How to Obtain the Redshift Distribution from Probabilistic Redshift Estimates

Alex I. Malz1 and David W. Hogg2,3,4,5

1 Ruhr-University Bochum, German Centre for Cosmological Lensing, Universitätsstraße 150, D-44801 Bochum, Germany; aimalz@ruw.edu
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

Received 2020 July 22; revised 2021 April 30; accepted 2021 May 26; published 2022 March 31

Abstract

A reliable estimate of the redshift distribution $n(z)$ is crucial for using weak gravitational lensing and large-scale structures 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 a decreasing signal-to-noise ratio, 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 that 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 cost. 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.

Unified Astronomy Thesaurus concepts: Observational cosmology (1146); Galaxy photometry (611); Astronomical methods (1043); Astrostatistics (1882); Hierarchical models (1925); Posterior distribution (1926); Prior distribution (1927); Bayesian statistics (1900); Redshift surveys (1378); Two-point correlation function (1951); Broad band photometry (184); Markov chain Monte Carlo (1889)

1. Introduction

Photometric redshift (photo-$z$) estimation has been a staple of studies of galaxy evolution, large-scale structures, 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 the photometric colors and redshifts may be constructed, forming a reliable 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 the surveyed galaxies. Many more photo-$z$ may be obtained in the time it would take to observe a smaller number of spec-$z$, and photo-$z$ 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-$z$ have thus enabled the era of precision cosmology, heralded by weak gravitational lensing tomography and baryon acoustic oscillation peak measurements.

However, photo-$z$ are susceptible to inaccuracy and imprecision in the form of inherent noisiness resulting from the decreasing signal-to-noise ratio with increasing redshift, 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-$z$ versus spec-$z$ illustrating the assumptions underlying photo-$z$ estimation in general, that spec-$z$ are a good approximation to true redshifts and photo-$z$ represent special nonlinear 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 populations 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 or 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 lying below the diagonal.

Once propagated through the calculations of the 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 in determining the influence of other sources of systematic error, the uncertainties associated with photo-$z$ 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 2018) states LSST’s requirements for photo-$z$ error tolerances for constraining cosmology with the main cosmological sample, reproduced in Table 1.

Much effort has been dedicated to improving photo-$z$, 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 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-$z$ for inference (Gorecki et al. 2014).

The probability density function (PDF) in redshift space for each galaxy, commonly written as $p(z)$, 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 $p(z)$ 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 Figure 1, the photo-$z$ PDFs are the horizontal cuts, probabilities of redshift conditioned on a specific value of data, i.e., posteriors $p(z|d)$, which constrain redshifts, whereas vertical cuts through this space are probabilities of data conditioned on a specific redshift, i.e., likelihoods $p(d|z)$, from which photometric data is actually drawn.

Table 1

| Number of galaxies | $\approx 10^7$ |
|--------------------|----------------|
| Rms error          | $<0.02 (1+z)$  |
| 3$\sigma$ outlier  | $<10\%$       |
| Canonical bias     | $<0.003 (1+z)$ |

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 (Shelden 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 (Figure 2).

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 $\{z_j^i\}$ of galaxies $j$ are known, then their redshift PDFs are well approximated by delta functions $\{\delta(z, z_j^i)\}$ centered at the true redshift, and the redshift distribution is effectively approximated by a histogram or other interpolation of the delta functions $\{\delta(z, z_j^i)\}$. When photo-$z$ posterior PDFs are available instead of true redshifts,

---

6 Note that spec-$z$ are not the same as known true redshifts; the PDFs of spec-$z$ would be narrow and almost always unimodal, but they would not be delta functions due to observational errors.
The Astrophysical Journal, 928:127 (18pp), 2022 April 1

Malz & Hogg

Figure 2. The DAG of the CHIPPR model, where circles indicate random variables and arrows indicate causal relationships. The redshift distribution \( n(z) \) parameterized by hyperparameters \( \phi \) exists independently 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 \( \phi \). The photometric data \( d_j \) for each galaxy is drawn from a function of its redshift \( z_j \) and observed, indicated by a shaded circle.

the simplest approach reduces them to point estimates \( \{z_j\} \) of redshift by using \( \delta(z, z_j) \) in place of \( \delta(z, z_j) \). Though it is most common for \( z_j \) to be the maximum or mode of the \( z \) 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 \( z \) posterior PDFs to point estimates discards valuable information about the uncertainty on redshift. \( z \) 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 (DiPompeo et al. 2015; Leung et al. 2017), but that procedure does not fully take advantage of all information contained in a probability distribution for parameter inference.

The most prevalent application of \( z \) posterior PDFs that preserves their information content is the estimation of the redshift distribution function \( N(z) \), or interchangeably, 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 \( g_L(\chi) \) in tomographic weak lensing by large-scale structures (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 structures, 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 \( z \)-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 \( z \) point estimates (Abruzzo & Haiman 2019), it has become more common to use \( z \) 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)

\[
\hat{n}(z) = \frac{1}{J} \sum_{j=0}^{J} p(z_j)
\]

for a sample of \( J \) galaxies \( j \), or equivalently, of the redshift distribution function \( \hat{N}(z) = J \hat{n}(z) \), by effectively averaging the \( z \) posterior PDFs. Despite its problems (Malz 2021), 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 \( z \) posterior PDF production, no implementation of inference using \( z \) posterior PDFs has yet been presented with a mathematically consistent methodology. This paper challenges the logically invalid yet pervasive analysis procedure of stacking \( z \) posterior PDFs by presenting and validating a hierarchical Bayesian technique for the use of \( z \) posterior PDFs in the inference of \( n(z) \), yielding a method applicable to arbitrary one-point statistics relevant to cosmology, large-scale structures, and galaxy evolution; future work will extend this methodology to higher-order statistics. We aim to develop a clear methodology guiding the use of \( z \) 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, 2019), the method presented here differs in that it makes use of any existing catalog of \( z \) posterior PDFs, rather than requiring a simultaneous derivation of the \( z \) 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 Cosmological Hierarchical Inference with Probabilistic Photometric Redshifts (CHIPPR) model for characterizing the full posterior probability landscape of \( N(z) \) using \( z \) posterior PDFs. In Section 3, we present the chippr implementation of the CHIPPR model and the experimental setup by which we validate it, including the forward modeling of mock \( z \) posterior PDFs. In Section 4, we present a number of informative test cases and compare the results of chippr with those of 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 $d_j$; thus the entire survey over some solid angle produces an ensemble of photometric magnitudes (or colors) and their associated observational errors $\{d_j\}$. 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 $d$. In Section 2.2, we present the directed acyclic graph (DAG) 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.

2.1. Forward Model

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

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

in terms of some parameter vector $\phi$. 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

$$z_j \sim p(z|\phi) \equiv f(z; \phi), \quad (4)$$

a probability density over redshift conditioned on the parameters $\phi$ 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 $p(z, d)$ of redshift $z$ and photometric data vectors $d$, which may be any combination of fluxes, magnitudes, colors, and their observational errors. Under this framework, $n(z)$ is equivalent to an integral

$$n(z) = \int p(z, d) \, dd \quad (5)$$

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 $p(z, d)$ need not be one-to-one. We assume a stronger version of statistical independence here, that draws $(z_j, d_j)$ are independent of draws $(z_{j'}, d_{j'})$ 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

$$d_j \sim p(d|z_j) \quad (6)$$

of photometry being conditioned on redshift, illustrated in Figure 1.

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 $\phi$.
2. Galaxy redshifts $\{z_j\}$ are independent draws from $p(z|\phi)$.
3. Galaxy photometry $d_j$ is drawn from the likelihoods $p(d_j|z_j)$.

2.2. Probabilistic Model

A forward model such as that of Section 2.1 corresponds to a probabilistic graphical model, represented by a 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 shorthand for mathematical relationships between variables. The photometric data $d_j$ of a galaxy is drawn from some function of its redshift $z_j$, independently 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 do not 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 $\phi$ that we seek are causally separated from the data by the latent variable of redshift, we call them hyperparameters.

The problem facing cosmologists is how to determine the true value of $\phi$ from observing the photometry $\{d_j\}$ 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 $p(\phi|\{d_j\})$ over all possible values of $\phi$ conditioned on all observed data $\{d_j\}$ available in a generic catalog. In order to use the DAG of Figure 2 to derive an expression for $p(\phi|\{d_j\})$ in terms of photo-z posterior PDFs, we must introduce two more concepts, confusingly named the implicit prior and the (hyper) prior PDF, elaborated upon below.

When we constrain the redshift of a galaxy using its observed photometric data $d_j$, we are effectively estimating a posterior $p(z|d_j)$, 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 $d = d_j$ of the
empirical frequency distribution \( p(z, d|\phi) \), with the joint probability density corresponding to the true redshift distribution parameterized by \( \phi \), which exists in nature but need not be known.

As the hyperparameters \( \phi \) of the true redshift distribution are in general unknown, the investigator seeking to estimate a posterior \( p(z|d_j) \) must have a model \( \phi \) 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 \( p(d, z) \), we would obtain a one-dimensional PDF \( p(z|\phi^*) \) over redshift, which can by definition be parameterized by the same functional form as \( n(z) \), for some \( \phi \) specific to the estimation procedure that may or may not bear any relation to the hyperparameters \( \phi \) of the true \( n(z) \). Rather, \( \phi \) 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 \( p(z|\phi) \) the implicit prior, as it is rarely explicitly known or chosen by the researcher. Because the implicit prior is unavoidable and almost inherently not uninformative, the photo-z posterior PDFs reported by any method must be implicit posteriors \( p(z|d, \phi^*) \) weighted by the implicit prior.

The prior probability density \( p(\phi) \) 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 \( \phi \). This prior need not be excessively prescriptive; for example, it may be chosen to enforce smoothness at physically motivated scales in redshift without imposing any particular region as over- or underdense.

With inputs of the photo-z implicit posterior catalog \( \{ p(z|d, \phi^*) \} \), the implicit prior \( p(z|\phi^*) \), and the prior PDF \( p(\phi) \), we thus aim to obtain the posterior probability \( p(\phi|d_j) \) of the redshift density function given all the photometric data. By performing the derivation in the appendix, we arrive at the desired expression:

\[
p(\phi|d_j) \propto p(\phi) \int \prod_{j=1}^t \frac{p(z|d_j, \phi^*)p(z|\phi)}{p(z|\phi^*)} dz.
\]

Equation (7) constitutes the heart of CHIPPR. A full mathematical derivation may be found in the appendix. Malz (2019) contains a less technical presentation geared toward the general public using the universal language of dance. 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 (A3) and Equation (A4) 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 \( \phi \) used to produce the photo-z implicit posteriors.

3. We must assume a hyperprior distribution \( p(\phi) \) 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 reliable; there is no consensus yet on the best technique to obtain photo-z posterior PDFs, and the implicit prior may not be appropriate or even known to us as users 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 \( p(z, d|\phi) \) has mutual coverage with the parameter values that real galaxies can take. In other words, any probability distribution over the \( z, d \) space must be nonzero where real galaxies can exist. Additionally, the hyperprior \( p(\phi) \) must be nonzero at the hyperparameters \( \phi \) 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 \( p(z, d|\phi^*) \) and \( p(\phi) \) must be sufficiently general as to not rule out plausible areas of parameter space.

3. Methods and 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 a prototype 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 public 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. For the sake of reproducibility, this paper presents the prototype-
level code used for the experiments discussed herein. However, a modern, flexible implementation appropriate for at-scale application to future data sets is currently under development.\(^9\)

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

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

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

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

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

The emcee (Foreman-Mackey et al. 2013) implementation of Monte Carlo Markov Chain (MCMC) ensemble sampling is used to sample the full log-posterior of Equation (A10). As a consequence of the central limit theorem being applied to a Markov chain, the sample distribution satisfies a statistical guarantee of normality (Brooks et al. 2011), meaning the output samples \(\{\phi_j\}\) of chippr naturally define a notion of standard deviation interpretable as conventional error bars.\(^10\) 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 \(\phi_i\) is drawn from the log-prior distribution and evaluated for acceptance to or rejection from the full log-posterior distribution. The resulting output includes \(2W\) accepted samples \(\phi_i\) for a prespecified chain thinning factor’s and their full posterior probabilities \(p(\phi_i|\{d_j\})\), as well as the autocorrelation times and acceptance fractions calculated for each element of \(\phi_i\) divided into separate files before and after the completion of the burn-in phase, as defined by the Gelman–Rubin statistic \(R\) (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-\(z\) implicit posteriors: the distribution \(n(z)\) of the redshifts at maximum posterior probability

\[
f^{\text{MMAP}}(z; \hat{\phi}) = \sum_{j=1}^{J} \delta(z, \text{ mode}[p(z|d_j, \phi^*)])
\]

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

\[
f^{\text{stack}}(z; \hat{\phi}) = \sum_{j=1}^{J} p(z|d_j, \phi^*)
\]

in terms of the photo-\(z\) implicit posteriors we have. These two approaches have been compared to each other 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 \(p(z|d_j, \phi^*)\) into delta functions with all probability at a single estimated redshift. Some variants of point estimation choose this single redshift to be that of the maximum a posteriori probability \(\text{mode}[p(z|d_j, \phi^*)]\) or the expected value of the redshift \(\langle z \rangle = \int \delta(z|d_j, \phi^*) dz\). Tanaka et al. (2018) direct 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-\(z\) implicit posteriors leads to the marginalized maximum a posteriori estimator and the marginalized expected value estimator, though only the former is included in this study since the latter has fallen out of favor in recent years.\(^11\)

It is crucial to discuss 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 (the marginalized maximum posterior estimate), the value of \(\phi\) maximizing the log-posterior given by Equation (A10) using any optimization code. The marginalized maximum posterior estimate can be obtained in substantially less time than enough samples to characterize the full log-posterior distribution of \(n(z)\). However, the marginalized maximum posterior estimate yields only a point estimate of \(n(z)\) rather than characterizing the full log-posterior on \(\phi\), 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 the MCMC samples of these 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),

\(^9\) A working prototype of a Python 3-based implementation built on the qp backend presented in Malz et al. (2018) may be found at https://github.com/LSSTDESC/qp, which is anticipated to evolve in sophistication over time.

\(^10\) Additionally, the proofs are tractable by hand in the special cases of simplified photo-\(z\) posterior PDFs introduced in Malz (2021).

\(^11\) 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.
given by

$$\text{KLD}_{\phi, \phi^*} = \int p(z|\phi) \ln \left( \frac{p(z|\phi)}{p(z|\phi^*)} \right) dz,$$

(12)

which measures the distance from the parameter values $\phi$ to the true parameter values $\phi^*$. The KLD is a measure of information loss, in units of nats, due to using $\phi$ to approximate the true value $\phi^*$ 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-$z$ implicit posterior maxima (modes) on mock data in the form of catalogs of emulated photo-$z$ implicit posteriors 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 $\phi^*$ of the hyperparameters governing $n(z)$, a photo-$z$ model $p(z, d)$ defining the space of redshift and photometry, and the prior values $\phi$ of the hyperparameters of $n(z)$. In the first phase, we sample $J = 10^4$ redshifts $z_j^* \sim p(z|\phi^*)$. In the second phase, we evaluate the photo-$z$ model at those redshifts, yielding a set of $J$ likelihoods $p(d|z_j^*)$, from which we then sample data $d^j$ for each galaxy. In the third phase, we evaluate the photo-$z$ model at data to obtain $J$ posteriors $p(z|d^j)$. In the fourth phase, we convolve the posteriors with the chosen prior $p(z|\phi^*)$, yielding implicit posteriors $p(z|d^j, \phi^*)$.

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

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

(13)

with $c_z = 0.3$, because it has been used in forecasting studies for LSST (Chang et al. 2013; LSST Dark Energy Science Collaboration et al. 2018).

The mock data emulates the three sources of error of highest concern to the photo-$z$ community, which 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, recalling Figure 1.

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

$$\Sigma_{k,k'} = q \exp \left[ -\frac{e\epsilon}{2} (\bar{\xi}_k - \bar{\xi}_{k'})^2 \right] + \delta(k, k')$$

(14)

based on the one used in Gaussian processes, where $k$ and $k'$ are indices ranging from 1 to $K$ and $q = 1.0$, $e\epsilon = 100.0$, 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 $\phi$. 

**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-$z$ implicit posterior catalogs.

**Figure 4.** The joint probability space of true and estimated redshifts for the three photo-$z$ systematics of concern at the level of the LSST requirements: intrinsic scatter, uniformly distributed catastrophic outliers, and bias. The main panel shows the samples (black points) in the space of mock data and redshift, akin to the standard scatterplots of true and estimated redshifts, the $z_{\text{spec}} = z_{\text{phot}}$ diagonal (gray line), and the posterior probabilities evaluated at the given estimated redshift (colored step functions). The insets show the marginal histograms (light gray) in each dimension, which 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).
We adapt the full log-posterior of Equation (A10) 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. The computational cost is dominated by the sampler’s burn-in phase and driven predominantly by the number of parameters $K$; for the tests presented in this study, with $J = 10^4$ galaxies, $K = 35$ parameters, $W = 100$ walkers, a Gelman–Rubin convergence threshold of $R = 1.5$, a thinning factor of $s = 1$, and $N = 10^3$ accepted post-burn samples, the single-threaded computation time on a 2017-era laptop is on the order of a few hours.

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 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 posterior 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 three to four times as many nats of information about $n(z)$ for the LSST requirements relative to the marginalized maximum posterior estimate of CHIPPR.

4.2. Catastrophic Outliers

As was covered in Section 1, catastrophic outliers tend to be distributed nonuniformly across the space of observed and true redshifts. However, the LSST requirements do not specify the details for the 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 from a Gaussian about the true redshift. Though this results in slightly less than a 10% catastrophic outlier rate, it can be done independently of the definition of the standard deviation and so was implemented for demonstrative purposes. Figure 7 shows 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 is not increased with redshift as indicated in Table 1 in order to isolate the effect of outliers, and is instead held at a constant $\sigma_{\text{r}} = 0.02$.

Figure 7 shows that at the level of the LSST requirements, the alternative estimators are overly broad, whereas CHIPPR’s marginalized maximum posterior 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 10 times the information loss of CHIPPR’s marginalized maximum posterior 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 a 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 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, which are truly multimodal. The testing conditions here, illustrated in the right panel of Figure 8, give 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 underestimating it (template fitting–like outliers).
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 10 times the information loss as compared to the CHIPPR marginalized maximum posterior estimate, 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 of the CHIPPR marginalized maximum posterior estimate. The CHIPPR marginalized maximum posterior 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 \( \Delta_z = \langle p(z|\hat{\phi}) \rangle - \langle p(z|\hat{\phi'}) \rangle \) representing a difference between the first moment of the estimated redshift density function and that of the true redshift density function. To distinguish other aforementioned manifestations of bias from this common form of bias, we refer to $\Delta_z$ as canonical bias.

In the context of photo-$z$ posterior PDFs, canonical bias represents an instance of model misspecification. Consider that if canonical bias were included in the framework of Figure 1, it could be trivially modeled out as a simple linear transformation of $z_{\text{phot}} \rightarrow z_{\text{phot}} - \Delta_z(1 + z_{\text{phot}})$ of the $(z_{\text{spec}}, z_{\text{phot}})$ space. In any case, for completeness, a test at 10 times the canonical bias of the LSST requirements, with no redshift-dependent intrinsic scatter or catastrophic outliers, is provided in Figure 10.

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

5. Discussion

The experiments of Section 4 quantify the influence on each estimator of $n(z)$ of each of the canonical types of photo-$z$ error one at a time in isolation. Now, we forecast the influence of these $n(z)$ estimators on cosmological parameter constraints in a tomographic analysis and stress-test CHIPPR by exploring the impact of the implicit prior, which has thus far not received 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-$z$ implicit posterior catalog. Section 5.3 presents the impact of $n(z)$ estimation methods on forecasted cosmological parameter constraints. These results provide compelling motivation for the photo-$z$ community to prioritize the study of implicit priors of existing and future photo-$z$ posterior PDF techniques.
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-\( z \) implicit posterior 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 number of templates, leading to a highly nonuniform but physically motivated interim prior. Machine-learning approaches tend to be trained on one or more previously observed data sets that include only galaxies for which spectroscopy is accessible, typically biasing the implicit prior toward atypically bright and/or low-redshift populations. 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-\( z \) implicit posteriors 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 that of a template-fitting approach (right panel). CHIPPR is robust to a nontrivial implicit prior, but the alternatives are biased toward the implicit prior.
One can see that the photo-z implicit posteriors 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 a 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 implicit posteriors generated with nontrivial implicit priors. In both cases, the CHIPPR marginalized maximum posterior 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 derivatives. The alternatives, on the other hand, are biased by the implicit prior except where it is flat (which is the case for high redshifts for the machine learning–like implicit prior), resulting in over 1000 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 posterior 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 implicit posterior catalogs.

5.2. Violations of the Model

In this test, the photo-z implicit posteriors are made according 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 implicit posterior 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 do not make use of the implicit prior and so do no worse than

---

**Figure 13.** 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) when they are run with an incorrectly specified implicit prior (gray curve). The data on which each panel’s results are based are provided in Figure 11, where the left panel 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 implicit posteriors, 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 \( n(z) \).

**Figure 14.** Same as Figure 4, but for four LSST-DESC-like tomographic bins (colors). Main panel: \( \{z_{\text{spec}}, z_{\text{phot}}\} \) pairs drawn from a model with LSST-like systematic errors. Top panel: the LSST-like tomographic \( n(z) \) model (thin, opaque line), after a reduction in the number of parameters of \( n(z) \) (dashed line), and histogram of \( z_{\text{spec}} \) values drawn from that model (thick, semitransparent line). Side panel: histograms of \( z_{\text{spec}} \) (thin, opaque line) and \( z_{\text{phot}} \) (semitransparent bars).
when the implicit prior is accurately provided, but CHIPPR is sensitive to prior misspecification, which violates the model on which it is based. It is thus crucial that photo-z implicit posterior methods always characterize and provide the implicit prior.

5.3. Propagation to Cosmology

It is of interest to explore the impact of incorrectly estimated \( n(z) \) on derived constraints on the cosmological parameters to answer the question of how wrong we will be in our understanding of the universe if we incorrectly constrain \( n(z) \). To test the impact of these uncertainties, we emulate mock data with all three effects with which LSST is concerned at the levels of Table 1 and propagate the results of CHIPPR and the other estimators of \( n(z) \) through a Fisher matrix forecast using CosmoLike (Krause & Eifler 2017), a publicly available cosmological forecasting code.

We consider as ground truth a set of known \( n(z) \) corresponding to each of four hypothetical samples of galaxies and the corresponding cosmological parameter covariance matrix. The \( n(z) \) of each galaxy subsample emulates that anticipated of galaxies binned by a redshift point estimate, as is common in tomographic redshift analyses, though our experimental procedure is agnostic to how the samples are identified. The cosmological parameter covariance matrices are those used for LSST-DESC forecasting with the ground truth \( n(z) \) in the same four bins. The true \( n(z) \) in each predefined bin is provided in the form of an evaluation of the function on a fine grid of 350 redshifts \( 0.0101 < z < 3.5001 \).

First, we bin down the true redshift distribution to a piecewise constant parameterization with a manageable 35 hyperparameters for chippr’s sampling capabilities. Next, we draw \( 10^4 \) true redshifts from the binned true \( n(z) \) for each tomographic bin. The original, binned, and drawn \( n(z) \) are shown in the top panel of Figure 14. We emulate photo-z posterior PDFs for the \( 10^4 \) true redshifts drawn from the true \( n(z) \) in each bin using the procedure of Figure 3 with all three effects of Table 1 at their given levels. The drawn \( \{z_{\text{true}}, z_{\text{phod}}\} \) pairs are shown in the main panel of Figure 14, and the side panel shows their marginal histograms.

We then make a point estimate of \( n(z) \) using chippr’s marginalized maximum posterior estimate optimization option as well as the alternative methods on the photo-z posterior PDF catalog for each tomographic bin, shown in Figure 15, because CosmoLike produces cosmology constraints from a single \( n(z) \) result, rather than samples from the full posterior probability density of possible \( n(z) \). The excessive breadth of the alternative estimators, especially in regions of low probability relative to the CHIPPR estimate, can be seen quite plainly.

We then use the different estimators of \( n(z) \) in a cosmological forecasting procedure with CosmoLike, constraining \( \Omega_m, \Omega_b, w_0, w_a, n_s, S_8, \) and \( H_0 \), shown in Figure 16. Though there are also slight differences in the angle of the error contours under the alternative estimators relative to CHIPPR, which are almost indistinguishable from those derived by using the true redshift distribution in each bin. The stacked estimator is significantly worse than the CHIPPR marginalized maximum posterior estimate for all parameters except \( \Omega_b \) and \( H_0 \). Stacking, however, outperforms the histogram of modes for all parameters except \( \Omega_m \) and \( S_8 \) for which their constraints are quite similar. Though the Fisher matrix approach cannot be used to evaluate the bias of point estimates of the cosmological parameters relative to their true values, we calculate the seven-dimensional KLD for the three \( n(z) \) estimators relative to the constraints derived from the true \( n(z) \), finding that CHIPPR preserves information 200–800 times better than the alternatives, and even the histogram of modes outperforms stacking by a factor of four.

---

**Figure 15.** The chippr-derived \( n(z) \) and other estimators of \( n(z) \) in each tomographic bin (panels), with the true \( n(z) \) (solid black), stacked estimator (thin dashed red), histogram of modes (dotted yellow), and CHIPPR marginalized maximum posterior estimate (thick dashed blue). The result of stacking is far too broad for LSST-like photo-z posterior PDFs, even more so than that of the simplistic histogram of modes, while the CHIPPR marginalized maximum posterior estimate more closely follows the amplitude of the true \( n(z) \).
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 on 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 prototype code and validated on mock data.

Using a flexible, self-consistent forward model of the relationship between true and estimated redshifts, capable 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 performance of each of two established \( n(z) \) estimation techniques relative to CHIPPR 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 can be made with confidence:

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

2. Both the CHIPPR marginalized maximum posterior 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.

3. The error bars of the posterior distribution over the \( n(z) \) hyperparameters are interpretable and arise naturally.

Figure 16. The result of propagating the estimators of \( n(z) \) by stacking (red horizontal hatching), the histogram of modes (yellow vertical hatching), and CHIPPR (blue shaded) and the true \( n(z) \) (black outline) of Figure 15 to a subset of cosmological parameters. For all parameters considered, CHIPPR yields contours no broader than those corresponding to the true \( n(z) \), whereas stacking and the histogram of modes yield broader contours.
under CHIPPR, unlike those that are sometimes assumed for conventional point estimators.

4. When the \( n(z) \) estimates for LSST-like data are propagated through a cosmological forecast, CHIPPR’s information loss in the \( (\Omega_m, \Omega_b, w_{\gamma}, w_0, \sigma_8, n_s, H_0) \) parameter space is two to three times lower than that corresponding to traditional \( n(z) \) estimators.

Not only is CHIPPR the only mathematically correct approach to the problem of obtaining \( n(z) \) from an external catalog of photo-\( z \) posterior PDFs, it also recovers the true values of both the hyperparameters defining \( n(z) \) and the cosmological parameters better than popular alternatives, as measured by the loss of information in \( n(z) \) KLD. However, the mathematically valid approach to inference with probabilistic data products incurs nontrivial computational costs, motivating future ongoing work to optimize the implementation.

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

1. 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.

2. 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 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 for 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 \( \psi \) for the SEDs. In other words, obtaining the implicit prior from a template-fitting code may be challenging or may even require consideration of higher-dimensional PDFs such as \( p(z, \text{SED}[\psi^*]) \).

The situation is more dire for data-driven techniques, whose training sets may not straightforwardly translate into a recoverable 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 in, for example, a histogram of training set redshifts as the implicit prior. Additionally, the weights may be stochastic, depending on the random seed used to initialize nondeterministic 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 the scope of this paper but should be a priority for all users of photo-\( z \) posterior PDFs. Alternatively, the trouble with the implicit prior would be avoided altogether if likelihoods were produced rather than posteriors. We thus encourage the community of those making photo-\( z \) posterior PDFs to consider developing methods yielding likelihoods rather than posteriors 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 coadding 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 determine and release the implicit prior used in generating this data product in order for any valid inference to be conducted by users 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.

While the chippr prototype code is publicly available, follow-up work is currently under way to build an at-scale implementation of the CHIPPR algorithm that will be applied to current and future data sets. The CHIPPR 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.

A.I.M. 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, A.I.M. was supported by National Science Foundation grant AST-1517237 and the U.S. Department of Energy (DOE), Office of Science, Office of Workforce Development for Teachers and Scientists, Office of Science Graduate Student Research program, administered by the Oak Ridge Institute for Science and Education for the DOE under contract No. 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

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 \( \phi \) given the catalog of photometry \( (\mathbf{d}_i) \).
By Bayes’ rule, the full log-posterior
\[
\ln[p(\phi|d_j)] = \ln[p(\phi|d_j)] + \ln[p(d_j)] - \ln[p(\phi)]
\]  
(A1)

may be expressed in terms of the full log-likelihood probability distribution (hereafter the full log-likelihood) \(\ln[p(d_j)|\phi]\) by way of a hyperprior log-probability distribution (hereafter the hyperprior) \(\ln[p(\phi)]\) over the hyperparameters and the log-evidence probability of the data \(\ln[p(d_j)]\). 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
\[
\ln[p(d_j)|\phi] = \ln\left[\int p(d_j|z_j)p(z_j|\phi)dz_j\right].
\]  
(A2)

We shall make two assumptions of independence in order to make the problem tractable; their limitations are discussed below. First, we take \(\ln[p(d_j)|z_j]\) to be the sum of \(J\) individual log-likelihood distribution functions \(\ln[p(d_j|z_j)]\), as in
\[
\ln[p(d_j|z_j)] = \sum_{j=1}^{J} \ln[p(d_j|z_j)].
\]  
(A3)

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 \(p(z|\phi)\).

Additionally, \(J\) itself is a Poisson random variable. The combination of these assumptions is given by
\[
\ln[p(z_j|\phi)] = -\int f(z; \phi)dz + \sum_{j=1}^{J} \ln[p(z_j|\phi)].
\]  
(A4)

The derivation differs when \(J\) is not known—for example, 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
\[
\ln[p(\phi|d_j)] \propto \ln[p(\phi)] - \int f(z; \phi)dz + \sum_{j=1}^{J} \ln[p(d_j|z)p(z|\phi)dz].
\]  
(A5)

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 (A5). To do so, we will need an explicit statement of this implicit prior \(\phi^*\) 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 (A5) as
\[
\ln[p(d_j|z)] = \ln[p(d_j|z)] + \ln[p(z|d_j, \phi^*)] - \ln[p(d_j|\phi^*)].
\]  
(A6)

Once the implicit prior \(\phi^*\) is explicitly introduced, we may expand the last term in Equation (A6) according to Bayes’ rule to get
\[
\ln[p(d_j|z)] = \ln[p(d_j|z)] + \ln[p(z|d_j, \phi^*)] - \ln[p(d_j|\phi^*)] - \ln[p(d_j|z, \phi^*)].
\]  
(A7)

Because there is no direct dependence of the data on the hyperparameters, we may again expand the term \(\ln[p(d_j|z, \phi^*)]\) to obtain
\[
\ln[p(d_j|z)] = \ln[p(d_j|z)] + \ln[p(z|d_j, \phi^*)] + \ln[p(d_j|\phi^*)] - \ln[p(d_j|z)] - \ln[p(z|\phi^*)] - \ln[p(d_j|\phi^*)] - \ln[p(d_j|z)].
\]  
(A8)

Canceling the undesirable terms for the inaccessible likelihood \(\ln[p(d_j|z)]\) and trivial \(\ln[p(d_j|\phi^*)]\) yields
\[
\ln[p(d_j|z)] = \ln[p(z|d_j, \phi^*)] - \ln[p(z|\phi^*)].
\]  
(A9)

We put this all together to get the full log-posterior probability distribution of
\[
\ln[p(\phi|d_j)] \propto \ln[p(\phi)] + \ln[p(z|d_j, \phi^*)] + \ln[p(z|\phi^*)] \]  
(A10)

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 (A10) 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 \(\ln[p(z|d_j, \phi^*)]\) and the implicit prior \(p(z|\phi^*)\) with which they were obtained, noting the relation
\[
p(z|\phi) = \frac{f(z; \phi)}{\int f(z; \phi)dz}.
\]  
(A11)

Since we cannot know the constant of proportionality, we sample the desired full log-posterior \(\ln[p(\phi|d_j)]\) using MCMC methods.

**ORCID iDs**

Alex I. Malz @ https://orcid.org/0000-0002-8676-1622
David W. Hogg @ https://orcid.org/0000-0003-2866-9403

**References**

Abell, P. A., Allison, J., Anderson, S. F., et al. 2009, arXiv:0912.0201
Abruzzo, M. W., & Haiman, Z. 2019, MNRAS, 486, 2730
Asorey, J., Kind, M. C., Sevilla-Noarbe, I., Brunner, R. J., & Thaler, J. 2016, MNRAS, 459, 1293
Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, ApJ, 683, 12
Baum, W. A. 1962, in IAU Symp. 15, Problems of Extra-Galactic Research (London: MacMillan Press), 390
