum_{j=1}^{J} \delta(z_j,\; z)\, dz = 1 \quad (2)$$

of finding a galaxy j in a catalog of J galaxies having a redshift z. We believe that galaxy redshifts are indeed sampled, or drawn, from $n(z)$, making it a probability density over redshift; this fact can also be confirmed by dimensional analysis of Equation 2, as suggested in Hogg (2012).

We may without loss of generality impose a parameterization

$$f(z; \boldsymbol{\phi}) \;\equiv\; n(z) \quad (3)$$

in terms of some parameter vector $\boldsymbol{\phi}$. At this point, the parameter vector is quite general and may represent coefficients in a high-order polynomial as a function of redshift, a set of means and variances defining Gaussians that sum to the desired distribution, a set of histogram heights that describe a binned version of the redshift distribution function, etc. Upon doing so, we may rewrite Equation 3 as

$$z_j \;\sim\; \mathrm{p}(z \mid \boldsymbol{\phi}) \;\equiv\; f(z; \boldsymbol{\phi}), \quad (4)$$

a probability density over redshift conditioned on the parameters $\boldsymbol{\phi}$ specifying $n(z)$. Note that z_j does not depend on the redshift $z_{j'}$ of some other galaxy $j' \neq j$, a statement of the causal independence of galaxy redshifts from one another.

In addition to believing $n(z)$ is a PDF from which redshifts are drawn, we also believe that there is some higher dimensional probability space $\mathrm{p}(z, \vec{d})$ of redshift z and photometric data vectors $\vec{d}$, which may be any combination of fluxes, magnitudes, colors, and their observational errors. Under this framework, $n(z)$ is equivalent to an integral

$$n(z) \;=\; \int \mathrm{p}(z, \vec{d}) \mathrm{d}\vec{d} \quad (5)$$

over the dimension of data in that joint probability space. Note that galaxies may have different observational data despite sharing the same redshift, and that galaxies at different redshifts may have identical photometry; the space $\mathrm{p}(z, \vec{d})$ need not be one-to-one. We assume a stronger version of statistical independence here, that draws $(z_j, \vec{d}_j)$ are independent of draws $(z_{j'}, \vec{d}_{j'})$ in this space; the data and redshift of each galaxy are independent of those of other galaxies.

However, this problem has additional causal structure that we can acknowledge. The photometry results from the redshifts, not the other way around. This is the fundamental assumption upon which photo-z estimation is based. The forward model corresponds to first drawing redshifts according to Equation 4 and then drawing data from the likelihood

$$\vec{d}_j \;\sim\; \mathrm{p}(\vec{d} \,|\, z_j) \quad (6)$$

of photometry conditioned on redshift, illustrated in Figure 1.

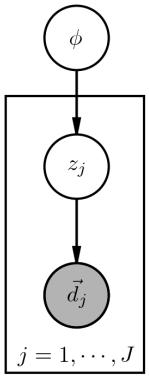

Figure 2. The directed acyclic graph of the CHIPPR model, where circles indicate random variables and arrows indicate causal relationships. The redshift distribution $n(z)$ parameterized by hyperparameters $\boldsymbol{\phi}$ exists independent of the survey of J galaxies, indicated as a box. The redshifts $\{z_j\}$ of all galaxies in the survey are latent variables independently drawn from the redshift distribution, which is a function of $\boldsymbol{\phi}$. The photometric data $\vec{d}_j$ for each galaxy is drawn from a function of its redshift z_j and observed, indicated by a shaded circle.

This description of the physical system corresponds to a forward model by which we actually believe photometry is generated:

1. There exists a redshift distribution $n(z)$ with parameters $\boldsymbol{\phi}$.
2. Galaxy redshifts $\{z_j\}$ are independent draws from $\mathrm{p}(z \mid \boldsymbol{\phi})$.
3. Galaxy photometry $\vec{d}_j$ is drawn from the likelihoods $\mathrm{p}(\vec{d}_j \mid z)$.

2.2. *Probabilistic Model*

A forward model such as that of Section 2.1 corresponds to a probabilistic graphical model (PGM), represented by a directed acyclic graph (DAG) as in Figure 2. A DAG conveys the causal relationships between physical parameters and, like a Feynman diagram in the context of particle physics, is a shorthand for mathematical relationships between variables. The photometric data $\vec{d}_j$ of a galaxy is drawn from some function of its redshift z_j, independent of other galaxies' data and redshift. Both data and redshift are random variables, but data is the one that we observe and redshift is not directly observable. In this problem, we don't care about further constraining the redshifts of individual galaxies, only the redshift distribution $n(z)$, so we consider redshift to be a *latent variable*. Because the parameters $\boldsymbol{\phi}$ that we seek are causally separated from the data by the latent variable of redshift, we call them *hyperparameters*.

The problem facing cosmologists is to determine the true value of $\boldsymbol{\phi}$ from observing the photometry $\{\vec{d}_j\}$ of a large sample of J galaxies j. To self-consistently propagate the uncertainty in the inference of redshift, however, it is more appropriate to estimate the posterior $\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})$ over all possible values of $\boldsymbol{\phi}$ conditioned on all the observed data $\{\vec{d}_j\}$ available in a generic catalog.

In order to use the DAG of Figure 2 to derive an expression for $\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})$ in terms of photo-z posterior PDFs, we must introduce two more concepts, confusingly named the *implicit prior* and the *prior probability density* (*prior PDF*), elaborated upon below.

When we constrain the redshift of a galaxy using its observed photometric data $\vec{d}_j$, we are effectively estimating a posterior $\mathrm{p}(z \mid \vec{d}_j)$, the probability of an unknown quantity conditioned on the quantity we have in hand, i.e the photometric data. This posterior is effectively a marginalization with respect to redshift at a given value of $\vec{d} = \vec{d}_j$ of the *empirical frequency distribution* $\mathrm{p}(z, \vec{d} \mid \boldsymbol{\phi}^\dagger)$, the joint probability density corresponding to the true redshift distribution parameterized by $\boldsymbol{\phi}^\dagger$, which exists in nature but need not be known.

As the hyperparameters $\boldsymbol{\phi}^\dagger$ of the true redshift distribution are in general unknown, the investigator seeking to estimate a posterior $\mathrm{p}(z \mid \vec{d}_j)$ must have a model ϕ^* for the general relationship between redshifts and photometry, whether empirical, as is the case for machine learning photo-z posterior PDF methods, or analytic, as is the case for template-based photo-z posterior PDF methods. If we were to marginalize over the photometry in $\mathrm{p}(\vec{d}, z)$, we would obtain a one-dimensional PDF $\mathrm{p}(z \mid \boldsymbol{\phi}^*)$ over redshift, which can by definition be parameterized by the same functional form as $n(z)$, for some $\boldsymbol{\phi}^*$ specific to the estimation procedure that may or may not bear any relation to the hyperparameters $\boldsymbol{\phi}^\dagger$ of the true $n(z)$. Rather, $\boldsymbol{\phi}^*$ is a consequence of the generative model for how photometry results from redshift, including the influence of intrinsic galaxy spectra and instrumental effects.

We call $\mathrm{p}(z \mid \boldsymbol{\phi}^*)$ the *implicit prior*, as it is rarely explicitly known nor chosen by the researcher[7] Because the implicit prior is unavoidable and almost inherently not uninformative, the photo-z posterior PDFs reported by any method must be *implicit posteriors* $\mathrm{p}(z \mid \vec{d}, \boldsymbol{\phi}^*)$ weighted by the implicit prior.

The prior probability density $\mathrm{p}(\boldsymbol{\phi})$ is a more familiar concept in astronomy; to progress, we will have to choose a prior probability density over all possible values of the hyperparameters $\boldsymbol{\phi}$. This prior need not be excessively proscriptive; for example, it may be chosen to enforce smoothness at physically motivated scales in redshift without imposing any particular region as over- or under-dense.

With inputs of the photo-z implicit posterior catalog $\{\mathrm{p}(z \mid \vec{d}, \boldsymbol{\phi}^*)\}$, the implicit prior $\mathrm{p}(z \mid \boldsymbol{\phi}^*)$, and the prior PDF $\mathrm{p}(\boldsymbol{\phi})$, we thus aim to obtain the posterior probability $\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})$ of the redshift density function given all the photometric data. By performing the derivation of the Appendix, we arrive at the desired expression

$$\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\}) \propto \mathrm{p}(\boldsymbol{\phi}) \int \prod_{j=1}^{J} \frac{\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)\mathrm{p}(z \mid \boldsymbol{\phi})}{\mathrm{p}(z \mid \boldsymbol{\phi}^*)} \mathrm{d}z, \quad (7)$$

[7] For template-based methods, the implicit prior is often an explicitly known input to the algorithm, engineered as an initial guess for the true $\boldsymbol{\phi}$, with an aim for a realistic choice guided by an earlier spectroscopic survey. (See Benítez (2000) for more detail.) It may thus be more appropriate to call it an *interim prior*, but we will use the former term throughout this paper for generality.

which is the very heart of CHIPPR, also given as Equation A.10. This in effect replaces the implicit prior with the sampled model hyperparameters, thereby converting the photo-z implicit posteriors into likelihoods in order to obtain unbiased posteriors.

2.3. *Model Limitations*

Finally, we explicitly review the assumptions made by this approach, which are as follows:

1. Photometric measurements of galaxies are statistically independent Poisson draws from the set of all galaxies such that Equation A.3 and Equation A.4 hold.

2. We take the reported photo-z implicit posteriors to be accurate, free of model misspecification; draws thereof must not be inconsistent with the distribution of photometry and redshifts. Furthermore, we must be given the implicit prior $\boldsymbol{\phi}^*$ used to produce the photo-z implicit posteriors.

3. We must assume a hyperprior distribution $\mathrm{p}(\boldsymbol{\phi})$ constraining the underlying probability distribution of the hyperparameters, which is informed by our prior beliefs about the true redshift distribution function.

These assumptions have known limitations. First, the photometric data are not a set of independent measurements; the data are correlated not only by the conditions of the experiment under which they were observed (instrument and observing conditions) but also by redshift covariances resulting from physical processes governing underlying galaxy spectra and their relation to the redshift distribution function. Second, the reported photo-z implicit posteriors may not be trustworthy; there is not yet agreement on the best technique to obtain photo-z posterior PDFs, and the implicit prior may not be appropriate or even known to us as consumers of photo-z implicit posteriors. Third, the hyperprior may be quite arbitrary and poorly motivated if the underlying physics is complex, and it can only be appropriate if our prior beliefs about $n(z)$ are accurate.

Furthermore, in Section 2.2, we have made an assumption of *support*, meaning the model $\mathrm{p}(z, \vec{d} \mid \boldsymbol{\phi})$ has mutual coverage with the parameter values that real galaxies can take. In other words, any probability distribution over the $(z, \vec{d})$ space must be nonzero where real galaxies can exist. Additionally, the hyperprior $\mathrm{p}(\boldsymbol{\phi})$ must be nonzero at the hyperparameters $\boldsymbol{\phi}^\dagger$ of the true redshift density function $n(z)$. This assumption cannot be violated under the experimental design of Section 2.1, but it is not generically guaranteed when performing inference on real data; thus the chosen $\mathrm{p}(z, \vec{d} \mid \boldsymbol{\phi}^*)$ and $\mathrm{p}(\boldsymbol{\phi})$ must be sufficiently general as to not rule out plausible areas of parameter space.

3. METHODS & DATA

Here we describe the method by which we demonstrate the CHIPPR model. In Section 3.1, we outline the implementation of the `chippr` code. In Section 3.2, we outline the procedure for emulating mock photo-z implicit posteriors.

3.1. *Implementation*

We implement the CHIPPR model in code in order to perform tests of its validity and to compare its performance to that of traditional alternatives. In Section 3.1.1, we describe the publicly available `chippr` library. In Section 3.1.2, we introduce the alternative approaches evaluated for comparison with CHIPPR. In Section 3.1.3, we describe the diagnostic criteria by which we assess estimators of $n(z)$.

3.1.1. *Code*

`chippr` is a *Python* 2 library[8] that includes an implementation of the CHIPPR model as well as an extensive suite of tools for comparing CHIPPR to other approaches.

Though there are plans for future expansion to more flexible parameterizations, the current version of `chippr` uses a log-space piecewise constant parameterization

$$f(z;\boldsymbol{\phi}) = \exp[\phi^k] \text{ if } z^k < z < z^{k+1} \tag{8}$$

for $n(z)$ and every photo-z posterior PDF, satisfying

$$\sum_{k=1}^{K} \exp[\phi^k]\delta z^k = 1 \tag{9}$$

with K bins of width $\delta z^1, \ldots, \delta z^K$ defined by endpoints $z^0, \ldots, z^K$. Thus each $\mathrm{p}(z \mid \vec{d}_j) = f(z;\boldsymbol{\phi}_j)$ has parameters $\boldsymbol{\phi}_j$ that are defined in the same basis as those of $n(z)$. To infer the full log-posterior distribution $\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})]$, one must provide a plaintext file with $K+1$ redshift bin endpoints $\{z_k\}$, the parameters $\boldsymbol{\phi}^*$ of the implicit log-prior, and the parameters $\{\boldsymbol{\phi}_j\}$ of the log-posteriors $\ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)]$.

The `emcee`(Foreman-Mackey et al. 2013) implementation of ensemble sampling is used to sample the full log-posterior of Equation A.10. `chippr` accepts a configuration file of user-specified parameters, among them the number W of walkers. At each iteration i and for each walker, a proposal distribution $\hat{\boldsymbol{\phi}}_i$ is drawn from the log-prior distribution and evaluated for acceptance to or rejection from the full log-posterior distribution.

The resulting output includes $\frac{I_0}{s}$ accepted samples $\boldsymbol{\phi}_i$ for a pre-specified chain thinning factor s and their full posterior probabilities $\mathrm{p}(\boldsymbol{\phi}_i \mid \{\vec{d}_j\})$, as well as the autocorrelation times and acceptance fractions calculated for each element of $\boldsymbol{\phi}$, divided into separate files before and after the completion of the burn-in phase, as defined by the Gelman-Rubin statistic (Gelman & Rubin 1992).

3.1.2. *Alternative approaches for comparison*

In this study, we compare the results of Equation 7 to those of the two most common approaches to estimating $n(z)$ from a catalog of photo-z posterior PDFs: the distribution $n(z_{\mathrm{max}})$ of the redshifts at maximum posterior probability

$$f^{MMAP}(z;\hat{\boldsymbol{\phi}}) = \sum_{j=1}^{J} \delta(z, \mathrm{mode}[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)]) \tag{10}$$

(i.e. the distribution of modes of the photo-z posterior PDFs) and the stacked estimator of Equation 11, which can be rewritten as

$$f^{stack}(z;\hat{\boldsymbol{\phi}}) = \sum_{j=1}^{J} \mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*) \tag{11}$$

in terms of the implicit photo-z posteriors we have. These two approaches have been compared to one another by Hildebrandt et al. (2012), Benjamin et al. (2013), and Asorey et al. (2016) in the past but not to CHIPPR.

Point estimation converts the implicit photo-z posteriors $\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)$ into delta functions with all probability at a single estimated redshift. Some variants of point estimation choose this single redshift to be that of maximum a posteriori probability $\mathrm{mode}[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)]$ or the expected value of redshift $\langle z \rangle = \int z\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)\mathrm{d}z$. Tanaka et al. (2018) directs attention to deriving an optimal point estimate reduction of a photo-z posterior PDF, but since the purpose of this paper is to compare against the most established alternative estimators of $n(z)$, its use will be postponed until a future study. Stacking these modified photo-z posterior PDFs leads to the marginalized maximum a posteriori (MMAP) estimator and the marginalized expected value (MExp) estimator, though only the former is included in this study since the latter has fallen out of favor in recent years[9].

It is worth discussing the relationship between point estimation and stacking. When the point estimator of redshift is equal to the true redshift, stacking delta function photo-z posterior PDFs will indeed lead to an accurate recovery of the true redshift distribution function. However, stacking is in general applied indiscriminately to broader photo-z posterior PDFs and imperfect point estimators of redshift. It is for these reasons that alternatives are considered here.

A final estimator of the hyperparameters is the maximum marginalized likelihood estimator (MMLE), the value of $\boldsymbol{\phi}$ maximizing the log posterior given by Equation A.10 using any optimization code. The MMLE can be obtained in substantially less time than enough samples to characterize the full log-posterior distribution of $n(z)$. However, the MMLE yields only a point estimate of $n(z)$ rather than characterizing the full log-posterior on $\boldsymbol{\phi}$, and it does not escape the dependence on the choice of hyperprior distribution. Furthermore, derivatives will not in general be available for the full posterior distribution, restricting optimization methods used, and, as is true for any optimization code, there is a risk of numerical instability.

3.1.3. *Performance metrics*

The results of the computation described in Section 3.1 are evaluated for accuracy on the basis of some quantitative measures. Beyond visual inspection of samples, we calculate summary statistics to quantitatively compare different estimators' precision and accuracy. Since

[8] `https://github.com/aimalz/chippr`

[9] And for good reason! Consider a bimodal photo-z posterior PDF; its expected value may very well fall in a region of very low probability, yielding a less probable point estimate than the point at which either peak achieves its maximum.

MCMC samples of hyperparameters are Gaussian distributions, we can quantify the breadth of the distribution for each hyperparameter using the standard deviation regardless of whether the true values are known.

In simulated cases where the true parameter values are known, we calculate the Kullback-Leibler divergence (KLD), given by

$$KL_{\boldsymbol{\phi},\boldsymbol{\phi}^\ddagger} = \int \mathrm{p}(z \mid \boldsymbol{\phi}) \ln \left[\frac{\mathrm{p}(z \mid \boldsymbol{\phi})}{\mathrm{p}(z \mid \boldsymbol{\phi}^\dagger)} \right] \mathrm{d}z, \qquad (12)$$

which measures a distance from parameter values $\boldsymbol{\phi}$ to true parameter values $\boldsymbol{\phi}^\dagger$. The KLD is a measure of the information loss, in units of nats, due to using $\boldsymbol{\phi}$ to approximate the true $\boldsymbol{\phi}^\dagger$ when it is known. A detailed exploration of the KLD may be found in the Appendix to Malz et al. (2018).

3.2. *Validation on mock data*

We compare the results of CHIPPR to those of stacking and the histogram of photo-z posterior PDF maxima (modes) on mock data in the form of catalogs of emulated photo-z posterior PDFs generated via the forward model discussed in Section 2.1. Figure 3 illustrates the implementation of the forward model, defined by the much simpler Figure 2, used for validating the method presented here. The irony of a simple model and complex validation procedure is not lost on the authors.

Figure 3 outlines the four phases of the generative model, which uses a total of three inputs. The experimental design requires our choice of true values $\phi^\dagger$ of the hyperparameters governing $n(z)$, a photo-z model $\mathrm{p}(z, \vec{d})$ defining the space of redshift and photometry, and prior values ϕ^* of the hyperparameters of $n(z)$. In the first phase, we sample $J = 10^4$ redshifts $z_j^\dagger \sim \mathrm{p}(z \mid \phi^\dagger)$. In the second phase, we evaluate the photo-z model at those redshifts, yielding a set of J likelihoods $\mathrm{p}(\vec{d} \mid z_j^\dagger)$, from which we then sample data $\vec{d}_j^\dagger \sim \mathrm{p}(\vec{d} \mid z_j^\dagger)$ for each galaxy. In the third phase, we evaluate the photo-z model at that data to obtain J posteriors $\mathrm{p}(z \mid \vec{d}_j^\dagger)$. In the fourth phase, we convolve the posteriors with the chosen prior $\mathrm{p}(z \mid \phi^*)$, yielding implicit posteriors $\mathrm{p}(z \mid \vec{d}_j^\dagger, \phi^*)$.

The true redshift distribution used in these tests is a particular instance of the gamma function

$$n^\dagger(z) = \frac{1}{2c_z} \left(\frac{z}{c_z} \right)^2 \exp\left[-\frac{z}{c_z} \right] \qquad (13)$$

with $c_z = 0.3$, because it has been used in forecasting studies for DES and LSST.

The mock data emulates the three sources of error of highest concern to the photo-z community that are explored in detail later in this section: intrinsic scatter (Section 4.1), catastrophic outliers (Section 4.2), and canonical bias (Section 4.3). Figure 4 illustrates these three effects simultaneously at the tolerance of LSST for demonstrative purposes, harking back to Figure 1.

The hyperprior distribution chosen for these tests is a multivariate normal distribution with mean $\vec{\mu}$ equal to the implicit prior $\boldsymbol{\phi}^*$ and covariance

$$\Sigma_{k,k'} = q \ \exp[-\frac{e}{2} \ (\bar{z}_k - \bar{z}_{k'})^2] \ + \ t\delta(k,k') \qquad (14)$$

inspired by one used in Gaussian processes, where k and k' are indices ranging from 1 to K and $q = 1.0$, $e = 100.0$, and $t = q \cdot 10^{-5}$ are constants chosen to permit draws from this prior distribution to produce shapes similar to that of a true $\tilde{\boldsymbol{\phi}}$. We adapt the full log-posterior of Equation A.10 to the chosen binning of redshift space.

The sampler is initialized with $W = 100$ walkers each with a value chosen from a Gaussian distribution of identity covariance around a sample from the hyperprior distribution.

4. RESULTS

Here, we compare the results of the CHIPPR methodology with those of established $n(z)$ estimators under the three traditional measures of photo-z uncertainty one at a time: Section 4.1 concerns the redshift-dependent intrinsic scatter, Section 4.2 concerns realistically complex catastrophic outlier populations, and Section 4.3 concerns the canonical bias in the mean redshift.

4.1. *Intrinsic scatter*

Figure 5 shows some examples of photo-z posterior PDFs generated with only the systematic of intrinsic scatter, at the level of the LSST requirements on the left and twice that on the right. One can see that the histogram of redshift estimates is broader than that of true redshifts, and that the effect is substantially more pronounced by just doubling the intrinsic scatter from the level of the LSST requirements.

Figure 6 shows the $n(z)$ recovered by CHIPPR and the alternative approaches. As expected, the estimates of $n(z)$ based on the modes of the photo-z posterior PDFs and stacking are broader than the marginalized maximum likelihood estimator from `chippr`, with more broadening as the intrinsic scatter increases. CHIPPR's marginalized maximum likelihood estimate is robust to intrinsic scatter and is unaffected by increased intrinsic scatter, though the CHIPPR posterior distribution on the redshift distribution is itself broader for the higher intrinsic scatter case than for the LSST requirements. The broadening of the alternative estimators corresponds to a loss of 3-4 times as many nats of information about $n(z)$ for the LSST requirements relative to the marginalized maximum likelihood estimate of CHIPPR.

4.2. *Catastrophic outliers*

As was covered in Section 1, catastrophic outliers tend to be distributed non-uniformly across the space of observed and true redshift. However, the LSST requirements do not specify details for a distribution of outliers to which they were tuned, and it is still instructive to examine the impact of uniform outliers on the inference of $n(z)$, so we begin by addressing uniformly distributed outliers before considering more realistic outlier distributions.

A uniformly distributed population of outliers was simulated by giving every sample in true redshift a 10% chance of having an observed redshift drawn from a uniform distribution rather than the Gaussian about the true redshift. Though this results in slightly less than the 10% catastrophic outlier rate, it can be done independently of the definition of the standard deviation so was implemented for demonstrative purposes. Figure 7 shows

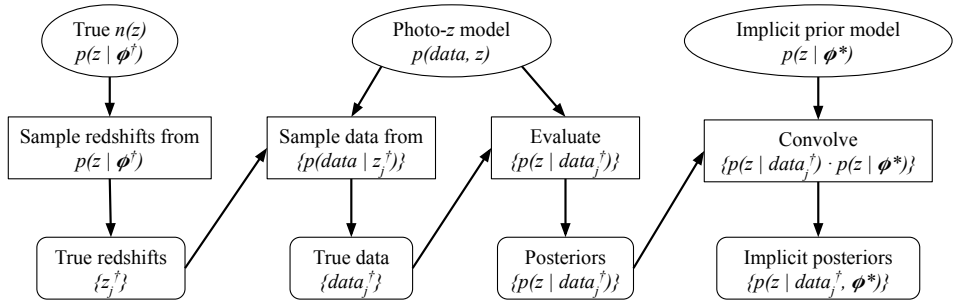

Figure 3. A flow chart illustrating the forward model used to generate mock data in the validation of CHIPPR, as described in Section 2.1. Ovals indicate a quantity that must be chosen in order to generate the data, rectangles indicate an operation we perform, and rounded rectangles indicate a quantity created by the forward model. Arrows indicate the inputs and outputs of each operation performed to simulate mock photo-z posterior PDF catalogs.

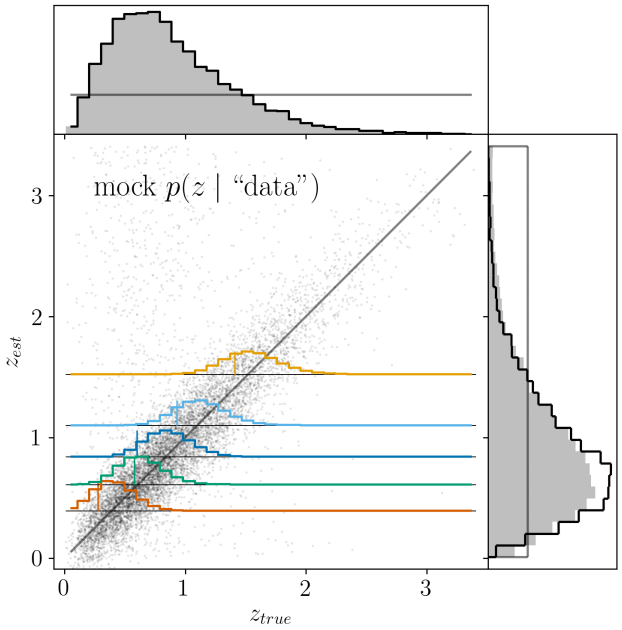

Figure 4. The joint probability space of true and estimated redshift for the three concerning photo-z systematics at the level of the LSST requirements: intrinsic scatter, uniformly distributed catastrophic outliers, and bias. The main panel shows samples (black points) in the space of mock data and redshift, akin to the standard scatterplots of true and estimated redshift, the $z_{\rm spec} = z_{\rm phot}$ diagonal (gray line), and posterior probabilities evaluated at the given estimated redshift (colored step functions). The insets show marginal histograms (light gray) in each dimension, that can be compared with the true $n(z)$ used to make the figure (black) to see the effect of these systematics, as well as the implicit prior (dark gray).

examples of photo-z posterior PDFs from a uniformly distributed outlier population at the level of the LSST requirements (left) as well as the results of CHIPPR and other $n(z)$ estimation methods (right). The intrinsic scatter of the tests in this section does not increase with redshift as indicated in Table 1 in order to isolate the effect of outliers, and is instead held at a constant $\sigma_z = 0.02$.

Figure 7 shows that at the level of the LSST requirements, the alternative estimators are overly broad, whereas CHIPPR's marginalized maximum likelihood estimate yields an unbiased estimate of $n(z)$. Further, the result of stacking is even broader than that of the histogram of modes, corresponding to ten times the information loss of CHIPPR's marginalized maximum likelihood estimate, making it worse than the most naive reduction of photo-z posterior PDFs to point estimates.

When one thinks of the photo-z posterior PDFs of catastrophic outliers, however, what comes to mind is multimodal photo-z posterior PDFs, wherein reducing photo-z posterior PDFs to point estimates to make a standard scatterplot of the true and observed redshifts leads to substantial probability density off the diagonal. These coordinated catastrophic outliers may be emulated in the joint probability space of true and estimated redshifts by using a mixture of the unbiased diagonal defined by the intrinsic scatter and an additional Gaussian in one dimension, with constant observed redshift for a template-fitting code and constant true redshift for a machine learning code.

In the case of a catastrophic outlier population like that anticipated of template-fitting codes, 10% of all galaxies have their observed redshift at a particular value unrelated to their true redshift, illustrated in the left panel of Figure 8. This case is subject to the same caveat as the uniformly distributed outliers when it comes to the LSST requirement. It is less straightforward to emulate catastrophic outliers like those anticipated of a machine learning code, those that are truly multimodal. The testing conditions here, illustrated in the right panel of Figure 8, gives 10% of galaxies at the redshift affected by outliers an observed redshift that is uniformly distributed relative to the true redshift, meaning that far fewer than 10% of all galaxies in the sample are catastrophic outliers.

The results of CHIPPR and the alternative estimators of $n(z)$ are presented in Figure 9. The most striking feature is that the histogram of modes is highly sensitive to both outlier populations, producing a severe overestimate in the case of an outlier population like those seen in template-fitting codes and a severe underestimate in the case of an outlier population like those seen in machine learning codes, corresponding to a twenty-fold loss of information compared to the CHIPPR marginalized maximum likelihood estimate in both cases. The effect on the stacked estimator of $n(z)$ is more subtle though still concerning. In the case of outliers like those resulting from template-fitting, the stacked estimator is overly broad even without realistic intrinsic scatter, resulting in ten times the information loss compared to the CHIPPR marginalized maximum likelihood estimate,

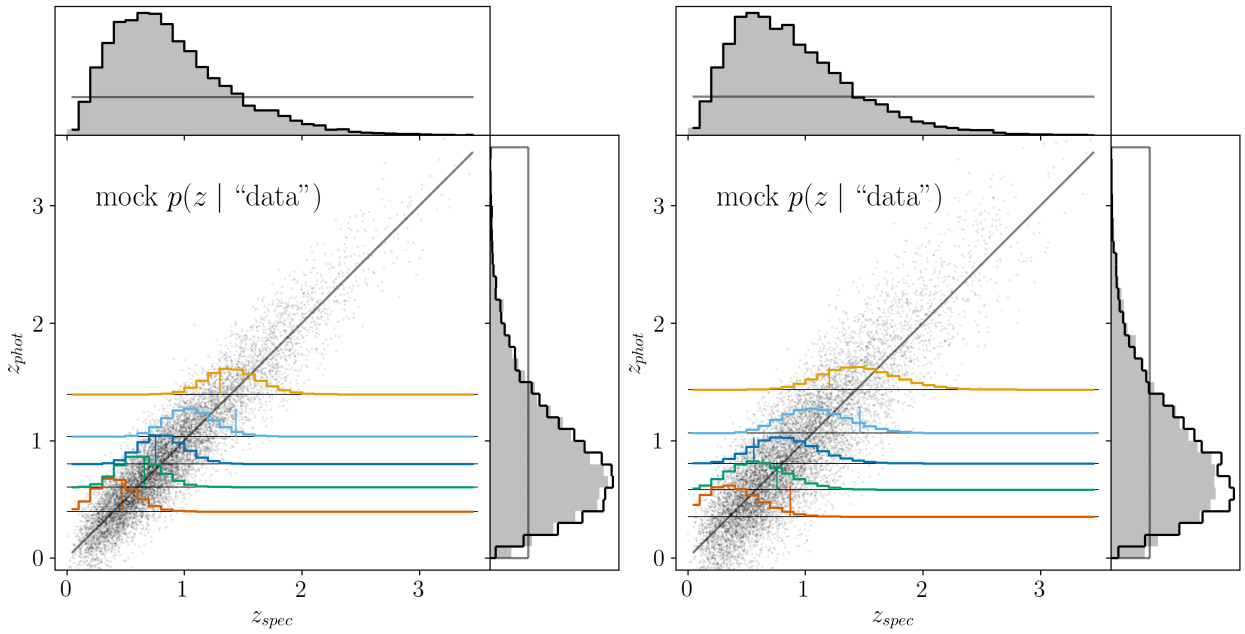

Figure 5. Examples of mock photo-z posterior PDFs generated with intrinsic scatter at the LSST requirements (left) and twice the LSST requirements (right), including samples from the probability space of true and observed redshift (black points), photo-z posterior PDFs (colored step functions), and the true redshifts of the example photo-z posterior PDFs (colored vertical lines). A histogram (light gray) of points in each dimension is shown in the respective inset, with the true redshift distribution (black) and implicit prior (dark gray).

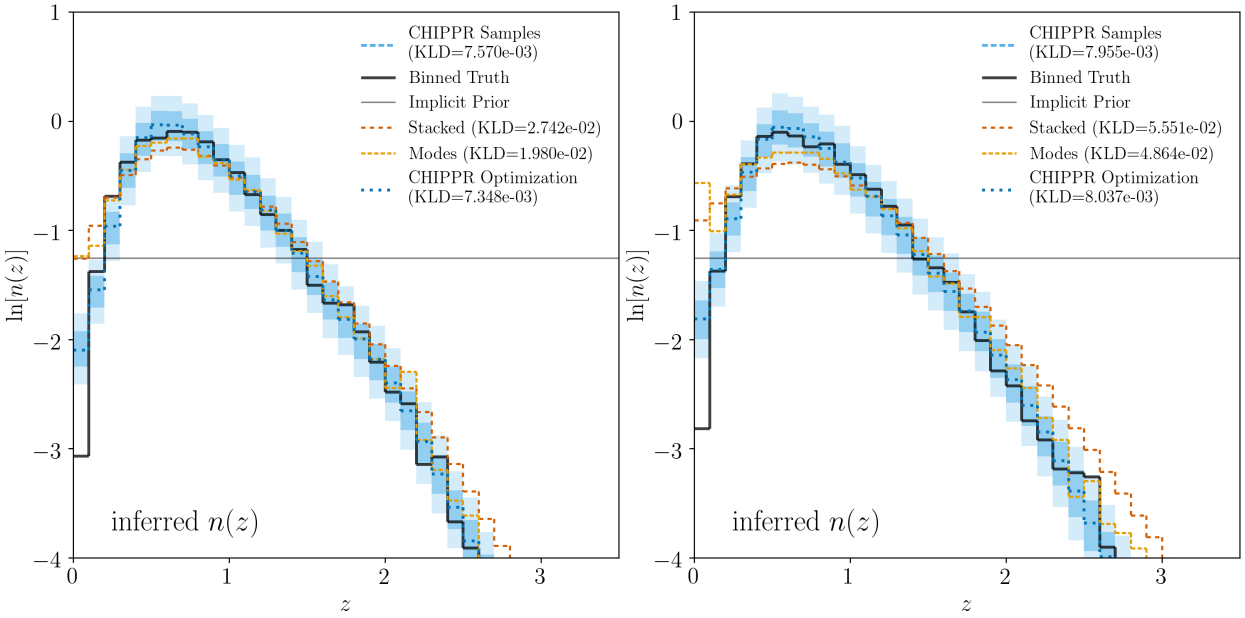

Figure 6. The results of CHIPPR (samples in light blue and optimization in dark blue) and the alternative approaches (the stacked estimator in red and the histogram of modes in yellow) on photo-z posterior PDFs with intrinsic scatter of the LSST requirements (left) and twice that (right), with the true redshift density (black curve) and implicit prior (gray curve). CHIPPR is robust to intrinsic scatter, but the alternatives suffer from overly broad $n(z)$ estimates that worsen with increasing intrinsic scatter.

and in the case of outliers like those resulting from machine learning, the stacked estimator features an overestimate at the redshift affected by the outlier population, resulting in about five times the information loss as the CHIPPR marginalized maximum likelihood estimate. The CHIPPR marginalized maximum likelihood estimate, however, appears unbiased and withstands these effects, and the breadth of the distribution of samples of $n(z)$ is invariant.

4.3. *Canonical bias*

Systematic bias in photo-z point estimates, is a concern for LSST's cosmology results, for the same reasons explored in Hoyle et al. (2018). This form of bias is typically summarized by a shift parameter $\Delta_z = (\langle \mathrm{p}(z \mid \hat{\boldsymbol{\phi}}) \rangle - \langle \mathrm{p}(z \mid \boldsymbol{\phi}^\dagger) \rangle)$ representing a difference between the first moment of the estimated redshift density function

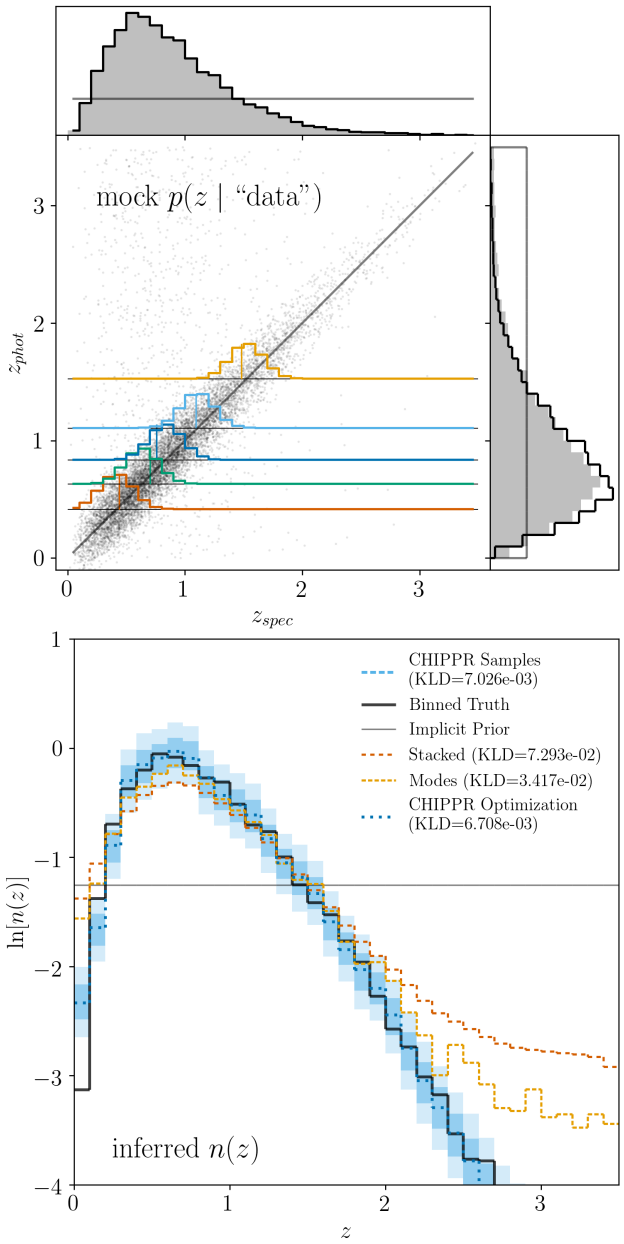

Figure 7. Top: Examples of photo-z posterior PDFs with a uniformly distributed catastrophic outlier population at the level of the LSST requirements, including samples from the probability space of true and observed redshift (black points), photo-z posterior PDFs (colored step functions), and the true redshifts of the example photo-z posterior PDFs (colored vertical lines), with marginal histograms (light gray) for each dimension with the true redshift distribution (black) and implicit prior (dark gray) in the insets. Bottom: The results of CHIPPR (samples in light blue, optimization in dark blue) and the alternative approaches (the stacked estimator in red, the histogram of modes in yellow) on photo-z posterior PDFs with uniformly distributed catastrophic outliers, with the true redshift density (black curve) and implicit prior (gray curve). The presence of the catastrophic outlier population broadens the histogram of modes and stacked estimator of the redshift distribution, but the result of CHIPPR is unbiased.

and that of the true redshift density function. To distinguish other aforementioned manifestations of bias from this common form of bias, we refer to Δ_z as the *canonical bias*.

In the context of photo-z posterior PDFs, the canonical bias represents an instance of model misspecification. Consider that if the canonical bias were included in the framework of Figure 1, it could be trivially modeled out as a simple linear transformation of $z_{\rm phot} \rightarrow z_{\rm phot} - \Delta_z(1 + z_{\rm phot})$ of the $(z_{\rm spec}, z_{\rm phot})$ space. Regardless, for completeness, a test at ten times the canonical bias of the LSST requirements, with no redshift-dependent intrinsic scatter nor catastrophic outliers, is provided in Figure 10.

As expected based on self-consistency of the forward-modeled photo-z posterior PDFs, CHIPPR is immune to linear bias of the form of Δ_z. Furthermore, the alternative estimators are only weakly affected, with information loss two and four times greater than that of the CHIPPR marginalized maximum likelihood estimate for the histogram of modes and stacked estimator respectively. (This general robustness may suggest that the canonical bias may not be the most relevant measure of performance of estimators of $n(z)$.)

5. DISCUSSION

The experiments of Section 4 quantify the influence on each estimator of $n(z)$ due to each of the canonical types of photo-z error one at a time in isolation. Now, we stress-test CHIPPR by exploring the impact of the implicit prior, which has thus far not received much attention in the literature. Section 5.1 demonstrates the sensitivity of $n(z)$ estimation methods to realistically complex implicit priors, and Section 5.2 demonstrates the consequences of mischaracterization of the implicit prior used to generate the photo-z posterior PDF catalog. These results provide compelling motivation for the photo-z community to prioritize the study of implicit priors of existing and developing photo-z posterior PDF techniques.

5.1. *Realistically complex implicit prior*

`chippr` can handle any implicit prior with support over the redshift range where $n(z)$ is defined, but some archetypes of implicit prior are more likely to be encountered in the wilds of photo-z posterior PDF codes. Ideally, an uninformative implicit prior would be used, although it may be complicated to compute from the covariances of the raw data. Template-fitting codes have an explicit prior input formed by redshifting a small number of templates, leading to a highly nonuniform but physically-motivated interim prior. Machine learning approaches tend to be trained on previously observed data sets that are biased towards low redshift, which biases the implicit prior towards low redshift. Some efforts have been made to modify an observationally informed implicit prior so that it is more representative of the photometric data for which redshifts are desired (Sheldon et al. 2012), but, unless it is equal to the true $n(z)$, it will propagate to the results of traditional $n(z)$ estimation methods.

Figure 11 shows examples of photo-z posterior PDFs with a low-redshift favoring implicit prior emulating that of a machine learning approach to photo-z estimation (left panel) and a more complex interim prior emulating

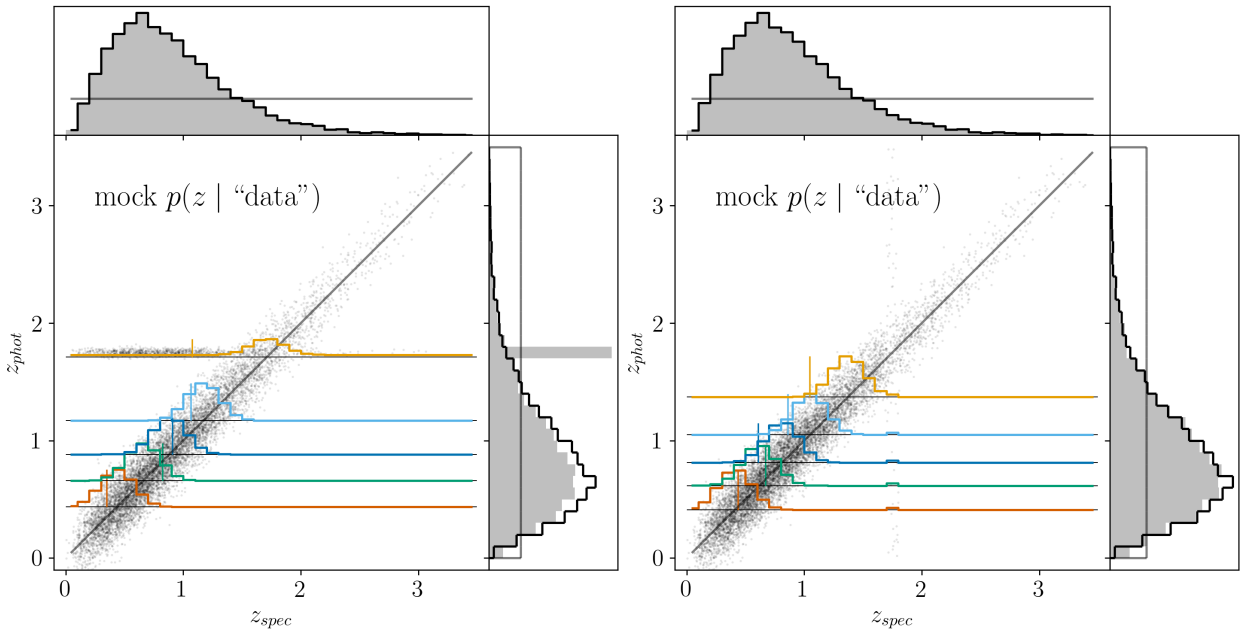

Figure 8. Examples of photo-z posterior PDFs with a catastrophic outlier population like that seen in template-fitting photo-z posterior PDF codes (left) and machine learning photo-z posterior PDF codes (right), including samples from the probability space of true and observed redshift (black points), photo-z posterior PDFs (colored step functions), and the true redshifts of the example photo-z posterior PDFs (colored vertical lines), with marginal histograms (light gray) for each dimension with the true redshift distribution (black) and implicit prior (dark gray) in the insets.

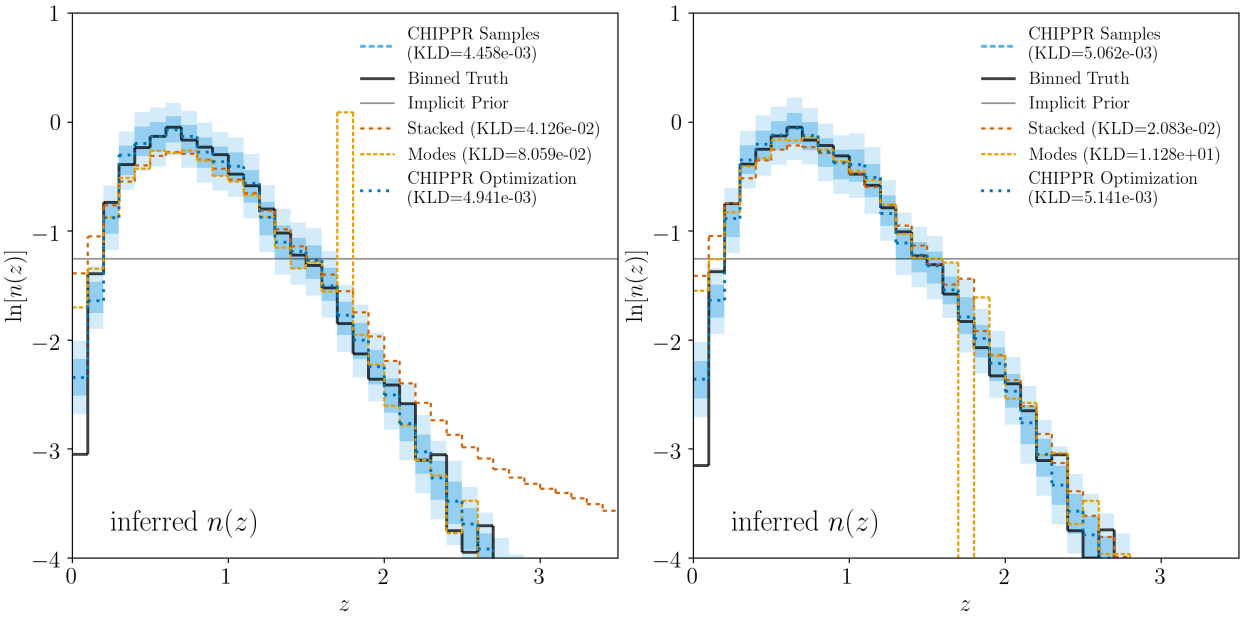

Figure 9. The results of CHIPPR (samples in light blue and optimization in dark blue) and the alternative approaches (the stacked estimator in red, the histogram of modes in yellow) on photo-z posterior PDFs with catastrophic outliers like those seen in template-fitting photo-z posterior PDF codes (left) and machine learning photo-z posterior PDF codes (right) to the LSST requirements, with the true redshift density (black curve) and implicit prior (gray curve). Though the histogram of modes is most sensitive to a catastrophic outlier population, the stacked estimator also overestimates $n(z)$ under (machine learning-like outliers) and beyond (template fitting-like outliers).

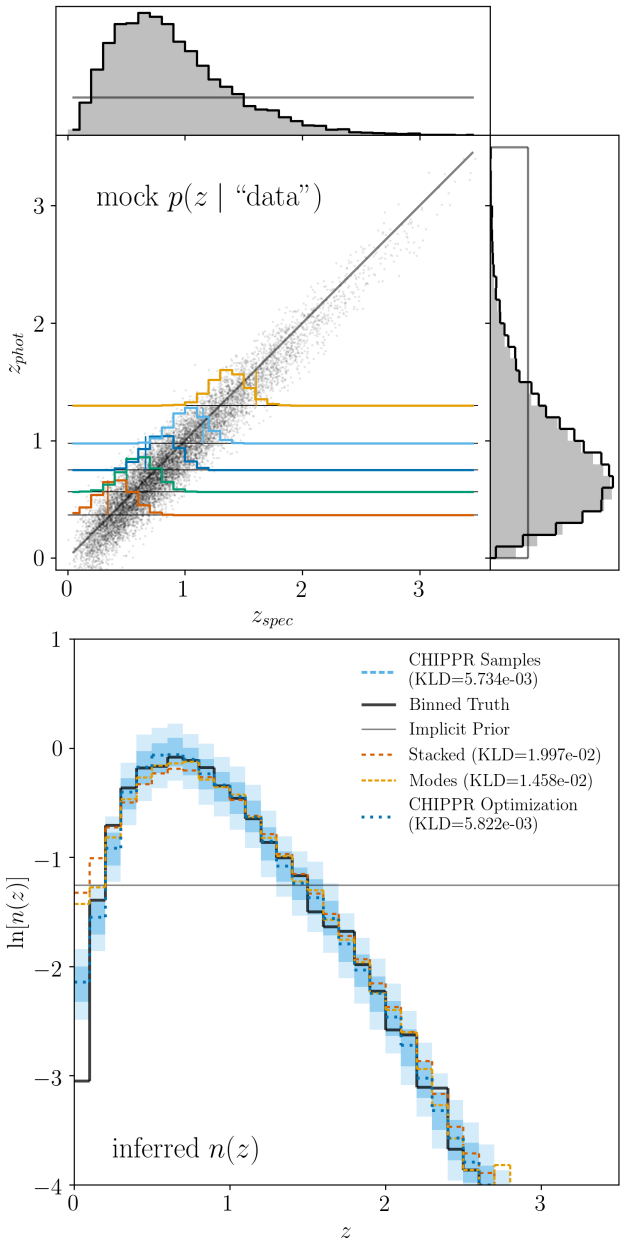

Figure 10. Top: Examples of photo-z posterior PDFs with ten times the bias of the LSST requirements, including samples from the probability space of true and observed redshift (black points), photo-z posterior PDFs (colored step functions), and the true redshifts of the example photo-z posterior PDFs (colored vertical lines), with marginal histograms (light gray) for each dimension with the true redshift distribution (black) and implicit prior (dark gray) in the insets. Bottom: The results of CHIPPR (samples in light blue, optimization in dark blue) and the alternative approaches (the stacked estimator in red, the histogram of modes in yellow) on photo-z posterior PDFs with ten times the bias of the LSST requirements, with the true redshift density (black curve) and implicit prior (gray curve). The impact of bias at even ten times the level of the LSST requirements is almost imperceptible on all estimators, though the CHIPPR marginalized maximum likelihood estimate minimizes the information loss regardless.

that of a template-fitting photo-z method (right panel). One can see that the photo-z posterior PDFs take different shapes from one another even though the marginal histograms of the points are identical. The machine learning-like implicit prior has been modified to have nonzero value at high-redshift because the implicit prior must be strictly positive definite for the CHIPPR model to be valid.

Figure 12 shows the performance of CHIPPR and the traditional methods on photo-z posterior PDFs generated with nontrivial implicit priors. In both cases, the CHIPPR marginalized maximum likelihood estimate effectively recovers the true redshift distribution, and the distribution of $n(z)$ parameter values reflects higher uncertainty where the implicit prior undergoes large changes in derivative. The alternatives, on the other hand, are biased by the implicit prior except where it is flat, in the case of high redshifts for the machine learning-like implicit prior, resulting in over $1,000$ times the information loss on $n(z)$ for the machine learning-like implicit prior and some $5-20$ times the information loss for the template fitting-like implicit prior, relative to the CHIPPR marginalized maximum likelihood estimate.

The main implication of the response of $n(z)$ estimates to a nontrivial implicit prior is that the implicit prior must be accounted for when using photo-z posterior PDF catalogs.

5.2. *Violations of the model*

In this test, the photo-z implicit posteriors are made to the LSST requirements but the implicit prior used for the inference is not the same as the implicit prior used for generating the data. Photo-z posterior PDF codes do not generally provide their implicit prior, with the exception of some template-fitting techniques for which it is a known input. If we naively used the photo-z posterior PDF catalog produced by a generic machine learning or template-fitting code and assumed a flat implicit prior, we would observe the contents of Figure 13.

The results of using a mischaracterized implicit prior are disastrous, causing every estimator, including CHIPPR, to be strongly biased. The stacked estimator and histogram of modes don't make use of the implicit prior so do no worse than when the implicit prior is accurately provided, but CHIPPR is sensitive to prior misspecification, which violates the model upon which it is based. It is thus crucial that photo-z posterior PDF methods always characterize and provide the implicit prior.

6. CONCLUSION

This study derives and demonstrates a mathematically consistent inference of a one-point statistic, the redshift density function $n(z)$, based on an arbitrary catalog of photo-z posterior PDFs. The fully Bayesian CHIPPR model, based in the fundamental laws of probability, begins with a probabilistic graphical model corresponding to equations for the full posterior distribution over the parameters for $n(z)$. The CHIPPR model is implemented in the publicly available `chippr` code. The method is implemented in the publicly available `chippr` code and validated on mock data.

Using a flexible, self-consistent forward model of the relationship between true and estimated redshifts, capa-

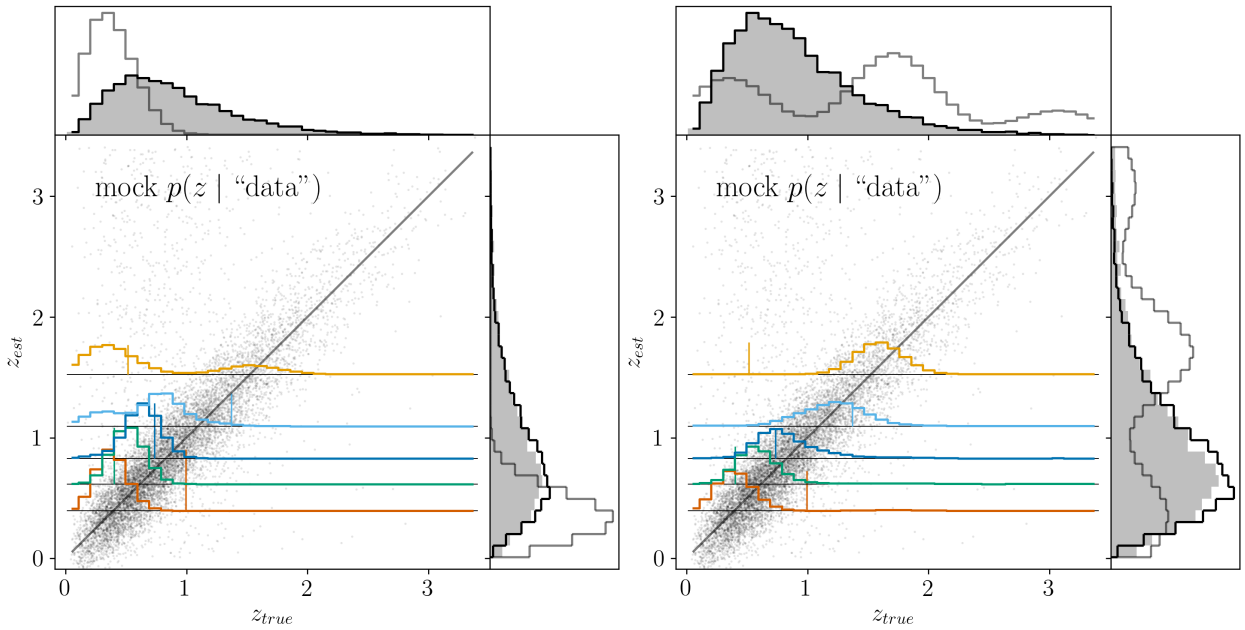

Figure 11. Examples of mock photo-z posterior PDFs generated with a machine learning-like implicit prior (left) and a template-fitting-like implicit prior (right), including samples from the probability space of true and observed redshift (black points), photo-z posterior PDFs (colored step functions), the true redshifts of the example photo-z posterior PDFs (colored vertical lines). A histogram (light gray) of points in each dimension is shown in the respective inset, with the true redshift distribution (black) and implicit prior (dark gray).

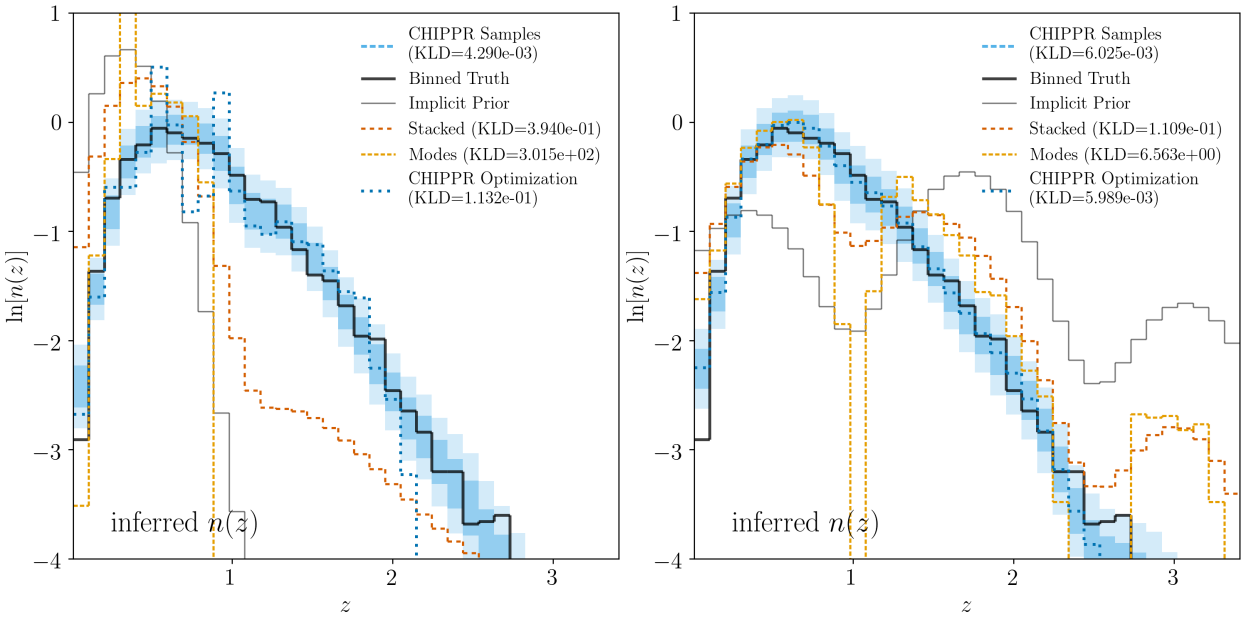

Figure 12. The results of CHIPPR (samples in light blue and optimization in dark blue) and the alternative approaches (the stacked estimator in red and the histogram of modes in yellow) on photo-z posterior PDFs with an implicit prior like that of machine learning photo-z posterior PDF approaches (left) and an implicit prior like that of template-fitting photo-z posterior PDF codes (right), with the true redshift density (black curve) and implicit prior (gray curve). CHIPPR is robust to a nontrivial implicit prior, but the alternatives are biased toward the implicit prior.

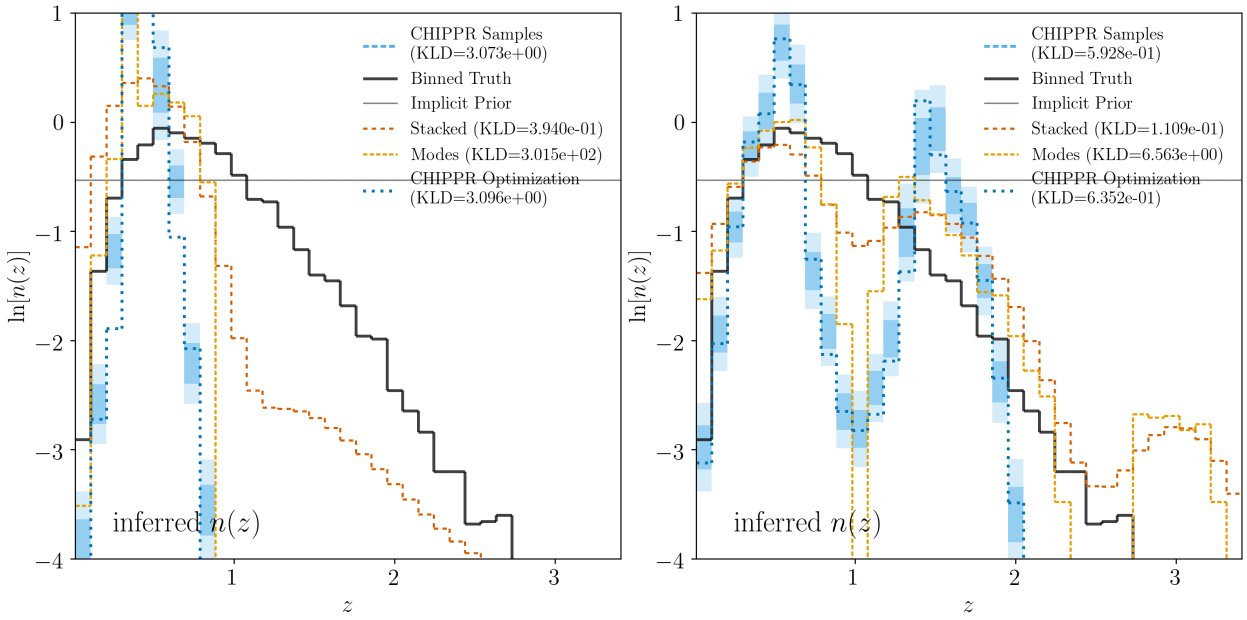

Figure 13. The results of CHIPPR (samples in light blue, optimization in dark blue) and the alternative approaches (the stacked estimator in red, the histogram of modes in yellow) when run with an incorrectly specified implicit prior (gray curve). The data upon which each panel's results are based are provided in Figure 11, where the left corresponds to the sort of implicit prior anticipated of machine learning approaches and the right corresponds to an implicit prior like that of a template-fitting code. Here, CHIPPR has been provided with a uniform implicit prior rather than those used to produce the mock photo-z posterior PDFs, and its performance is notably worse than when it is provided an accurate implicit prior, as in Figure 12. When the incorrect implicit prior is provided to `chippr`, even Bayesian inference cannot recover the true $n(z)$.

ble of encapsulating the complexity of observed redshift-photometry relations (e.g. Figure 1), we emulate the canonical photo-z error statistics, intrinsic scatter (Section 4.1), catastrophic outliers (Section 4.2), and canonical bias (Section 4.3) one at a time. Though these test cases may appear overly simplistic, they enable rigorous quantification of the relative performance of each $n(z)$ estimation techniques under the controlled conditions of each type of error in isolation, at levels equal to and beyond those of LSST.

Based on our tests, the following statements about the CHIPPR methodology may be made with confidence:

- CHIPPR outperforms traditional estimators of $n(z)$ under realistically complex conditions, even at pessimistic levels relative to future survey requirements on the traditional photo-z error statistics, as demonstrated both by eye and according to KLD values corresponding to 10% the information loss of alternative methods.

- Both the CHIPPR marginalized maximum likelihood estimate and the mean of `chippr` samples are good point estimators of $n(z)$, whereas the histogram of modes is very sensitive to outliers and the stacked estimator is always excessively broad.

- The error bars on the posterior distribution over $n(z)$ hyperparameters are interpretable and arise naturally under CHIPPR, unlike those that may be assumed for the conventional point estimators.

Not only is CHIPPR the only mathematically correct approach to the problem, it also recovers the true values of the hyperparameters defining $n(z)$ better than popular alternatives, as measured by the loss of information in $n(z)$. However, the mathematically valid approach to inference with probabilistic data products incurs nontrivial computational expense, motivating future work to optimize the implementation.

Additionally, this work highlights a crucial and almost entirely overlooked complication to the usage of photo-z posterior PDFs, namely the implicit prior, motivating the following recommendations:

- In the presence of a nontrivial implicit prior corresponding to the specifics of the architecture of the method by which photo-z posterior PDFs are obtained, established methods cannot recover $n(z)$; a principled hierarchical inference such as CHIPPR is the only way to recover $n(z)$ from photo-z posterior PDFs.

- Neither CHIPPR nor traditional alternatives can recover $n(z)$ in the presence of a misspecified implicit prior; the implicit prior used to produce the photo-z posterior PDF catalog must be known and provided to CHIPPR in order to recover the true $n(z)$.

Given the significance of the implicit prior (Schmidt et al. 2020), it is therefore imperative that those developing codes to obtain photo-z posterior PDFs provide a way to isolate the implicit prior and that those publishing photo-z posterior PDF catalogs provide the implicit prior to users. This mandate is easier said than done, both for template fitting and machine learning approaches.

While the implicit prior is often an explicit input to model-based routines, it may be defined in a space of redshift and SED templates. In this case, it may not be possible to apply CHIPPR without marginalizing over additional variables ψ for the SEDs. In other words, obtaining the implicit prior from a template fitting code may be challenging or even require consideration of higher-dimensional PDFs such as $\mathrm{p}(z, \mathrm{SED} \mid \psi^*)$.

The situation appears more dire for data-driven techniques, whose training sets may not straightforwardly translate into an implicit prior. For example, some training set galaxies may contribute to the photo-z posterior PDFs more than others, resulting in different effective weights when factoring into, say, a histogram of training set redshifts as the implicit prior. Additionally, the weights may be stochastic, depending on the random seed used to initialize non-deterministic methods, precluding reproducibility. It is thus unclear whether the implicit prior can be meaningfully obtained from such methods at all.

A thorough investigation of the degree to which the implicit prior can be meaningfully obtained is outside this paper but should be a priority for all consumers of photo-z posterior PDFs. As an alternative, however, we must point out that if likelihoods were available rather than posteriors, the trouble with the implicit prior would be avoided altogether. We thus encourage the community of those making photo-z posterior PDFs to consider developing such methods so that the resulting data products may be correctly used in scientific inference more generically.

By showing that CHIPPR is effective in recovering the true redshift distribution function and posterior distributions on its parameters from catalogs of photo-z posterior PDFs, this work supports the production of photo-z posterior PDFs by upcoming photometric surveys such as LSST to enable more accurate inference of the cosmological parameters. We discourage researchers from co-adding photo-z posterior PDFs or converting them into point estimates of redshift and instead recommend the use of Bayesian probability to guide the usage of photo-z posterior PDFs. We emphasize to those who produce photo-z posterior PDFs from data that it is essential to release the implicit prior used in generating this data product in order for any valid inference to be conducted by consumers of this information. Methodologies for obtaining photo-z posterior PDFs must therefore be designed such that there is a known implicit prior, i.e. one that is not implicit at all, so that likelihoods may be recovered.

The technique herein developed is applicable with minimal modification to other one-point statistics of redshift to which we will apply this method in the future, such as the redshift-dependent luminosity function and weak lensing mean distance ratio. Future work will also include the extension of this fully probabilistic approach to higher-order statistics of redshift such as the two-point correlation function.

AIM acknowledges support from the Max Planck Society and the Alexander von Humboldt Foundation in

the framework of the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. During the completion of this work, AIM was supported by National Science Foundation grant AST-1517237 and the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists, Office of Science Graduate Student Research (SCGSR) program, administered by the Oak Ridge Institute for Science and Education for the DOE under contract number DE-SC0014664. The authors thank Phil Marshall for advice on relevant examples, Elisabeth Krause for assistance with the `CosmoLike` code, Mohammadjavad Vakili for statistical insights, Geoffrey Ryan for programming advice, and Boris Leistedt for other helpful comments in the development of CHIPPR. This work was completed with generous nutritional support from the Center for Computational Astrophysics.

APPENDIX

DERIVATION

We perform the derivation of Equation 7 using log-probabilities. What we wish to estimate is then the full log-posterior probability distribution (hereafter the full log-posterior) of the hyperparameters $\boldsymbol{\phi}$ given the catalog of photometry $\{\vec{d}_j\}$.

By Bayes' Rule, the full log-posterior

$$\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})] = \ln[\mathrm{p}(\{\vec{d}_j\} \mid \boldsymbol{\phi})] + \ln[\mathrm{p}(\boldsymbol{\phi})] - \ln[\mathrm{p}(\{\vec{d}_j\})] \tag{A.1}$$

may be expressed in terms of the full log-likelihood probability distribution (hereafter the full log-likelihood) $\ln[\mathrm{p}(\{\vec{d}_j\} \mid \boldsymbol{\phi})]$ by way of a hyperprior log-probability distribution (hereafter the hyperprior) $\ln[\mathrm{p}(\boldsymbol{\phi})]$ over the hyperparameters and the log-evidence probability of the data $\ln[\mathrm{p}(\{\vec{d}_j\})]$. However, the evidence is rarely known, so we probe the full log-posterior modulo an unknown constant of proportionality.

The full log-likelihood may be expanded in terms of a marginalization over the redshifts as parameters, as in

$$\ln[\mathrm{p}(\{\vec{d}_j\} \mid \boldsymbol{\phi})] = \ln\left[\int \mathrm{p}(\{\vec{d}_j\} \mid \{z_j\})\mathrm{p}(\{z_j\} \mid \boldsymbol{\phi})\mathrm{d}\{z_j\}\right]. \tag{A.2}$$

We shall make two assumptions of independence in order to make the problem tractable; their limitations are be discussed below. First, we take $\ln[\mathrm{p}(\{\vec{d}_j\} \mid \{z_j\})]$ to be the sum of J individual log-likelihood distribution functions $\ln[\mathrm{p}(\vec{d}_j \mid z_j)]$, as in

$$\ln[\mathrm{p}(\{\vec{d}_j\} \mid \{z_j\})] = \sum_{j=1}^{J} \ln[\mathrm{p}(\vec{d}_j \mid z_j)], \tag{A.3}$$

a result of the definition of probabilistic independence encoded by the box in Figure 2. Second, we shall assume the true redshifts $\{z_j\}$ are J independent draws from the true $\mathrm{p}(z \mid \boldsymbol{\phi})$. Additionally, J itself is a Poisson random variable. The combination of these assumptions is given by

$$\ln[\mathrm{p}(\{z_j\} \mid \boldsymbol{\phi})] = -\int f(z;\boldsymbol{\phi})\mathrm{d}z + \sum_{j=1}^{J} \ln[\mathrm{p}(z_j \mid \boldsymbol{\phi})]. \tag{A.4}$$

The derivation differs when J is not known, say, when we want to learn about a distribution in nature rather than a distribution specific to data in hand, but for a photometric galaxy catalog where the desired quantity is $n(z)$ for the galaxies entering a larger cosmology calculation, it is a fixed quantity. A detailed discussion of this matter may be found in Foreman-Mackey et al. (2014). Applying Bayes' Rule, we may combine terms to obtain

$$\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})] \propto \ln[\mathrm{p}(\boldsymbol{\phi})] - \int f(z;\boldsymbol{\phi})\mathrm{d}z + \sum_{j=1}^{J} \ln\left[\int \mathrm{p}(\vec{d}_j \mid z)\mathrm{p}(z \mid \boldsymbol{\phi})\mathrm{d}z\right]. \tag{A.5}$$

Since we only have access to photo-z implicit posteriors, we must be able to write the full log-posterior in terms of log photo-z implicit posteriors rather than the log-likelihoods of Equation A.5. To do so, we will need an explicit statement of this implicit prior $\boldsymbol{\phi}^*$ for whatever method is chosen to produce the photo-z implicit posteriors.

To perform the necessary transformation from likelihoods to posteriors, we follow the reasoning of Foreman-Mackey et al. (2014). Let us consider the probability of the parameters conditioned on the data and an interim prior and rewrite the problematic likelihood of Equation A.5 as

$$\ln[\mathrm{p}(\vec{d}_j \mid z)] = \ln[\mathrm{p}(\vec{d}_j \mid z)] + \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)]. \tag{A.6}$$

Once the implicit prior $\boldsymbol{\phi}^*$ is explicitly introduced, we may expand the last term in Equation A.6 according to Bayes' Rule to get

$$\ln[\mathrm{p}(\vec{d}_j \mid z)] = \ln[\mathrm{p}(\vec{d}_j \mid z)] + \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] + \ln[\mathrm{p}(\vec{d}_j \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(\vec{d}_j \mid z, \boldsymbol{\phi}^*)]. \tag{A.7}$$

Because there is no direct dependence of the data upon the hyperparameters, we may again expand the term $\ln[\mathrm{p}(\vec{d}_j \mid z, \boldsymbol{\phi}^*)]$ to obtain

$$\ln[\mathrm{p}(\vec{d}_j \mid z)] = \ln[\mathrm{p}(\vec{d}_j \mid z)] + \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] + \ln[\mathrm{p}(\vec{d}_j \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(\vec{d}_j \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(\vec{d}_j \mid z)]. \tag{A.8}$$

Canceling the undesirable terms for the inaccessible likelihood $\ln[\mathrm{p}(\vec{d}_j \mid z)]$ and trivial $\ln[\mathrm{p}(\vec{d}_j \mid \boldsymbol{\phi}^*)]$ yields

$$\ln[\mathrm{p}(\vec{d}_j \mid z)] = \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)]. \tag{A.9}$$

We put this all together to get the full log-posterior probability distribution of

$$\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})] \propto \ln[\mathrm{p}(\boldsymbol{\phi})] + \ln\left[\int \exp\left[\sum_{j=1}^{J}\Big(\ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] + \ln[\mathrm{p}(z \mid \boldsymbol{\phi})] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)]\Big)\right] \mathrm{d}z\right], \tag{A.10}$$

which is equivalent to that of Hogg et al. (2010), though the context differs.

The argument of the integral in the log-posterior of Equation A.10 depends solely on knowable quantities (and those we must explicitly assume) and can be calculated for a given sample of log photo-z implicit posteriors $\{\ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)]\}$ and the implicit prior $\mathrm{p}(z \mid \boldsymbol{\phi}^*)$ with which they were obtained, noting the relation of

$$\mathrm{p}(z \mid \boldsymbol{\phi}) = \frac{f(z;\boldsymbol{\phi})}{\int f(z;\boldsymbol{\phi})\mathrm{d}z}. \tag{A.11}$$

Since we cannot know constant of proportionality, we sample the desired full log-posterior $\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})]$ using Monte Carlo-Markov chain (MCMC) methods.

REFERENCES

Abell, P. A., Allison, J., Anderson, S. F., et al. 2009

Abruzzo, M. W., & Haiman, Z. 2019, Mon Not R Astron Soc, 486, 2730

Asorey, J., Kind, M. C., Sevilla-Noarbe, I., Brunner, R. J., & Thaler, J. 2016, Monthly Notices of the Royal Astronomical Society, 459, 1293

Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, The Astrophysical Journal, 683, 12

Baum, W. A. 1962, Proceedings from IAU Symposium, 15, 390

Benítez, N. 2000, ApJ, 536, 571

Benjamin, J., Van Waerbeke, L., Heymans, C., et al. 2013, Mon Not R Astron Soc, 431, 1547

Bonnett, C. 2015, Monthly Notices of the Royal Astronomical Society, 449, 1043

Bonnett, C., Troxel, M. A., Hartley, W., et al. 2016, Phys. Rev. D, 94, 042005

Budavári, T. 2009, The Astrophysical Journal, 695, 747

Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, The Astrophysical Journal, 712, 511

Carrasco Kind, M., & Brunner, R. J. 2013, Monthly Notices of the Royal Astronomical Society, 432, 1483

—. 2014a, Monthly Notices of the Royal Astronomical Society, 442, 3380

—. 2014b, Mon Not R Astron Soc, 441, 3550

Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, The Astrophysical Journal, 775, 93

DiPompeo, M. A., Bovy, J., Myers, A. D., & Lang, D. 2015, Monthly Notices of the Royal Astronomical Society, 452, 3124

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, The Astrophysical Journal, 795, 64

Gelman, A., & Rubin, D. B. 1992, Statist. Sci., 7, 457

Gorecki, A., Abate, A., Ansari, R., et al. 2014, Astronomy & Astrophysics, 561, A128

Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, Astronomy & Astrophysics, 523, A31

Hildebrandt, H., Erben, T., Kuijken, K., et al. 2012, Mon Not R Astron Soc, 421, 2355

Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, Mon Not R Astron Soc, 465, 1454

Hogg, D. W. 2012, arXiv, 1205.4446

Hogg, D. W., Myers, A. D., & Bovy, J. 2010, ApJ, 725, 2166

Hoyle, B., Gruen, D., Bernstein, G. M., et al. 2018, Mon Not R Astron Soc, 478, 592

Jain, B., Spergel, D., Bean, R., et al. 2015, arXiv, 1501.07897

Kelly, P. L., von der Linden, A., Applegate, D. E., et al. 2014, Mon Not R Astron Soc, 439, 28

Koo, D. C. 1999, in Photom. Redshifts High Redshift Galaxies, ed. R. Weymann, L. Storrie-Lombardi, M. Sawicki, & R. Brunner (ASP Conference Series)

Leistedt, B., Mortlock, D. J., & Peiris, H. V. 2016, Mon Not R Astron Soc, 460, 4258

Leung, A. S., Acquaviva, V., Gawiser, E., et al. 2017, ApJ, 843, 130

Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, Monthly Notices of the Royal Astronomical Society, 390, 118

Malz, A. I., Marshall, P. J., DeRose, J., et al. 2018, AJ, 156, 35

Mandelbaum, R. 2017, arXiv, 1710.03235

Mandelbaum, R., Seljak, U., Hirata, C. M., et al. 2008, Mon Not R Astron Soc, 386, 781

Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53

Ménard, B., Scranton, R., Schmidt, S., et al. 2013, arXiv, 1303.4722

Norberg, P., Cole, S., Baugh, C. M., et al. 2002, Monthly Notices of the Royal Astronomical Society, 336, 907

Rohatgi, A. 2019, WebPlotDigitizer

Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502

Sanchez, A. G., Kazin, E. A., Beutler, F., et al. 2013, Mon. Not. R. Astron. Soc., 433, 1202

Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, arXiv, 2001.03621

Sheldon, E. S., Cunha, C. E., Mandelbaum, R., Brinkmann, J., & Weaver, B. A. 2012, ApJS, 201, 32

Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, Publ Astron Soc Jpn Nihon Tenmon Gakkai, 70, S9

van Breukelen, C., & Clewley, L. 2009, Monthly Notices of the Royal Astronomical Society, 395, 1845

Viironen, K., Marín-Franch, A., López-Sanjuan, C., et al. 2015, Astronomy & Astrophysics, 576, A25

Yang, S., & Pullen, A. R. 2018, Mon Not R Astron Soc, 481, 1441