How to obtain the redshift distribution from probabilistic redshift estimates

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

aimalz@nyu.edu

ABSTRACT

The redshift distribution $n(z)$ is a crucial ingredient for weak lensing cosmology. Spectroscopically confirmed redshifts for the dim and numerous galaxies observed by weak lensing surveys are expected to be inaccessible, making photometric redshifts (photo-zs) the next best alternative. Because of the nontrivial inference involved in their determination, photo-z point estimates are being superseded by photo-z probability distribution functions (PDFs). However, analytic methods for utilizing these new data products in cosmological inference are still evolving.

This paper presents a novel approach to estimating the posterior distribution over $n(z)$ from a survey of galaxy photo-z PDFs based upon a probabilistic graphical model of hierarchical inference. We present the Cosmological Hierarchical Inference with Probabilistic Photometric Redshifts (CHIPPR) code implementing this technique, as well as its validation on mock data and testing on a subset of BOSS DR10. CHIPPR yields an accurate characterization of $n(z)$ containing information beyond the best-fit estimator produced by traditional procedures. The publicly available code is easily extensible to other one-point statistics that depend on redshift.

Subject headings: catalogs — cosmology: cosmological parameters — galaxies: statistics — gravitational lensing: weak — methods: analytical — methods: data analysis — methods: statistical — techniques: photometric

1. Introduction

The redshift distribution $n(z)$ is necessary for calculating two-point statistics of galaxy properties used to determine the cosmological parameter values that inform our understanding of the evolution of large-scale structure in the universe (Masters et al. 2015). Inaccurate estimates of

[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

$n(z)$ can significantly impact the constraining power of a galaxy survey, biasing recovery of the cosmological parameters. For example, if the $n(z)$ used in analysis is offset from the true $n(z)$ by a small, negative constant, it results in underestimating w_0 and overestimating σ_8 (Samuroff et al. 2017).

Though the redshift density function has traditionally been determined from spectroscopically observed redshifts, modern galaxy surveys including DES, LSST, Euclid, and WFIRST seek to obtain two-point statistics of redshift from samples of galaxies for which spectroscopic redshifts are unavailable, either due to their large numbers or their low brightnesses. For decades, photometrically estimated redshifts (photo-zs) have been the leading alternative to spectroscopically observed redshifts, though they suffer from issues of precision in the form of an intrinsic scatter and accuracy in the form of catastrophic outliers, as well as other systematics imparted by the properties of the survey, data reduction pipeline(s), and assumptions underlying the analysis (Baum 1962). These weaknesses are illuminated by a probabilistic interpretation of photo-zs; if these nontrivial uncertainties were expressed as a probability distribution function (PDF) over redshift, photo-zs could be replaced by photo-z PDFs containing more information than a simple point estimate (Koo 1999). Such data products have been commonly released by photometric galaxy surveys since SDSS DR7 (Abazajian & Survey 2009) using a great variety of methods.

Methods for using photo-z PDFs in cosmological inference remain underdeveloped, with many survey pipelines reducing them to familiar point estimators that are compatible with existing technology or engaging with them heuristically in a manner inconsistent with their probabilistic nature. Stacking photo-z PDFs, or other mathematically equivalent methods, to obtain an estimator of $n(z)$ is especially popular (Cunha et al. 2009; Sheldon et al. 2012). However, photo-z PDFs are probabilistic data products that must be handled in a fully consistent, probabilistic manner; hierarchical inference is the only mathematically valid way to do inference with photo-z PDFs and other probabilistic quantities.

It is desirable to create rigorous methods for using photo-z PDFs in inference, beginning with the simplest one-point statistic of redshift, $n(z)$. In the spirit of Hogg et al. (2010); Foreman-Mackey et al. (2014), this paper derives a mathematically rigorous approach to inferring $n(z)$ in Sec. 2. The presentation of a public code implementing this novel technique is given in Sec. **??**. The code is validated on mock data and tested on a subset of BOSS DR10 data in Sec. 3. We conclude and discuss future directions in Sec. 4.

2. Method

The redshift distribution $n(z)$ may be understood as the probability of finding a galaxy at a redshift z. We may express $n(z)$ in terms of some functional form $f_{\vec{\theta}}(z)$ with parameters comprising

$\vec{\theta}$ according to Eq. 1.

$$n(z) \equiv f_{\vec{\theta}}(z) \;\equiv\; p(z|\vec{\theta}) \tag{1}$$

In this work, we wish to characterize the posterior distribution $p(\vec{\theta}|\underline{D})$ of the parameters contained in $\vec{\theta}$ defining $n(z)$ given all available data $\underline{D}$. The likelihood $p(z|\vec{\theta})$ represents the probability of a random variable, in this case redshift z, given the parameters in $\vec{\theta}$ defining the distribution $n(z)$ from which it is drawn. Our model for inferring $p(\vec{\theta}|\underline{D})$ will be introduced in Sec. 2.1, and alternative methods of estimating $\vec{\theta}$ will be presented in Sec. 2.2.

2.1. Model Specifics

Inference of $p(\vec{\theta}|\underline{D})$ is perfectly suited to a hierarchical Bayesian model. In this problem, the available data $\underline{D}$ is a catalog of photometry $\{\vec{d}_i\}$ of individual galaxies i where each $\vec{d}_i$ is a random variable that is a function of its redshift parameter z_i; the redshift parameters $\{z_i\}$ are independently drawn from a function defined by hyperparameters in $\vec{\theta}$. These physical relationships may be illustrated in Fig. 1 by a directed acyclic graph representing a probabilistic graphical model (PGM). This PGM may be used as the basis for a derivation of the desired hyperposterior $p(\vec{\theta}|\{\vec{d}_i\})$, assuming our physical model is complete. From this point on, we will work solely with log probabilities.

We begin with Eq. 2 by applying Bayes' Rule to express the log-hyperposterior $\ln\Big[p(\vec{\theta}|\{\vec{d}_i\})\Big]$ in terms of the log-hyperlikelihood $\ln\Big[p(\{\vec{d}_i\}|\vec{\theta})\Big]$.

$$\ln\Big[p(\vec{\theta}|\{\vec{d}_i\})\Big] \propto \ln\Big[p(\vec{\theta})\Big] + \ln\Big[p(\{\vec{d}_i\}|\vec{\theta})\Big] \tag{2}$$

To do this, we will choose a log-hyperprior distribution $\ln\Big[p(\vec{\theta})\Big]$ representing our beliefs about the distribution of the hyperparameters $\vec{\theta}$ in terms of fixed variables that we will assume are known.

The log-hyperlikelihood $\ln\Big[p(\{\vec{d}_i\}|\vec{\theta})\Big]$ contains no explicit reference to redshifts, therefore, we employ the PGM to write the marginalization over the redshifts in Eq. 3.

$$\ln\Big[p(\{\vec{d}_i\}|\vec{\theta})\Big] = \ln\left[\int\; p(\{\vec{d}_i\}|\{z_i\})\; p(\{z_i\}|\vec{\theta})\; d\{z_i\}\right] \tag{3}$$

We shall handle the two terms in the integral separately.

The first term is the likelihood of all galaxy photometry given all galaxy redshifts. If we assume independence of galaxy photometry, such that $p(\vec{d}_i|\{\vec{d}_{i'\neq i}\},\{z_i\}) = p(\vec{d}_i|z_i)$, then we may write Eq. 4.

$$\ln\Big[p(\{\vec{d}_i\}|\{z_i\})\Big] = \sum_i \ln\Big[p(\vec{d}_i|z_i)\Big] \tag{4}$$

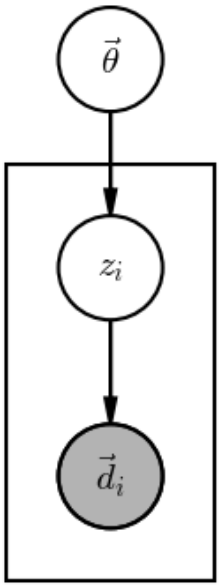

Fig. 1.— This directed acyclic graph corresponds to a PGM for a hierarchical inference of $p(\vec{\theta}|\{\vec{d}_i\})$. In this graph, all random variables are shown in circles, with observed variables shown in shaded circles. Relationships between variables are indicated by arrows. The box indicates that there are a number of copies of the relationships between boxed parameters, each independent of all others. The hyperparameters $\vec{\theta}$ representing $n(z)$ are at the top. Independently drawn from a function of the hyperparameters $\vec{\theta}$ are galaxy redshifts $\{z_i\}$ below. The observed galaxy photometry $\{\vec{d}_i\}$, shown in shaded circles, is determined by the redshifts above.

The second term may also be expanded as in Eq. 5 by assuming that all galaxy redshifts are drawn independently from $n(z)$.

$$\ln\left[p(\{z_i\}|\vec{\theta})\right] = \sum_i \ln\left[p(z_i|\vec{\theta})\right] \tag{5}$$

Since all galaxy redshifts are independent of one another, an integral over the aggregate $\int \ldots d\{z_i\}$ is simply the product of the integrals over each one $\prod_i \int \ldots dz_i$. Thus, the log-hyperlikelihood can be written according to Eq. 6.

$$\ln\left[p(\{\vec{d}_i\}|\vec{\theta})\right] = \sum_i \ln\left[\int\ p(\vec{d}_i|z_i)\ p(z_i|\vec{\theta})\ dz_i\right] \tag{6}$$

The expression of Eq. 6 contains two types of quantities. The $\{p(z_i|\vec{\theta})\}$ are known likelihoods that will be equal to $\{f_{\vec{\theta}}(z_i)\}$. The $\{p(\vec{d}_i|z_i)\}$, however, are likelihoods that are unknown and in general unknowable, and it is worth discussing these facts.

Photo-z PDFs are commonly written simply as $p(z)$, but this notation oversimplifies their substance. Whether determined by way of template-fitting or machine learning, photo-z PDFs are dependent on the data $\vec{d}$ from which they are calculated. This data $\vec{d}$ must be considered to be a quantity upon which the redshift z is conditioned, otherwise normalizing a photo-z PDF would require integrating over all possible values of $\vec{d}$. This means that photo-z PDFs are posteriors, probabilities of parameters z conditioned on data $\vec{d}$. However, where there is a posterior, there is always a prior. Photo-z PDFs are in general *interim* posteriors, because in addition to being conditioned on observations, they are also conditioned on an interim prior in the form of a particular value $\vec{\theta}^*$ of $\vec{\theta}$ necessary for computing the photo-z PDF. In the case of template-fitting methods, the interim prior is usually specified as an input to the photo-z PDF production code, and it often takes the form of an initial guess for $\vec{\theta}$ based on the results of previous galaxy surveys or simulations thereof; in the case of machine learning methods, the interim prior may be explicitly derived from the training set or implicitly produced in the process of determining the photo-z PDFs, and the interim prior is not always revealed to the user. Hierarchical inference may be performed only when the interim prior is known, which means it will only be possible for photo-z PDFs made via certain methods.

Since the data products from which we hope to probe the log-hyperposterior are themselves posteriors, we must express the likelihoods $\{p(\vec{d}_i|z_i)\}$ of Eq. 6 in terms of the photo-z PDFs $\{p(z_i|\vec{d}_i,\vec{\theta}^*)\}$. We start by multiplying the likelihood by an inspired factor of unity written in terms of the photo-z interim posterior we have at hand, as in Eq. 7.

$$p(\vec{d}_i|z_i) = p(\vec{d}_i|z_i)\ \frac{p(z_i|\vec{d}_i,\vec{\theta}^*)}{p(z_i|\vec{d}_i,\vec{\theta}^*)} \tag{7}$$

Next, we expand the denominator in terms of Bayes' Rule to obtain Eq. 8.

$$p(\vec{d}_i|z_i) = p(\vec{d}_i|z_i)\ p(z_i|\vec{d}_i,\vec{\theta}^*)\ \frac{p(\vec{d}_i|\vec{\theta}^*)}{p(z_i|\vec{\theta}^*)\ p(\vec{d}_i|z_i,\vec{\theta}^*)} \tag{8}$$

Because the redshift z_i is independent of the interim prior $\vec{\theta}^*$, the interim likelihood $p(\vec{d}_i|z_i, \vec{\theta}^*)$ may be expanded further as in Eq. 9.

$$p(\vec{d}_i|z_i) = p(\vec{d}_i|z_i)\ p(z_i|\vec{d}_i, \vec{\theta}^*)\ \frac{p(\vec{d}_i|\vec{\theta}^*)}{p(z_i|\vec{\theta}^*)\ p(\vec{d}_i|z_i)\ p(\vec{d}_i|\vec{\theta}^*)} \tag{9}$$

We cancel the terms $p(\vec{d}_i|z_i)$ and $p(\vec{d}_i|\vec{\theta}^*)$ that appear in both the numerator and denominator of Eq. 9 to obtain Eq. 10.

$$p(\vec{d}_i|z_i) = \frac{p(z_i|\vec{d}_i, \vec{\theta}^*)}{p(z_i|\vec{\theta}^*)} \tag{10}$$

Finally we may return to Eq. 2 to express the log-hyperposterior $\ln\left[p(\vec{\theta}|\{\vec{d}_i\})\right]$ in terms of the photo-z interim posteriors $\{p(z_i|\vec{d}_i, \vec{\theta}^*)\}$, as in Eq. 11.

$$\ln\left[p(\vec{\theta}|\{\vec{d}_i\})\right] \propto \ln\left[p(\vec{\theta})\right] + \sum_i \ln\left[\int\ p(z_i|\vec{d}_i, \vec{\theta}^*)\ \frac{p(z_i|\vec{\theta})}{p(z_i|\vec{\theta}^*)}\ dz_i\right] \tag{11}$$

Now we have an expression for a quantity proportional to the log-hyperposterior we desire! Though the constant of proportionality, $p(\{\vec{d}_i\})$ is in general not possible to calculate, we may still characterize the log-hyperposterior by way of MCMC sampling.

Several assumptions were noted in the above derivation; for clarity, they will be enumerated below, where their limitations will be discussed.

1. The PGM is an expression of our beliefs about the physics of the problem, and the inference will only be valid to the degree that the model is correct. One can easily think of ways in which the PGM of Fig. 1 is incomplete; for example, intrinsic galaxy SEDs will evolve with redshift. However, we assume that physical processes not represented in the PGM of Fig. 1 are subdominant and may thus be neglected.

2. We must choose a hyperprior $p(\vec{\theta})$; we assume that our choice of a sufficiently general hyperprior that will not be a dominant source of information in the hyperposterior is successful. If we choose poorly, the hyperprior may dominate over the hyperlikelihood, downweighting the significance of the data in determining the hyperposterior or biasing the result based on a misunderstanding of the underlying physics.

3. We assume independence of galaxies in our catalog such that each galaxy's photometry $\vec{d}_i$ is independent from all other galaxies' photometry $\{\vec{d}_{i'\neq i}\}$ and all other galaxies' redshifts $\{z_{i'\neq i}\}$. However, the photometry $\{\vec{d}_i\}$ will inherently share instrumental and systematic effects due to being observed with the same telescope as part of a single survey project. Furthermore, there may be blended objects or other correlations between the redshifts of different galaxies. We must assume that such effects are negligible and will not consider them in this treatment.

4. We assume that the interim prior is known to us, either as input to or output of the method producing the photo-z interim posteriors. Furthermore, it must be representable in the parametrization of $f_{\vec{\theta}}(z)$, i.e. our model $f_{\vec{\theta}}(z)$ for $n(z)$ must be flexible enough to encompass both the true distribution we aim to probe and the interim prior.

5. Finally, we assume that the photo-z interim posteriors $\{p(z_i|\vec{d}_i, \vec{\theta}^*)\}$ are accurate. There is some evidence that photo-z PDFs derived by different methods disagree with one another or are strongly dependent on the interim prior. However, the choice of the technique by which the photo-z interim posteriors are determined is outside the scope of this paper.

2.2. Alternative methods

3. Experiments

4. Discussion

AIM thanks Phil Marshall for advising on the production of usable code, Mohammadjavad Vakili for insightful input on statistics, Geoffrey Ryan for assistance in debugging code, and Boris Leistedt for helpful comments provided in the preparation of this paper.

REFERENCES

Abazajian, K., & Survey, f. t. S. D. S. 2009, The Astrophysical Journal Supplement Series, 182, 543, arXiv: 0812.0649

Baum, W. A. 1962, Proc. from IAU Symp.

Cunha, C. E., Lima, M., Oyaizu, H., Frieman, J., & Lin, H. 2009, Mon. Not. R. Astron. Soc., 396, 2379

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, Astrophys. J., 795, 64

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

Koo, D. C. 1999, in Photometric Redshifts and the Detection of High Redshift Galaxies, Vol. 191, eprint: arXiv:astro-ph/9907273, 3

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

Samuroff, S., Troxel, M. A., Bridle, S. L., et al. 2017, Monthly Notices of the Royal Astronomical Society: Letters, 465, L20

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