Draft version May 24, 2019
Preprint typeset using LaTeX style emulateapj v. 01/23/15

HOW TO OBTAIN THE REDSHIFT DISTRIBUTION FROM PROBABILISTIC REDSHIFT ESTIMATES

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

Draft version May 24, 2019

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 weak-lensing surveys are expected to be unavailable, making photometric redshifts (photo-zs) the next-best alternative. The nontrivial systematics affecting photo-z estimation have motivated the weak-lensing community to favor photo-z probability density functions (PDFs) as a more comprehensive alternative to photo-z point estimates. However, analytic methods for utilizing these new data products in cosmological inference are still evolving. The ubiquitous methodology known as stacking produces a systematically biased estimator of $n(z)$ that worsens with decreasing signal-to-noise, the very regime where photo-z PDFs are most necessary. We introduce a mathematically rigorous probabilistic graphical model (PGM) of $n(z)$ for hierarchical inference, which is provably the only self-consistent way to combine photo-z PDFs to produce an estimator of $n(z)$. Our Cosmological Hierarchical Inference with Probabilistic Photometric Redshifts (CHIPPR) model yields a more accurate characterization of $n(z)$ by correctly propagating the redshift uncertainty information beyond the best-fit estimator produced by traditional procedures. The CHIPPR approach is applicable to any one-point statistic of any random variable. We conclude by propagating these effects to constraints in the space of cosmological parameters, . *AIM*: be clear that rigor comes at a cost, chippr makes explicit the requirements that the pzpdfs be posteriors under a known prior (which would be required by any justifiable method), be honest that we're not ready for prime time

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 decades 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 galaxy redshifts 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 via a number of effects, particularly their inherent noisiness due to the coarseness of photometric filters, catastrophic errors in which galaxies of one type at one redshift are mistaken for galaxies of another type at a different redshift, and systematics introduced by observational techniques, data reduction processes, and training or template set limitations. Figure 1 illustrates the relationship between photometry and redshift generically by showing "data" in one dimension for visualization purposes, suggesting that a special nonlinear projection of the photometry could more-or-less yield a one-to-one relationship with true redshifts. This plot looks very much like the traditional $z_{\rm spec}$ versus $z_{\rm phot}$ plots of photo-z point estimates and was indeed adapted from one such plot in Jain et al. (2015), because photo-z point estimates effectively are a special nonlinear projection of the data.

There are several varieties of generally non-Gaussian deviation from a perfect correspondence between redshift and data in Figure 1, represented by a $y = x$ diagonal line. There is scatter about the diagonal due to the coarseness of the photometric filters, with larger scatter perpendicular to the diagonal due to the specific wavelengths where highly identifiable spectral features pass between the filters, as well as higher scatter at high redshifts due to larger photometric errors on fainter galaxies. There are populations of outliers, far from the diagonal, comprised of galaxies for which the redshift estimate is

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

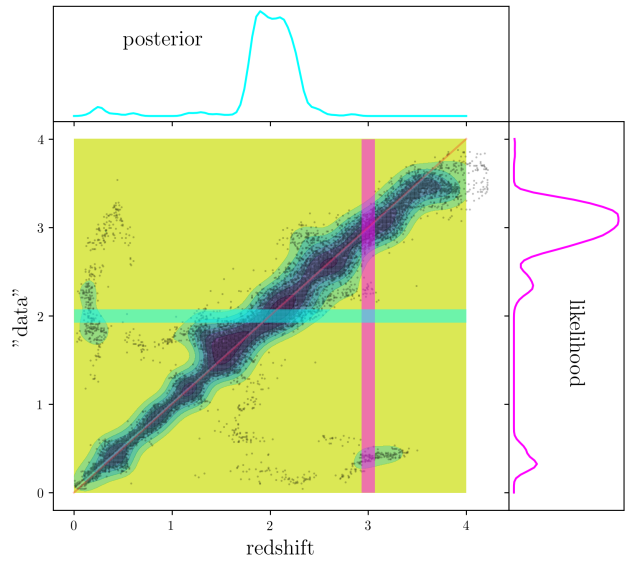

Figure 1. A generic probability space (darker in areas of higher probability density) of redshift (x-axis) and data (y-axis), where the data is projected into a single dimension, with vertical cuts and marginals (cyan) indicating the construction of likelihoods and horizontal cuts and marginals (magenta) indicating the construction of posteriors. The data (black points) used to generate the contours were extracted from Jain et al. (2015) using WebPlotDigitizer (Rohatgi 2019), with the ideal redshift estimation provided for reference (red diagonal). *AIM*: Choose better colors?

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	$< 10\%$
Bias	$< 0.003(1+z)$

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 amount, suggested by the trend of points to lie below the diagonal at high redshift. 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.

Once propagated through the calculations of correlation functions of cosmic shear and galaxy positions, photo-z errors are not insignificant contributors to the total uncertainties reported on cosmological parameters. 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. 2017), HSC (Tanaka et al. 2018), and KiDS (Hildebrandt et al. 2017).

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. This approach tends to lead to catastrophic outliers like the horizontally oriented population of Figure 1, due to the presence of galaxies whose SEDs are not represented by the template library. Data-driven approaches tend to result in catastrophic outliers like the vertically oriented population of Figure 1, due to training sets that are incomplete in redshift coverage. Recent advances have focused on identifying and removing catastrophic outliers when using photo-zs for inference (Gorecki et al. 2014). The approaches of using a training set versus a template library are related to one another by Budavári (2009). Sophisticated Bayesian techniques and cutting-edge machine learning methods have been employed to improve precision (Carliles et al. 2010) and accuracy (Sadeh et al. 2016).

An alternative to point estimation of photo-zs is redshift probability distribution function (PDF) estimation that reports the probability $\mathrm{p}(z)$ over all possible redshifts for every galaxy rather than an 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 notions of precision, accuracy, and systematic errors. However, denoting photo-zPDFs 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-zPDFs are *posterior* PDFs, conditioned on the photometric data and prior knowledge. In terms of Figure 1, photo-zPDFs are horizontal cuts, probabilities of redshift conditioned on a specific value of data, i.e. posteriors, which constrain redshifts, whereas vertical cuts through this space are probabilities of data conditioned on a specific redshift, i.e. likelihoods, from which data is actually drawn.

Photo-zPDFs 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-zPDFs are not without their own weaknesses, however, including the resources necessary to calculate and record them for large galaxy surveys (Carrasco Kind & Brunner 2014b) 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). The most important of these issues, however, is that use of them in the literature is inconsistent at best and incorrect at worst. The most common application of photo-zPDFs is their use in estimating $N(z)$, the distribution of redshifts of a sample of galaxies, a quantity essential to the calculation of the power spectra of weak gravitational lensing and large-scale structure that are used to constrain the parameters of cosmological models Mandelbaum et al. (2008); Sheldon et al. (2012); Bonnett et al. (2016).

If the true redshifts $\{z_i^\dagger\}$ of galaxies i were known, then the redshift PDFs would be delta functions $\{\delta(z, z_i^\dagger)\}$ centered at the true redshift, and the redshift distribution could be effectively approximated by a histogram or other interpolation of the delta functions $\{\delta(z, z_i^\dagger)\}$. When photo-$z$PDFs are available instead of true redshifts, the simplest approach reduces them to point estimates $\{\hat{z}_i\}$ of redshift by using $\delta(z, \hat{z}_i)$ in place of $\delta(z, z_i^\dagger)$.

Though it is most common for $\hat{z}_i$ to be the mode (maximum) of the photo-zPDF, there are other, more principled point estimate reduction procedures (Tanaka et al. 2018).

To circumvent the reduction of uncertainty information due to point estimation, however, it has grown to be more common to calculate the mathematically inconsistent but conceptually simple *stacked estimator* of the redshift density function (Lima et al. 2008),

$$\hat{n}(z) = \frac{1}{N_{\rm tot}} \sum_{i=0}^{N_{\rm tot}} \mathrm{p}(z)_i \tag{1}$$

for a sample of $N_{\rm tot}$ galaxies i, or, equivalently, the redshift distribution function $\hat{N}(z) = N_{\rm tot}\hat{n}(z)$. Much of this thesis centers around the problems with this logically invalid yet pervasive quantity.

AIM: TODO: Pull in content from global intro of thesis.

Though their potential to improve estimates of physical parameters is tremendous, photo-zPDFs have been applied only to a limited extent. They have 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.

Despite the growing prevalence of photo-zPDF production, no implementation of inference using photo-zPDFs has yet been presented with a mathematically consistent methodology. We present and validate a hierarchical Bayesian technique for the use of photo-zPDF in inference of arbitrary statistics relevant to cosmology, large-scale structure, and galaxy evolution. For simplicity, we consider only one-point statistics, though future work will extend this methodology to higher-order statistics.

The *redshift distribution function* $N(z)$, or, almost interchangably, its normalized cousin the *redshift density function* $n(z)$, serves as an ideal statistic upon which to demonstrate this novel approach. $N(z)$ is necessary for calculations of two-point correlation functions of weak gravitational lensing and counting statistics that are used to probe dark energy (Masters et al. 2015). The observed $N(z)$ also been used to validate survey selection functions usnfenjed in generation of realistic, multi-purpose mock catalogs (Norberg et al. 2002). Additionally, $N(z)$ has been the subject of inference using photo-zPDFs before (Sheldon et al. 2012; Hildebrandt et al. 2012; Kelly et al. 2014; Benjamin et al. 2013; Bonnett 2015; Viironen et al. 2015; Asorey et al. 2016; Leistedt et al. 2016), so comparisons to the literature may easily be made.

The prevailing way to estimate $n(z)$ for a sample of galaxies j is to "stack" photo-zPDFs according to Equation 1. The stacked estimator of the redshift density function is effectively an average of the photo-zPDF catalog and, as we show in this paper, not in general mathematically valid. Stacking is nonetheless considered the preferred method for obtaining $N(z)$ from a photo-zPDF catalog (Sheldon et al. 2012; Kelly et al. 2014; Benjamin et al. 2013; Bonnett 2015; Viironen et al. 2015; Asorey et al. 2016). However, it must be noted here that Equation 1 is not in general mathematically valid. (See Hogg (2012) for a complete discussion.)

We aim to develop a clear methodology guiding the use of photo-zPDFs 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-zPDFs, rather than requiring a simultaneous derivation of the photo-zPDFs and the redshift distribution, making it preferable to ongoing surveys that may be resistant to restructuring their analysis pipelines.

In Section 2, we present the CHIPPR model and `chippr` implementation for characterizing the full posterior probability landscape of $N(z)$ using photo-zPDFs. In Section 3, we describe the experimental design for testing the fully probabilistic approach to mock and real datasets, the results of which are found in Section 4. In Section 4, we stress-test the CHIPPR model against stacking and its other competitors in the context of cosmology.

2. METHOD

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 z_j that we would like to learn; redshift is a parameter in this case. The distribution of the ensemble of redshifts $\{z_j\}$ may be described by the hyperparameters defining the redshift distribution function $n(z)$ that we would like to quantify. This situation may be considered to be a probabilistic generative model, illustrated by the directed acyclic graph of Figure 2.

The redshift distribution function $n(z)$ is the number of galaxies per unit redshift, effectively defining the evolution in the number of galaxies (Ménard et al. 2013). In the following sections, we present and compare methods for estimating $n(z)$ from photo-zPDFs. Section 2.2 contains the mathematical derivation of a probabilistic model for $n(z)$ dependent on photo-z probability distribution functions, and Section 2.4 contrasts the probabilistic model with alternative methods.

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

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

of finding a galaxy j in a catalog of J galaxies having a redshift z. We believe that galaxy redshifts are indeed 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

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

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

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

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 and photometric data vectors $\vec{d}$, which may be any combination of fluxes, magnitudes, colors, and their observational errors. In that sense $n(z)$ is equivalent to an integral

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

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

$$\vec{d}_j \;\sim\; \mathrm{p}(\vec{d} \mid z_j) \tag{6}$$

of photometry 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 $\boldsymbol{\phi}$.

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*

AIM: TODO: fix implicit posterior vs. implicit-weighted posterior

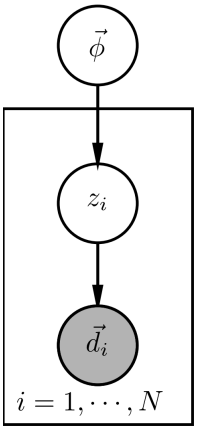

Figure 2. The directed acyclic graph of the CHIPPR model, where circles indicate random variables and arrows indicate causal relationships. The redshift distribution function parameterized by hyperparameters $\boldsymbol{\phi}$ exists independent of the survey of J galaxies, indicated as a box. The redshifts $\{z_j\}$ 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.

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, 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*. *AIM*: Am I adequately explaining the distinction between hyperparameters and parameters?

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 redshifts, 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-zPDFs, we must introduce two more concepts, confusingly named the implicit prior and the prior probability density.

When we constrain the redshift of a galaxy using its observed photometric data $\vec{d}_j$, we are effectively estimating a posterior $\mathrm{p}(z \mid \vec{d}_j)$. However, to do this, we must have a model for the general relationship between redshifts and photometry, whether empirical, as is the case

for machine learning photo-zPDF methods, or analytic, as is the case for template-based photo-zPDF methods. Such a relationship is defined in the space of probability density over redshift, so it must be able to be parameterized by the same functional form $f(z;\boldsymbol{\phi})$ as $n(z)$. It is thus natural to write it as $\mathrm{p}(z \mid \boldsymbol{\phi}^*)$, where $\boldsymbol{\phi}^*$ is the parameters for this relationship under some generic photo-zPDF method. We call $\mathrm{p}(z \mid \boldsymbol{\phi}^*)$ the *implicit prior*, as it is rarely explicitly known nor chosen by the researcher; for template-based methods where it can be chosen to be "realistic," it may be appropriate to call it an *interim prior*. Because the implicit prior is unavoidable, the photo-zPDFs reported by any method are really *implicit-weighted posteriors* $\mathrm{p}(z \mid \vec{d}, \boldsymbol{\phi}^*)$.

The implicit prior $\boldsymbol{\phi}^*$ may be thought of as an initial guess for the parameters contained in $\boldsymbol{\phi}$, inspired by the generative model for photometry from the redshift distribution functions and including some parameters defining intrinsic galaxy spectra and instrumental effects. (See Benítez (2000) for more detail.) For statistical purposes, we would like any interim prior to be uninformative, but this is rarely achievable. In the case of estimating $n(z)$ photometrically, it is common to use $\boldsymbol{\phi}^*$ corresponding to $n(z)$ derived from some different, spectroscopically confirmed sample or from a cosmological simulation.

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 parameters $\boldsymbol{\phi}$. 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.

AIM: TODO: Clarify what we want to compute ($\mathrm{p}(\boldsymbol{\phi} \mid \vec{d})$), what we're given ($\{\mathrm{p}(z \mid \vec{d}, \boldsymbol{\phi}^*)\}$). *AIM*: TODO: Give answer here, say this is correct, move derivation to appendix. *AIM*: TODO: Give human-readable description of ratios, explain that CHIPPR replaces implicit prior with sampled model, converting posterior into likelihood.

With these definitions, we obtain the desired expression for $\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})$,

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

the heart of CHIPPR. The derivation is provided in Appendix app:math.

2.2.1. *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 $\boldsymbol{\phi}^*$ 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-zPDFs, 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.

2.3. *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 2.3.1, we describe the publicly available `chippr` library. In Section **??**, we outline how `chippr` can be used to sample the full log-posterior distribution $\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})]$.

2.3.1. *Code*

`chippr` is a *Python* 2 library[5] 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

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

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

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

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 A10. `chippr` accepts a configuration file of user-specified parameters, among them

[5] https://github.com/aimalz/chippr

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.

Two threshold conditions are defined, one designating all previous samples to be ignored as as products of a burn-in phase and another indicating when a sufficient number of post-burn samples have been accepted. In this case, the first threshold (described in Section **??**) is defined in terms of sub-runs of 10^3 accepted samples, and the second is defined as an accumulation of 10^4 samples.

Though previous versions used `HDF5` for the primary I/O format due to its efficiency for large quantities of data, it was abandoned in favor of `pickle` in the working release due to the instability of the *Python* implementation of the format on high-performance computing systems. The resulting output is a set of I ordered `hickle` files enumerated by ρ containing the state information after each sub-run. The state information includes $\frac{I_0}{s}$ actual 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 $\boldsymbol{\phi}$ over the entire sub-run.

2.4. *Comparison with Alternative Approaches*

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-zPDFs: the distribution $n(z_{\mathrm{max}})$ of the redshifts at maximum posterior probability

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

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

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

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 $\langle z \rangle = \int z \mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*) \mathrm{d}z$. *AIM*: cite tanaka 2018 here directs attention to deriving an optimal point estimate reduction of a photo-zPDF, 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-zPDFs 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[6].

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-zPDFs will indeed lead to an accurate recovery of the true redshift distribution function. However, stacking is in general applied indiscriminately to broader photo-zPDFs 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 value of $\boldsymbol{\phi}$ maximizing the log posterior given by Equation A10 using any optimization code. To compare with sampling, the MMLE also depends on the choice of the hyperprior distribution, and it does not produce a full posterior probability distribution over the parameters of interest, only point estimators. It must be noted that computation of the maximum marginalized likelihood estimator may be unstable depending on the strengths and weaknesses of the optimizer. In general, derivatives will not be available for the full posterior distribution, restricting optimization methods used.

In Section 2.4.1 we outline the measures used to evaluate the performance of the method.

2.4.1. *Performance metrics*

The results of the computation described in Section 2.3.1 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 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

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

which measures a distance from parameter values $\boldsymbol{\phi}'$ to true parameter values $\boldsymbol{\phi}^{\dagger}$, which is invariant under changes of variables. We note that $KL_{\prime,\dagger} \neq KL_{\dagger,\prime}$ and is only interpretable when there is a notion that $\boldsymbol{\phi}^{\dagger}$ is closer to the truth than $\boldsymbol{\phi}'$. The KLD is explored in detail in the Appendix to Malz et al. (2018). In simulated tests, $\boldsymbol{\phi}^{\dagger}$ is the true value and $\boldsymbol{\phi}'$ is that produced by one of the methods in question.

3. VALIDATION

We compare the results of CHIPPR to those of stacking and the histogram of photo-zPDF maxima (modes) on mock data in the form of catalogs of emulated photo-zPDFs generated via the forward model discussed in Sec-

[6] And for good reason! Consider a bimodal photo-zPDF; 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.

tion 2.1. Figure 3 illustrates the implementation of this forward model used in this paper.

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

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

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 3.1), catastrophic outliers (Section 3.2), and systematic bias (Section 3.3). Tests including all three effects at the tolerance levels of LSST (see Table 1) are presented in Section 4. Figure 4 illustrates these three effects individually at twice the tolerance of LSST for demonstrative purposes, hearkening back to Figure 1. We also test nontrivial implicit priors in Section 3.4, which ought to be a priority for the photo-z community.

AIM: TODO: add in the plot from thesis intro

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

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

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 A10 to the chosen binning of redshift space.

AIM: TODO: Add back the figure of prior samples?

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.

3.1. *Intrinsic scatter*

Figure 5 shows some examples of photo-zPDFs 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-zPDFs 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.

3.2. *Catastrophic outliers*

As was covered in Section **??**, 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 examples of photo-zPDFs 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-zPDFs to point estimates.

When one thinks of the photo-zPDFs of catastrophic outliers, however, what comes to mind is multimodal photo-zPDFs, wherein reducing photo-zPDFs 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

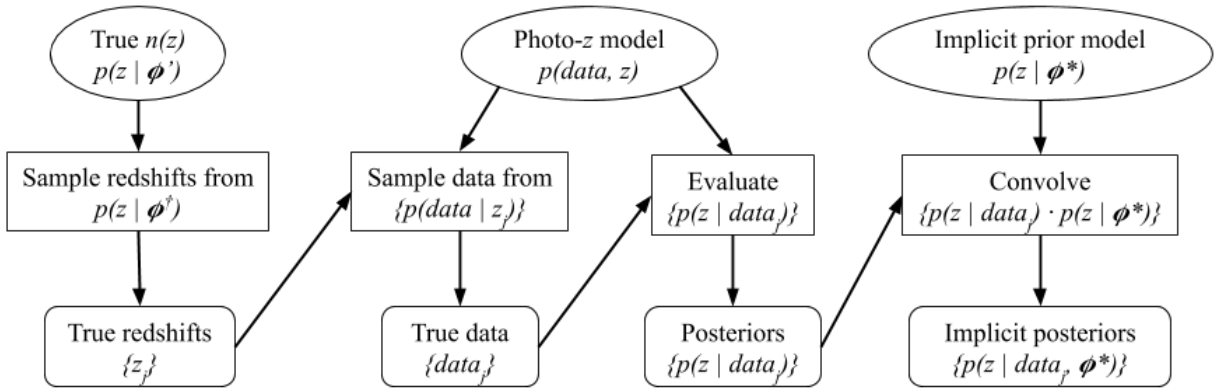

Figure 3. *AIM*: Check that all names and formulae are defined in the text and the same way as in diagram, say the whole thing in words in the text as well. Explain that this is the forward model of the validation, not of CHIPPR! contrast this with Fig. 1 (joke about simple model and complex validation?) 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-zPDF catalogs.

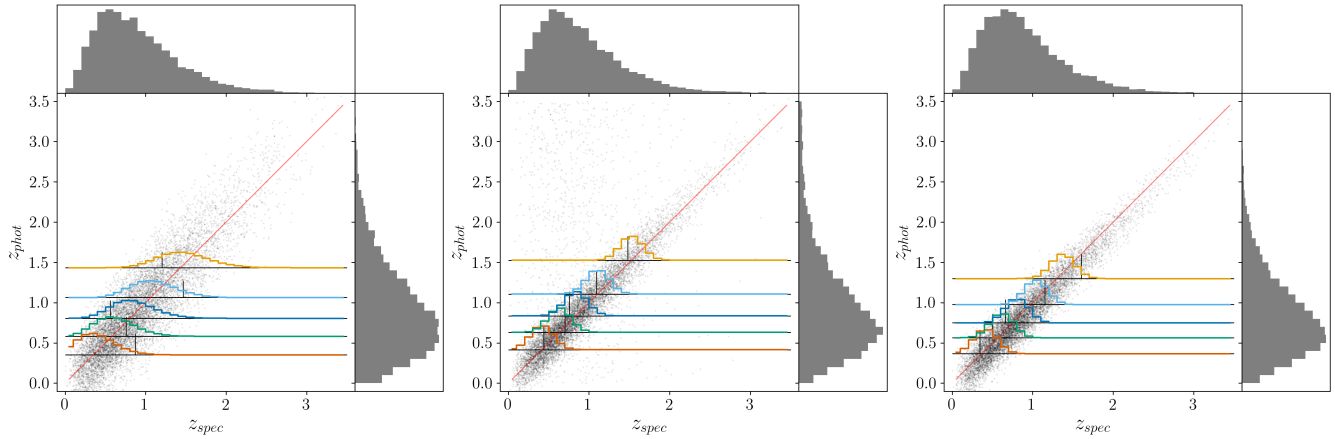

Figure 4. *AIM*: TODO: Add watermark of "mock data" in UL corner, same for "results of inference" on other kind of plot. The joint probability space of true and estimated redshift for the three concerning photo-z systematics at twice the level of the LSST requirements: intrinsic scatter (left), uniform outliers (center), and bias (right). The main panel of each shows samples (black points) in the space of mock data and redshift, akin to the standard scatterplots of true and estimated redshift, the $z_{\rm true} = z_{\rm phot}$ diagonal (thin red line), and posterior probabilities evaluated at the given estimated redshift (colored step functions). The insets show marginal histograms (gray) in each dimension, that can be compared with the true $n(z)$ used to make the figure (blue curve) to see the effect of the isolated systematic, as well as the implicit prior (red line).

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, 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.

3.3. *Systematic 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. (2017). However, in the context of photo-zPDFs, the notion of redshift bias is a form of model misspecification. Consider that if bias were included in the framework of Figure 1; a simple linear transformation of $z_{\rm phot} \to z_{\rm phot} - \Delta_z(1 + z_{\rm phot})$ would eliminate the bias. Regardless, for completeness, a test at ten times the bias of the LSST requirements, with no redshift-dependent intrinsic scatter nor catastrophic outliers, is provided in Figure 10.

As expected based on consistency with the forward model, CHIPPR is completely resistant to bias, and

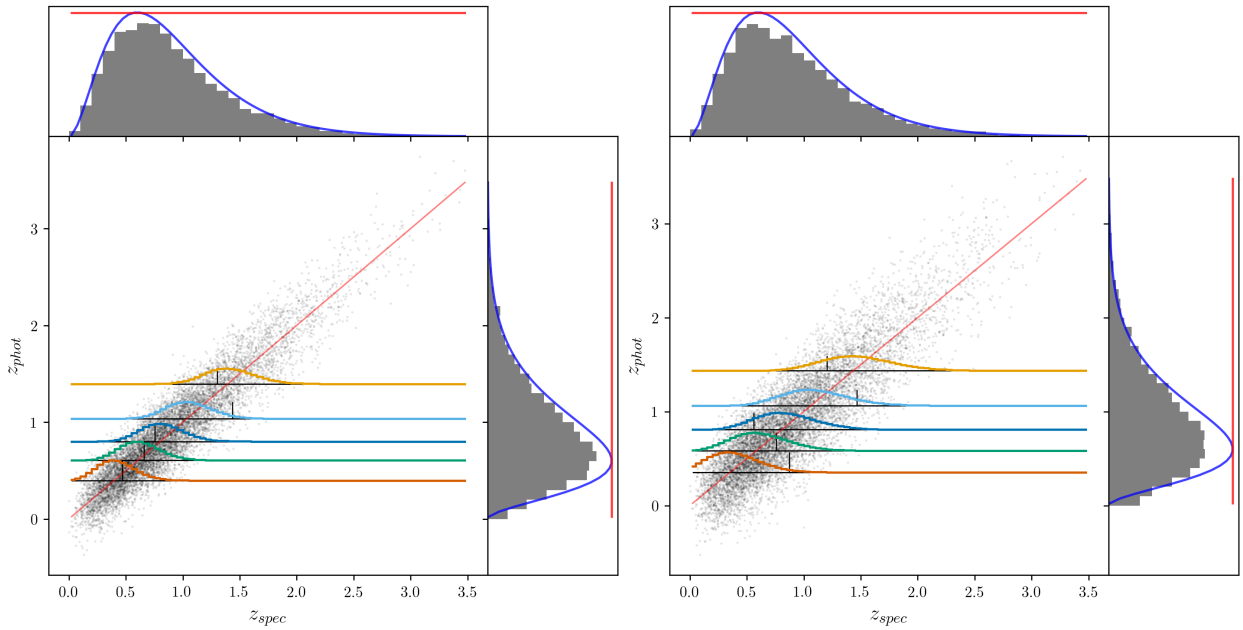

Figure 5. Examples of mock photo-zPDFs (colored lines) generated with intrinsic scatter at the LSST requirements (left) and twice the LSST requirements (right), including samples from the probability space of true and observed redshift (black points), photo-zPDFs (colored lines), the true redshifts of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

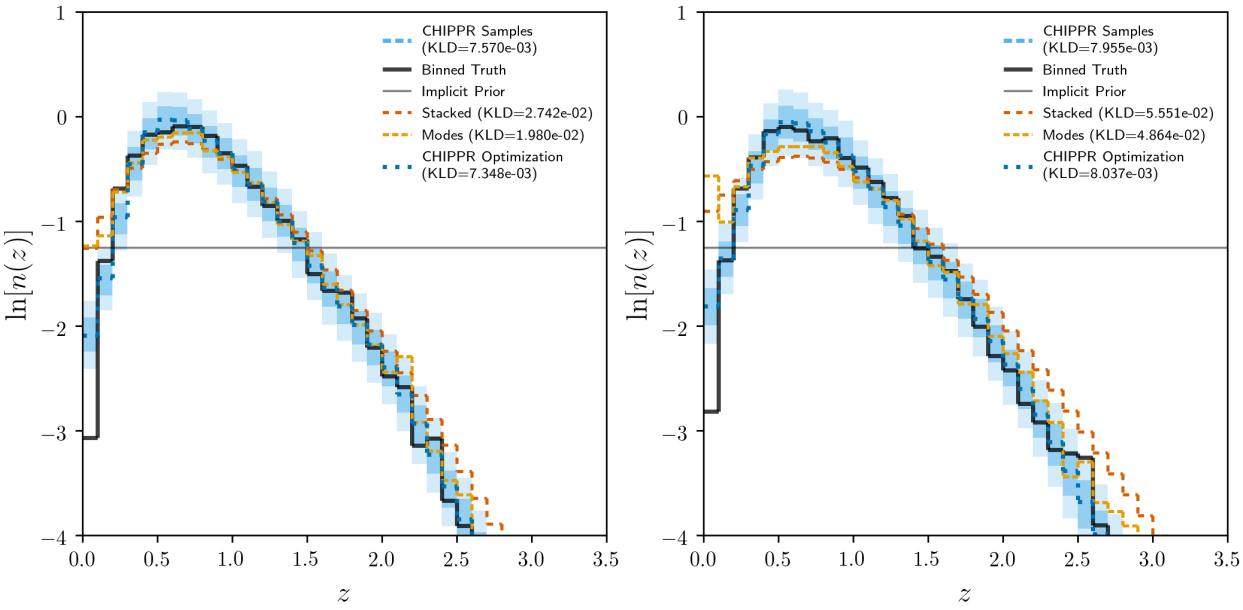

Figure 6. The results of CHIPPR (samples in light blue and optimization in dark blue) and the alternative approaches (the stacked estimator in red and the histogram of modes in yellow) on photo-zPDFs with intrinsic scatter of the LSST requirements (left) and twice that (right), with the true redshift 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.

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. *AIM*: I'm not sure I effectively explained why CHIPPR is expected to be unaffected by bias of this form.

3.4. *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-zPDF 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

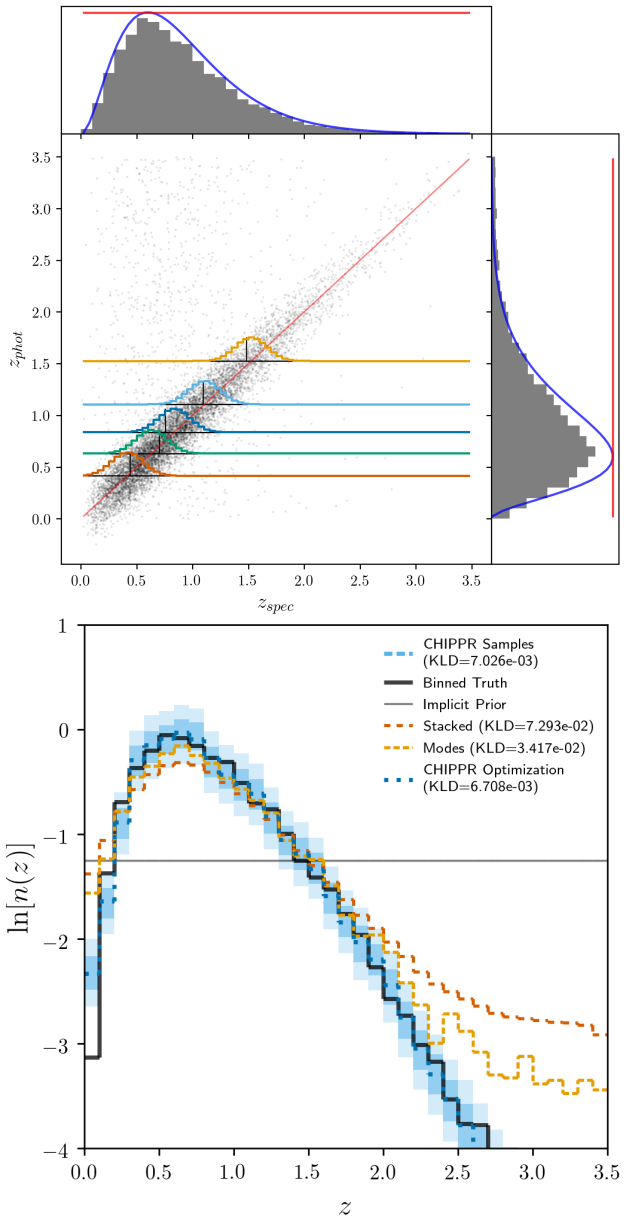

Figure 7. Top: Examples of photo-zPDFs with a uniform catastrophic outlier population at the level of the LSST requirements, including samples from the probability space of true and observed redshift (black points), photo-zPDFs (colored curves), and the true redshifts of the example photo-zPDFs (black vertical lines), with marginal histograms (gray) for each dimension with the true redshift distribution (blue curve) and implicit prior (red curve) 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-zPDFs 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.

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-zPDFs 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 photo-z method (right panel). One can see that the photo-zPDFs 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-zPDFs 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-zPDF catalogs.

4. DISCUSSION

The experiments of Section 3 isolate the potential sources of error in $n(z)$ estimation one at a time. Now, we stress-test CHIPPR by investigating two realistically complex cases, one in which the $n(z)$ estimates are made tomographically as in a modern cosmological analysis (Section 4.1) and one in which the $n(z)$ estimators are not provided with the same implicit prior used to generate the photo-zPDF catalog (Section 4.2).

4.1. *LSST Requirements*

AIM: TODO: clarify the scope: the "binned" samples are already determined by some observational property but redshift distributions are probabilistic as follows.

It is of interest to explore the impact of incorrectly estimated $n(z)$ on the cosmological inference 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 simulate mock data with all three effects with which LSST is concerned at the levels of Table 1 and propagated the results of CHIPPR and the other estimators to a Fisher matrix forecast using `CosmoLike` (Krause & Eifler 2017), a publicly available cosmological forecasting code. Though redshift tomography is non-physical, as redshift is a continuous random variable, and binning in estimated redshift introduces poorly understood systematic error, we perform this analysis as an example of how it affects the current standard in how cosmological parameters are constrained by galaxy surveys, rather than how we think they ought to be constrained.

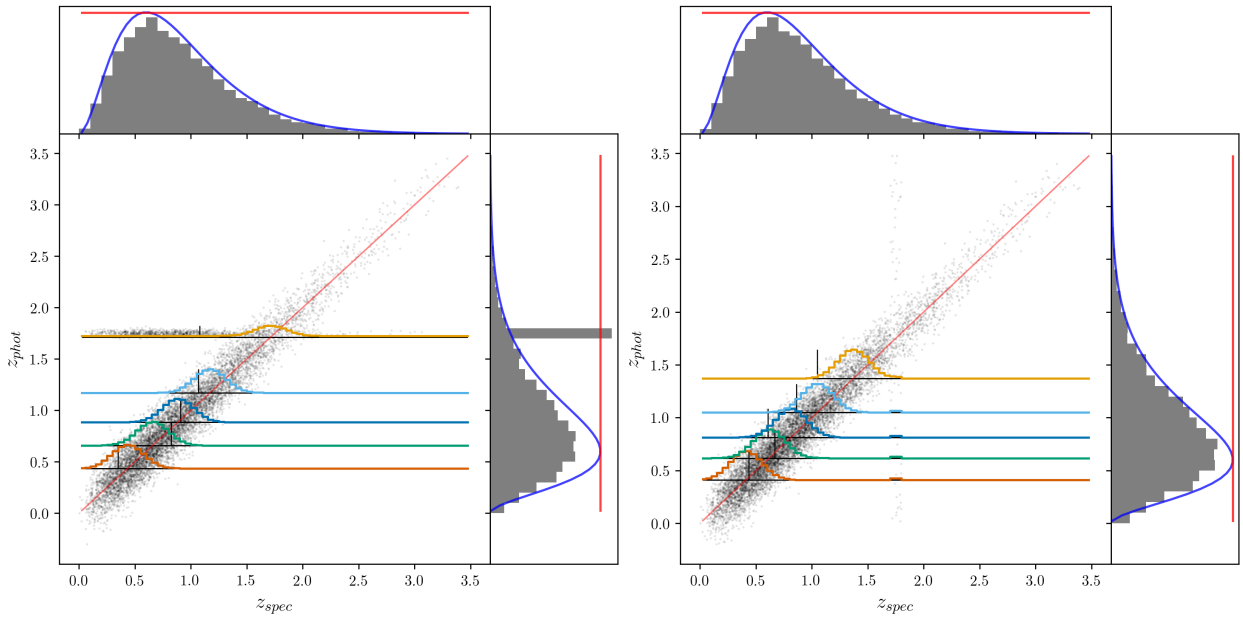

Figure 8. Examples of photo-zPDFs with a catastrophic outlier population like that seen in template-fitting photo-zPDF codes (left) and machine learning photo-zPDF codes (right), including samples from the probability space of true and observed redshift (black points), photo-zPDFs (colored curves), and the true redshifts of the example photo-zPDFs (black vertical lines), with marginal histograms (gray) for each dimension with the true redshift distribution (blue curve) and implicit prior (red curve) in the insets.

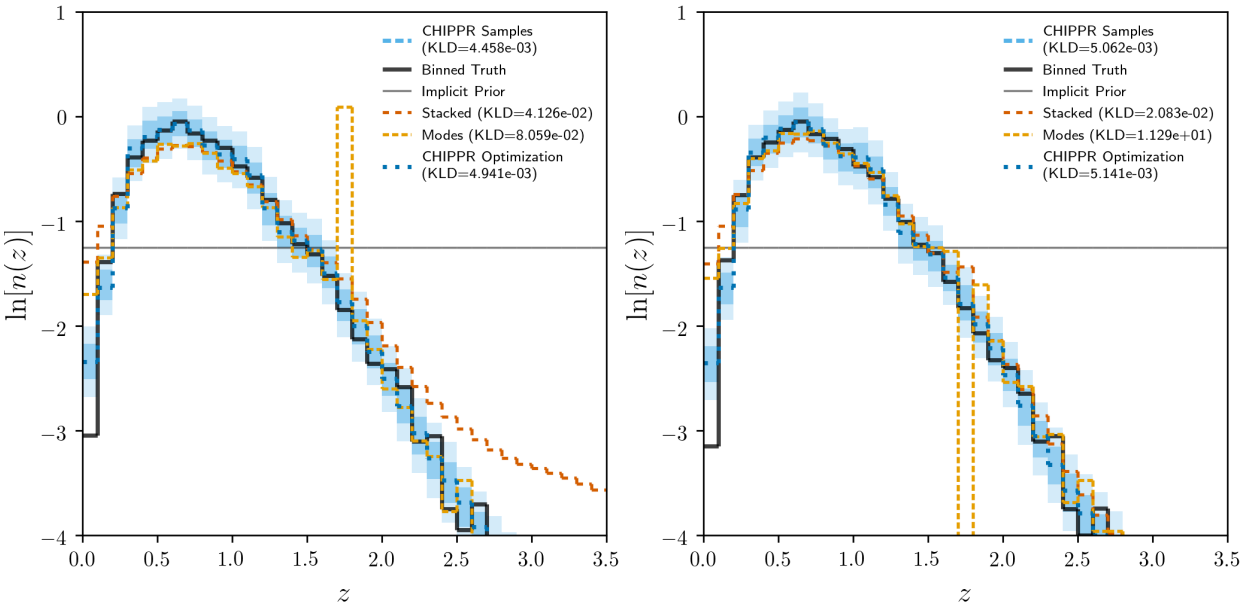

Figure 9. The results of CHIPPR (samples in light blue and optimization in dark blue) and the alternative approaches (the stacked estimator in red, the histogram of modes in yellow) on photo-zPDFs with catastrophic outliers like those seen in template-fitting photo-zPDF codes (left) and machine learning photo-zPDF 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).

We consider a set of tomographically binned $n(z)$ and cosmological parameter covariance matrices used for LSST-DESC forecasting, for which the true $n(z)$ in each pre-defined bin is already provided in the form of an evaluation of a function on a fine grid of 350 redshifts $0.0101 < z < 3.5001$. First, we bin them down to a piecewise constant parameterization with a manageable 35 parameters 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 Figure 13. We emulate photo-zPDFs 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. Illustrations of this process are provided in Figure 14.

AIM: TODO: Ask Elisabeth how to get the bias here

We then make a point estimate of $n(z)$ using CHIPPR as well as the alternative methods on the photo-zPDF

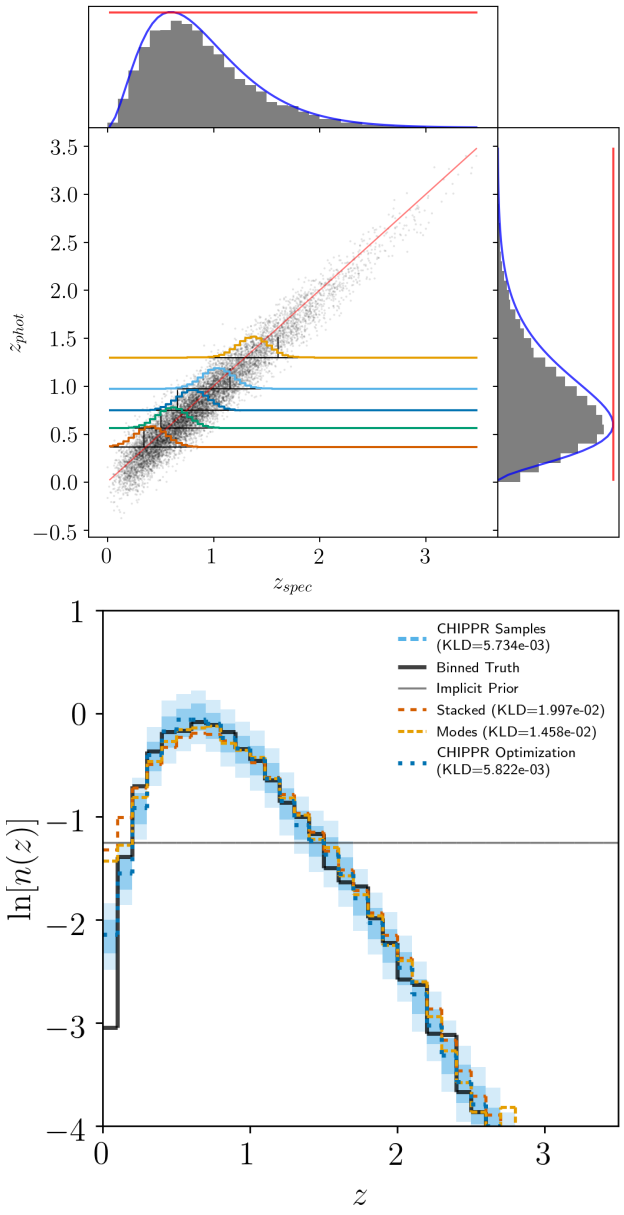

Figure 10. Left: Examples of photo-zPDFs with ten times the bias of the LSST requirements, including samples from the probability space of true and observed redshift (black points), photo-zPDFs (colored curves), and the true redshifts of the example photo-zPDFs (black vertical lines), with marginal histograms (gray) for each dimension with the true redshift distribution (blue curve) and implicit prior (red curve) in the insets. Right: 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-zPDFs with uniformly distributed catastrophic outliers, 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.

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 $n(z)$. Note that Figure 15 is shown in linear rather than log probability units, unlike all other plots in this paper, to better show the behavior at low probability. The excessive breadth of the alternative estimators can be seen quite plainly.

We then use the different estimators of $n(z)$ in a cosmological forecasting procedure with `CosmoLike` , constraining Ω_m, Ω_b, w_a, w_0, n_s, S_8, and H_0. Though there are also slight differences in the angle of the error ellipses, most of the differences are due to broadening of the 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 likelihood estimate for all parameters except Ω_b and H_0. Stacking, however, outperforms the histogram of modes for all parameters except Ω_m and S_8, for which their constraints are quite similar. Though the true values of the parameters themselves were not accessible with the Fisher matrix-based framework, we calculate the seven-dimensional KLD for the three $n(z)$ estimators relative to the constraints derived from the true $n(z)$, showing that CHIPPR preserves information $200 - 800$ times better than the alternatives, with the histogram of modes doing about four times better than stacking.

AIM: Hogg says "I think it would be good to talk a little quantitatively about where people need to know N(z) and other one-point statistics, and how much they will get various things wrong if they don't know these correctly. And situate that discussion within the current context of cosmological parameter estimation and precision cosmology." Did I successfully answer the question?

4.2. *Violations of the model*

AIM: Move Section 3.4 here?

In this test, without tomographic binning, 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-zPDF 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-zPDF catalog produced by a generic machine learning code and assumed a flat implicit prior, we would observe the contents of Figure 17.

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-zPDF methods always characterize and provide the implicit prior.

5. CONCLUSION

AIM: TODO: new paragraph for what can go wrong, call to community for what to do about it, what aspects of implicit prior are knowable and not knowable? if testable how so? outside the scope of paper, data producers be warned! a likelihood is better – give us that if you can! focus on methods that acknowledge probabilistic structure of problem

AIM: claim: implications for tomographic binning are severe

This study derives and demonstrates a mathematically consistent inference of a one-point statistic, the redshift

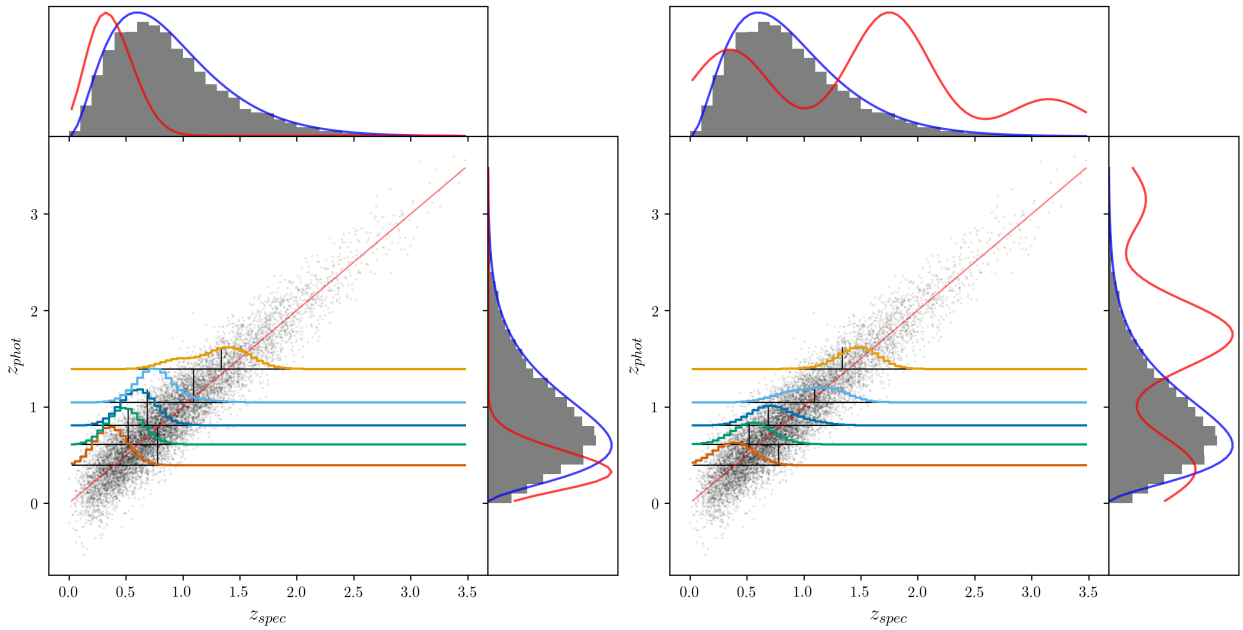

Figure 11. Examples of mock photo-zPDFs (colored lines) generated with a machine learning-like implicit prior (left) and a template-fitting-like implicit prior (right), including samples from the probability space of true and observed redshift (black points), photo-zPDFs (colored lines), the true redshifts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

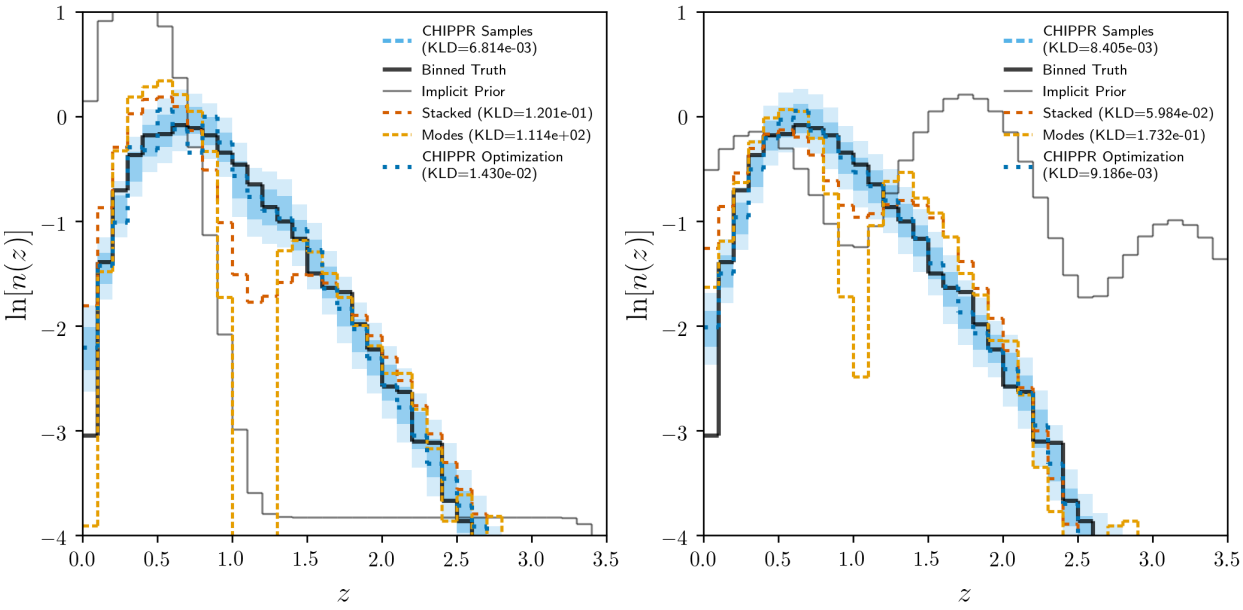

Figure 12. The results of CHIPPR (samples in light blue and optimization in dark blue) and the alternative approaches (the stacked estimator in red and the histogram of modes in yellow) on photo-zPDFs with an implicit prior like that of machine learning photo-zPDF approaches (left) and an implicit prior like that of template-fitting photo-zPDF 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.

density function $n(z)$, based on an arbitrary catalog of photo-zPDFs. The fully Bayesian method, 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 method is validated on mock data and tested in the regime of LSST with promising results, outperforming the traditional stacking estimator at the level of $n(z)$ as well as in terms of constraining power on the cosmological parameters. Not only is this the only mathematically correct approach to the problem, it also recovers the true parameter values better than popular alternatives, as measured by the loss of information in $n(z)$ and the size of error ellipses in the space of cosmological parameters.

The following conclusions and recommendations can be made with confidence:

1. Both the CHIPPR marginalized maximum likeli-

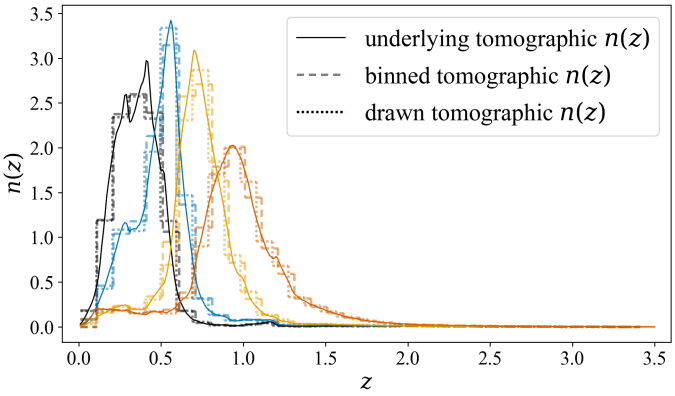

Figure 13. The LSST-like tomographic binning and true redshift distribution, where the truth (solid) is a PDF evaluated on a fine grid of 350 redshifts $0.0101 < z < 3.5001$, and the binned (dashed) and drawn (dotted) $n(z)$ are piecewise constant functions evaluated in 35 evenly spaced bins, for four different tomographic bins (colors).

hood 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.

2. 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.

3. When the implicit prior is known to be a poor match to the data, only the results of CHIPPR are satisfactory estimators of the redshift distribution function because they are the only methods that can account for the bias induced on the photo-zPDF catalog by the method that produces it; this is the most compelling case for the sampler because of the ubiquity of inappropriate interim priors.

By showing that CHIPPR is effective in recovering the true redshift distribution function and posterior distributions on its parameters from catalogs of photo-zPDFs, this work supports the production of photo-zPDFs by upcoming photometric surveys such as LSST to enable more accurate inference of the cosmological parameters. We discourage researchers from co-adding photo-zPDFs or converting them into point estimates of redshift and instead recommend the use of Bayesian probability to guide the usage of photo-zPDFs. We emphasize to those who produce photo-zPDFs from data that it is essential to release the implicit prior used in generating this data product in order for proper inference to be conducted by consumers of this information.

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.

APPENDIX
DERIVATION

In this paper, we work exclusively with 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

$$\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})] = \ln[\mathrm{p}(\{\vec{d}_j\} \mid \boldsymbol{\phi})] + \ln[\mathrm{p}(\boldsymbol{\phi})] - \ln[\mathrm{p}(\{\vec{d}_j\})] \tag{A1}$$

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

$$\ln[\mathrm{p}(\{\vec{d}_j\} \mid \boldsymbol{\phi})] = \ln\left[\int \mathrm{p}(\{\vec{d}_j\} \mid \{z_j\})\mathrm{p}(\{z_j\} \mid \boldsymbol{\phi})\mathrm{d}\{z_j\}\right]. \tag{A2}$$

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

$$\ln[\mathrm{p}(\{\vec{d}_j\} \mid \{z_j\})] = \sum_{j=1}^{J} \ln[\mathrm{p}(\vec{d}_j \mid z_j)], \tag{A3}$$

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 $\mathrm{p}(z \mid \boldsymbol{\phi})$. Additionally, J itself is a Poisson random variable. The combination of these assumptions is given by

$$\ln[\mathrm{p}(\{z_j\} \mid \boldsymbol{\phi})] = -\int f(z;\boldsymbol{\phi})\mathrm{d}z + \sum_{j=1}^{J} \ln[\mathrm{p}(z_j \mid \boldsymbol{\phi})]. \tag{A4}$$

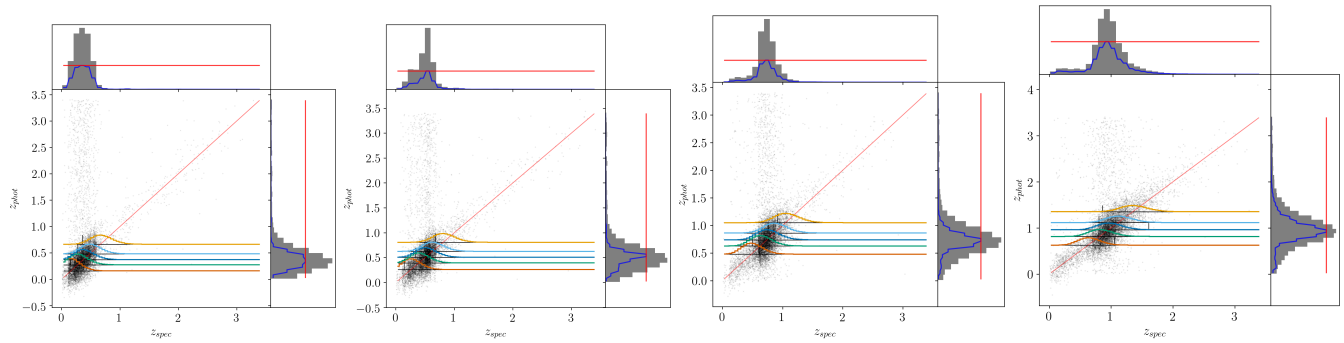

Figure 14. As in Figure 4, with a different tomographic bin in each panel and the three effects of intrinsic scatter, uniformly distributed catastrophic outliers, and bias at the levels of the LSST SRD, given in Table 1. *AIM*: TODO: Make this one big plot instead of four little ones to eliminate repeated insets and legend, and make the axis labels bigger.

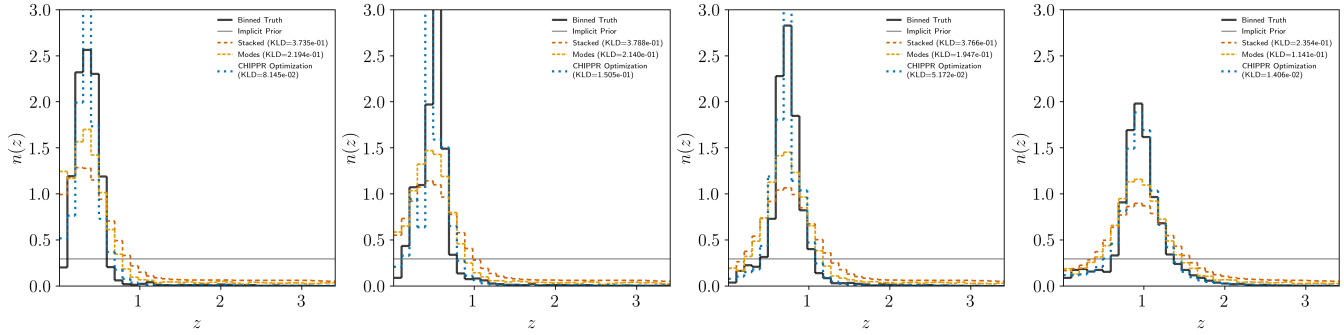

Figure 15. The `chippr` -derived and other estimators of $n(z)$ in each tomographic bin, with the true $n(z)$ (black), the implicit prior (gray), stacked estimator (red), histogram of modes (yellow), and CHIPPR marginalized maximum likelihood estimate (blue). The result of stacking is far too broad for LSST-like photo-zPDFs, even moreso than the simplistic histogram of modes. *AIM*: TODO: Make this one big plot instead of four little ones to eliminate repeated axis labels and legend, and make the legend text bigger.

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

$$\begin{aligned}\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})] \propto \ln[\mathrm{p}(\boldsymbol{\phi})] - \int f(z;\boldsymbol{\phi})\mathrm{d}z \\ + \sum_{j=1}^{J} \ln\left[\int \mathrm{p}(\vec{d}_j \mid z)\mathrm{p}(z \mid \boldsymbol{\phi})\mathrm{d}z\right].\end{aligned} \tag{A5}$$

Since we only have access to implicit photo-z posteriors, we must be able to write the full log-posterior in terms of implicit photo-z log-posteriors rather than the log-likelihoods of Equation A5. To do so, we will need an explicit statement of this implicit prior $\boldsymbol{\phi}^*$ for whatever method is chosen to produce the implicit photo-z 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

$$\begin{aligned}\ln[\mathrm{p}(\vec{d}_j \mid z)] = \ln[\mathrm{p}(\vec{d}_j \mid z)] + \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] \\ - \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)].\end{aligned} \tag{A6}$$

Once the implicit prior $\boldsymbol{\phi}^*$ is explicitly introduced, we may expand the last term in Equation A6 according to Bayes' Rule to get

$$\begin{aligned}\ln[\mathrm{p}(\vec{d}_j \mid z)] = \ln[\mathrm{p}(\vec{d}_j \mid z)] + \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] \\ + \ln[\mathrm{p}(\vec{d}_j \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)] \\ - \ln[\mathrm{p}(\vec{d}_j \mid z, \boldsymbol{\phi}^*)].\end{aligned} \tag{A7}$$

Because there is no direct dependence of the data upon the hyperparameters, we may again expand the term $\ln[\mathrm{p}(\vec{d}_j \mid$

Figure 16. The result of propagating the estimators of $n(z)$ by stacking (red), the histogram of modes (yellow), CHIPPR (blue), and the true $n(z)$ (black) 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 for most parameters, stacking and the histogram of modes yield broader contours. *AIM*: Are the contours any easier to see now?

$z, \boldsymbol{\phi}^*)]$ to obtain

$$\begin{aligned} \ln[\mathrm{p}(\vec{d}_j \mid z)] = &\ln[\mathrm{p}(\vec{d}_j \mid z)] + \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] \\ &+ \ln[\mathrm{p}(\vec{d}_j \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)] \\ &- \ln[\mathrm{p}(\vec{d}_j \mid \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(\vec{d}_j \mid z)]. \end{aligned} \tag{A8}$$

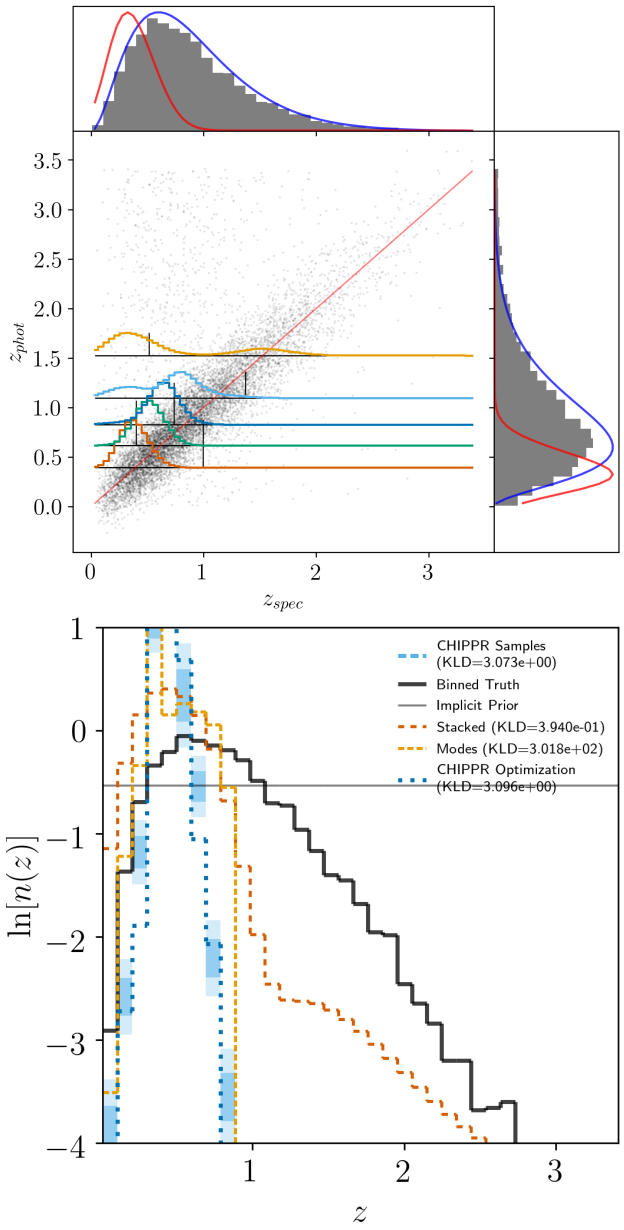

Figure 17. Left: Examples of photo-zPDFs with systematics at the level of the LSST requirements with a machine learning-like implicit prior, including samples from the probability space of true and observed redshift (black points), photo-zPDFs (colored curves), and the true redshifts of the example photo-zPDFs (black vertical lines), with marginal histograms (gray) for each dimension with the true redshift distribution (blue curve) and implicit prior (red curve) in the insets. Right: 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-zPDFs with uniformly distributed catastrophic outliers, with the true redshift density (black curve) and the uniform implicit prior given to `chippr` (gray curve). When the incorrect implicit prior is provided to `chippr` , even Bayesian inference cannot recover the true $n(z)$.

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

$$\ln[\mathrm{p}(\vec{d}_j \mid z)] = \ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)]. \tag{A9}$$

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

$$\ln[\mathrm{p}(\boldsymbol{\phi} \mid \{\vec{d}_j\})] \propto \ln[\mathrm{p}(\boldsymbol{\phi})] + \ln\left[\int \exp\left[\sum_{j=1}^{J} \Big(\ln[\mathrm{p}(z \mid \vec{d}_j, \boldsymbol{\phi}^*)] + \ln[\mathrm{p}(z \mid \boldsymbol{\phi})] - \ln[\mathrm{p}(z \mid \boldsymbol{\phi}^*)]\Big)\right] \mathrm{d}z\right], \tag{A10}$$

AIM: TODO: fix the trailing parentheses/brackets.

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 photo-z log-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

$$\mathrm{p}(z \mid \boldsymbol{\phi}) = \frac{f(z; \boldsymbol{\phi})}{\int f(z; \boldsymbol{\phi}) \mathrm{d}z}. \tag{A11}$$

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. The method outlined here is valid regardless of how the implicit photo-z log-posteriors are obtained so the many approaches to producing photo-zPDFs will not be rehashed; 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), and will be briefly reviewed in *AIM*: cite PZDC1 paper here.

CONVERGENCE CRITERIA

In addition to qualitative visual inspection of the chains, two quantities that probe the convergence of the sampler are used in this study, the autocorrelation time and the Gelman-Rubin convergence criterion.

The autocorrelation time is effectively a measure of the efficiency of the method and can be described as the expected number of iterations necessary to accept a new sample independent of the current accepted sample. A sampler that converges faster will have a smaller autocorrelation time, and smaller autocorrelation times are preferable because it means fewer iterations are wasted on non-independent samples when independent samples are desired. See Foreman-Mackey et al. (2013) for a more complete exploration of the autocorrelation time. In all tests discussed here, autocorrelation times across walkers and parameters were approximately 20, meaning two samples 20 or more iterations apart were independent, a satisfactory level of efficiency. Low autocorrelation times are a necessary but not always sufficient convergence condition, as the autocorrelation times calculated for tests in this paper were constant across all sub-runs, even those that were obviously burning in.

The Gelman-Rubin statistic

$$R_k = \sqrt{\frac{(1 - \frac{2}{I_0})w_k + \frac{2}{I_0} b_k}{w_k}}, \tag{B1}$$

a weighted sum of the mean w_k of the variances within individual walkers' chains and the variance b_k between chains of different walkers m, is calculated over each sub-run i to determine the duration of the burn-in period. Convergence is achieved when the statistic approaches unity.

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 DESC0014664. 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. *AIM*: TODO: Send draft around to Foreman-Mackey, Leistedt, others for feedback. This work was completed with generous nutritional support from the Center for Computational Astrophysics.

REFERENCES

Abell, P. A., Allison, J., Anderson, S. F., et al. 2009
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, Astrophys. J., 683, 12
Baum, W. A. 1962, Proc. from IAU Symp.
Benítez, N. 2000, The Astrophysical Journal, 536, 571
Benjamin, J., Van Waerbeke, L., Heymans, C., et al. 2013, Mon. Not. R. Astron. Soc., 431, 1547
Bonnett, C. 2015, Mon. Not. R. Astron. Soc., 449, 1043
Bonnett, C., Troxel, M. A., Hartley, W., et al. 2016, Physical Review D, 94, 042005
Budavári, T. 2009, Astrophys. J., 695, 747
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, Astrophys. J., 712, 511
Carrasco Kind, M., & Brunner, R. J. 2013, Mon. Not. R. Astron. Soc., 432, 1483
—. 2014a, Mon. Not. R. Astron. Soc., 442, 3380
—. 2014b, Monthly Notices of the Royal Astronomical Society, 441, 3550
Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, \apj, 775, 93
DiPompeo, M. A., Bovy, J., Myers, A. D., & Lang, D. 2015, Mon. Not. R. Astron. Soc., 452, 3124

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publ. Astron. Soc. Pacific, 125, 306
Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, Astrophys. J., 795, 64
Gorecki, A., Abate, A., Ansari, R., et al. 2014, Astron. Astrophys., 561, A128
Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, \aap, 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, Monthly Notices of the Royal Astronomical Society, 465, 1454
Hogg, D. W. 2012, arXiv
Hoyle, B., Gruen, D., Bernstein, G. M., et al. 2017, Dark Energy Survey Year 1 Results: Redshift distributions of the weak lensing source galaxies, Tech. Rep. Fermilab PUB-17-293-AE
Jain, B., Spergel, D., Bean, R., et al. 2015, arXiv:1501.07897 [astro-ph]
Kelly, P. L., von der Linden, A., Applegate, D. E., et al. 2014, Mon. Not. R. Astron. Soc., 439, 28
Koo, D. C. 1999, arXiv:astro-ph/9907273
Krause, E., & Eifler, T. 2017, Monthly Notices of the Royal Astronomical Society, 470, 2100
Leistedt, B., Mortlock, D. J., & Peiris, H. V. 2016, Monthly Notices of the Royal Astronomical Society, 460, 4258
Leung, A. S., Acquaviva, V., Gawiser, E., et al. 2017, The Astrophysical Journal, 843, 130
Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, Mon. Not. R. Astron. Soc., 390, 118
Malz, A. I., Marshall, P. J., DeRose, J., et al. 2018, The Astronomical Journal, 156, 35
Mandelbaum, R. 2017, arXiv:1710.03235 [astro-ph]
Mandelbaum, R., Seljak, U., Hirata, C. M., et al. 2008, Monthly Notices of the Royal Astronomical Society, 386, 781
Masters, D., Capak, P., Stern, D., et al. 2015, The Astrophysical Journal, 813, 53
Ménard, B., Scranton, R., Schmidt, S., et al. 2013, arXiv, 10
Norberg, P., Cole, S., Baugh, C. M., et al. 2002, Mon. Not. R. Astron. Soc., 336, 907
Rohatgi, A. 2019, WebPlotDigitizer
Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, Publications of the Astronomical Society of the Pacific, 128, 104502
Sanchez, A. G., Kazin, E. A., Beutler, F., et al. 2013, Mon. Not. R. Astron. Soc., 433, 1202
Sheldon, E. S., Cunha, C. E., Mandelbaum, R., Brinkmann, J., & Weaver, B. A. 2012, The Astrophysical Journal Supplement Series, 201, 32
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, Publications of the Astronomical Society of Japan, 70, S9
van Breukelen, C., & Clewley, L. 2009, Mon. Not. R. Astron. Soc., 395, 1845
Viironen, K., Marín-Franch, A., López-Sanjuan, C., et al. 2015, Astron. Astrophys., 576, A25