Draft version June 29, 2018
Preprint typeset using LaTeX style emulateapj v. 01/23/15

HOW TO OBTAIN THE REDSHIFT DISTRIBUTION FROM PROBABILISTIC REDSHIFT ESTIMATES

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

Draft version June 29, 2018

ABSTRACT

A trustworthy estimate of the redshift distribution $n(z)$ is crucial for weak-lensing cosmology as we know it. Spectroscopic redshifts for the dim and numerous galaxies of weak-lensing surveys are expected to be inaccessible, 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 hierarchical inference of $n(z)$, which is provably the only self-consistent way to combine photo-z PDFs to produce an estimator of $n(z)$. The novel 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. We conclude by propagating these effects to constraints in the space of cosmological parameters.

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

1. INTRODUCTION

Brief literature review addressing how photo-z PDFs are currently used in cosmology

Q: Why should we question existing methods?
A: Stacking is bad and only looks like it works because of assumptions that don't hold when the data is as bad as we anticipate. Cite pedantic doc?

overview of next questions: How can we improve the effectiveness of using photo-z PDFs in inference? How does the result of CHIPPR compare to established estimators in terms of the accuracy of $n(z)$? Reach goal: How significant is the effect of the discrepancy between $n(z)$ estimators on cosmological constraints?

2. METHODS

Q: How can we improve the effectiveness of using photo-z PDFs in inference?
A: Hierarchical inference is the only self-consistent way.

2.1. *Model generalities*

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

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

of finding a galaxy j in a catalog of J galaxies having a redshift z. We may without loss of generality impose a parametrization

$$f_\phi(z) \;\equiv\; n(z) \qquad (2)$$

in terms of some parameter vector ϕ. We believe that galaxy redshifts are indeed drawn from $n(z)$, making it a probability density over redshift; this fact can also be confirmed from dimensional analysis of Equation 1. Therefore, it can be rewritten as

$$z_j \;\sim\; \Pr(z \mid \phi) \;\equiv\; f_\phi(z), \qquad (3)$$

a probability density over redshift conditioned on the parameters defining $n(z)$. Note that z_j does not depend on $z_{j'}$, 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 PDF from which photometric data d, which may be any combination of fluxes, magnitudes, colors, and their observational errors, are drawn. Such a PDF over data is a likelihood

$$d \;\sim\; \Pr(d \mid z) \qquad (4)$$

conditioned on redshift. This assumption that the data are drawn from some function of the redshift forms the foundation upon which photo-z estimation is based. Note that galaxies may have different observational data d despite sharing the same redshift and that the data d_j of one galaxy is causally independent of the redshifts $z_{j'}$ and data $d_{j'}$ of other galaxies.

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 ϕ.

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

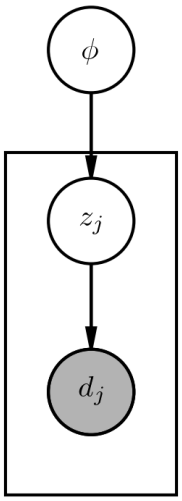

Figure 1. The PGM

2. Galaxy redshifts $\{z_j\}$ are independent draws from $\Pr(z \mid \phi)$.

3. Galaxy photometry d_j is drawn from the likelihoods $\Pr(d|z_j)$.

A forward model such as this corresponds to a probabilistic graphical model (PGM), represented by a directed acyclic graph (DAG) as in Figure 1. 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.

2.2. *Application to cosmology*

The problem facing cosmologists is to determine the true parameters ϕ_0 of $n(z)$ from observing the photometry $\{d_j\}$ of a large sample of galaxies j. To self-consistently propagate the uncertainty in the redshifts, however, it is more appropriate to estimate the posterior $\Pr(\phi \mid \{d_j\})$ over all possible parameters ϕ (and thus potential redshift distributions $n(z)$) conditioned on all the observed data $\{d_j\}$ available in a generic catalog.

In order to use the DAG of Figure 1 to derive an expression for $\Pr(\phi \mid \{d_j\})$ in terms of photo-z PDFs, 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 d_j, we are effectively estimating a posterior $\Pr(z \mid d_j)$. 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-z PDF methods, or analytic, as is the case for template-based photo-z PDF 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 $f_\phi(z)$ as $n(z)$. It is thus natural to write it as $\Pr(z \mid \phi^*)$, where ϕ^* is the parameters for this relationship under some generic photo-z PDF method. We call $\Pr(z \mid \phi^*)$ 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-z PDFs reported by any method are really *implicit-weighted posteriors* $\Pr(z \mid d, \phi^*)$.

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

With these definitions, we obtain the desired expression for $\Pr(\phi \mid \{d_j\})$,

$$\ln[\Pr(\phi \mid \{d_j\})] \propto \ln[\Pr(\phi)] + \ln\left[\int dz \right. \\ \exp\left[\sum_{j=1}^{J} (\ln[\Pr(z \mid d_j, \phi^*)] \right. \\ \left. \left. + \ln[\Pr(z \mid \phi)] - \ln[\Pr(z \mid \phi^*)] \;) \;] \;] \right], \quad (5)$$

which is the heart of CHIPPR. The entire derivation of Equation 5 is provided in Appendix A.

2.3. *Implementation*

In this study, we compare the results of Equation 5 to those of the two most common approaches to estimating $n(z)$ from a catalog of photo-z PDFs: the distribution $n(z_{\rm max})$ of the redshifts at maximum posterior probability (i.e. the modes of the photo-z PDFs), and the stacked estimator

$$\hat{n}(z) \equiv \frac{1}{J} \sum_{j=1}^{J} \Pr(z \mid d_j, \phi^*). \quad (6)$$

We perform this comparison on mock data in the form of catalogs of emulated photo-z PDFs generated via the forward model discussed above and presented in detail in Appendix B. The mock data emulates the three sources of error of highest concern to the photo-z community: intrinsic scatter, catastrophic outliers, and systematic bias. Figure 2 illustrates these three effects individually at ten times the tolerance of the upcoming Large Synoptic Survey Telescope (LSST). Tests including all three effects at the tolerance levels of LSST are presented in Section 3.

2.4. *Limitations*

Finally, we explicitly review the assumptions made by this approach.

- prior $\Pr(\phi)$
- implicit prior ϕ^* known
- photo-z PDFs are accurate posteriors

3. RESULTS

Q: How does the result of CHIPPR compare to established estimators in terms of the accuracy of $n(z)$?
A: CHIPPR yields the best possible $n(z)$, conditional on the accuracy of the photo-z PDFs used.

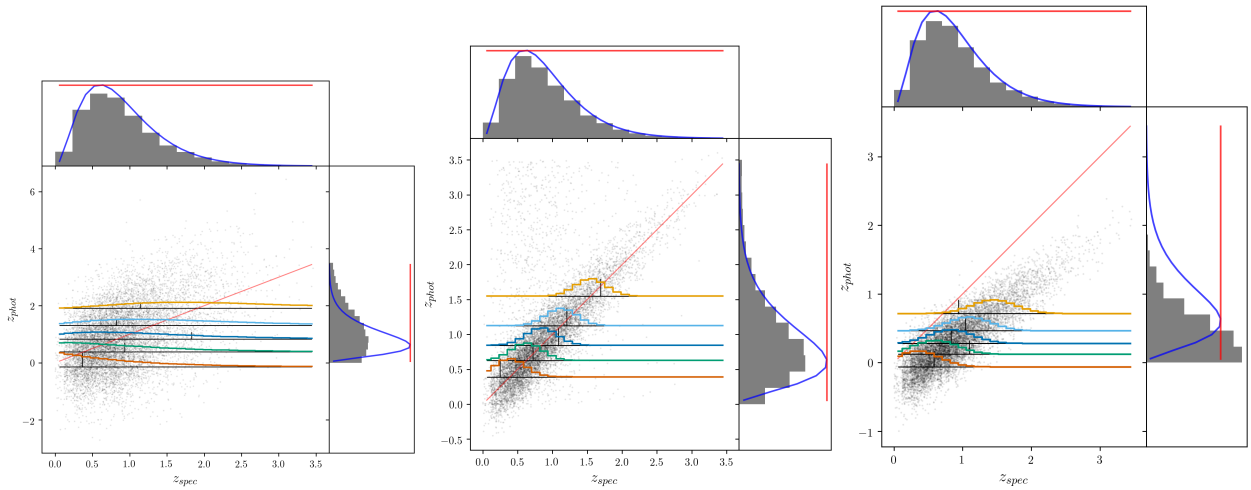

Figure 2. intrinsic scatter (left), uniform outliers (center), bias (right); Not sure how best to broach the subject of nonuniform outliers and bias as model misspecification

3.1. *LSST Requirements*

LSST specs: biased to 0.003, 10% uniform outliers, z-dependent scatter of 0.05

3.2. *Violations of the model*

mischaracterized interim prior, with idealized or realistic data?

4. DISCUSSION

5. CONCLUSION

APPENDICES

A. MATHEMATICAL DERIVATION

B. MOCK DATA GENERATION

Plots: flow chart of forward model

AIM thanks Elisabeth Krause for assistance with the `CosmoLike` code, Mohammadjavad Vakili for insightful input on statistics, Geoffrey Ryan for advice on debugging, and Boris Leistedt for helpful comments provided in the preparation of this paper. This work was completed under the generous nutritional support of the Center for Computational Astrophysics.

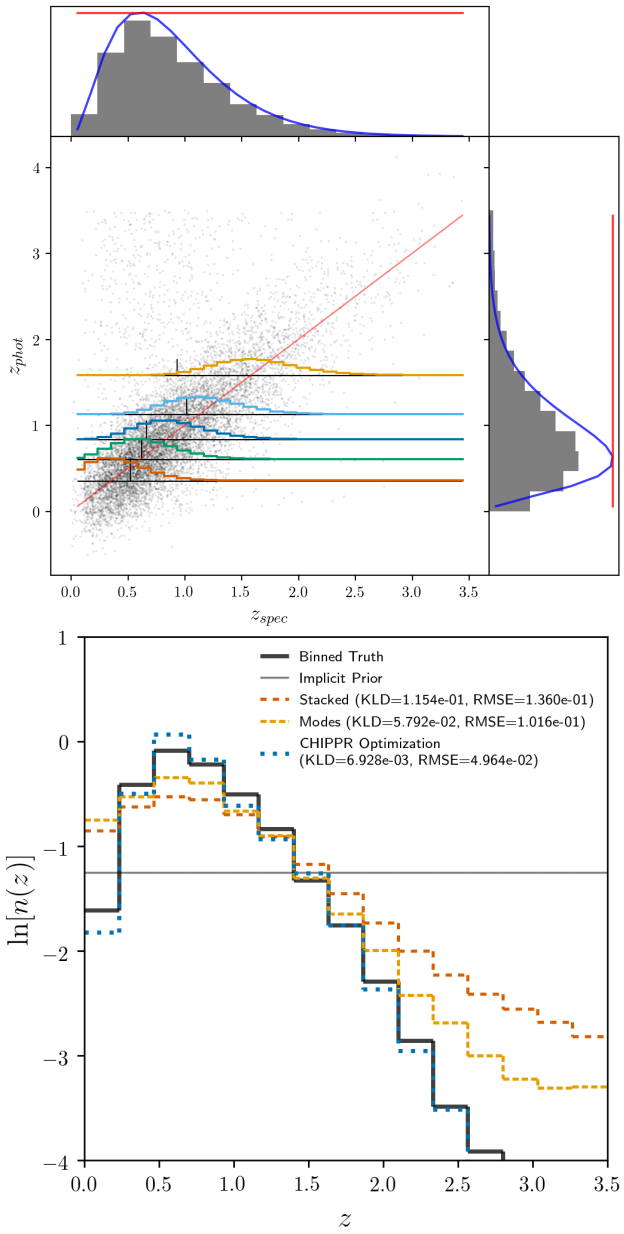

Figure 3. Using LSST numbers

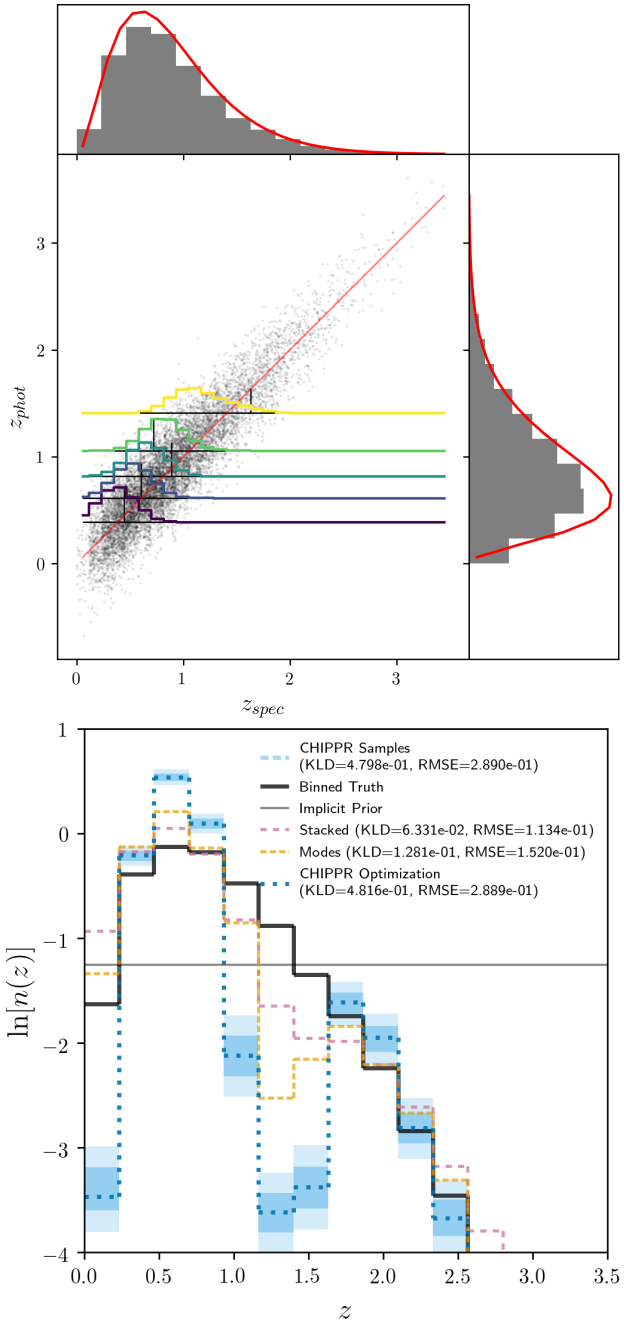

Figure 4. this is for wrong implicit prior, not bias, haven't done that one yet

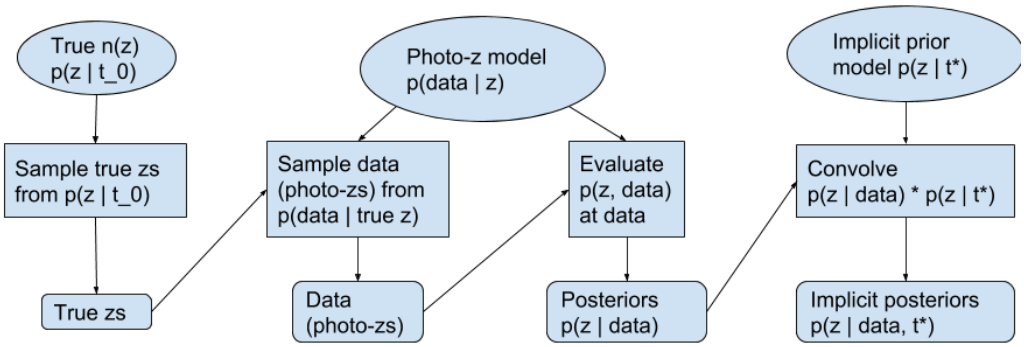

Figure 5. In appendix?