Draft version June 20, 2017
Preprint typeset using LaTeX style emulateapj v. 12/16/11

HOW TO OBTAIN THE REDSHIFT DISTRIBUTION FROM PROBABILISTIC REDSHIFT ESTIMATES

Alex Malz[1], David W. Hogg[1,2,3,4], Phil Marshall[5], & Others

Draft version June 20, 2017

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 Because of the nontrivial inference involved in their determination, photo-z point estimates are being superseded by photo-z probability distribution functions (PDFs).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 the `Buzzard` simulations. `CHIPPR` yields a more 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.

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

1. INTRODUCTION

After a brief literature review addressing how photo-z PDFs are currently used in cosmology, this paper aims to answer the following questions:

- Why should we question existing methods?
- 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 $n(z)$?
- How significant is the effect of the discrepancy between $n(z)$ estimators on cosmological constraints?

2. METHOD

This paper presents a mathematically consistent method for obtaining the posterior distribution over the redshift density function $n(z)$ using a catalog of photo-z PDFs.

2.1. *Model*

The directed acyclic graph of Fig. 1 represents a probabilistic graphical model for hierarchical inference of $n(z)$, the mathematical interpretation of which will be presented below.

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
[5] [SLAC]

Math from `prob-z` version will go here.

This framework entails a number of choices and assumptions that must be addressed explicitly.

1. While we advocate for the approach of hierarchical inference, the probabilistic graphical model presented here is not the only one that could be proposed.

2.2. *Alternative Approaches*

It is useful to translate some popular existing methods for deriving $n(z)$ from photo-z PDFs into the mathematical framework of Sec. 2.1.

2.3. *Implementation*

The publicly available `CHIPPR` code implements the probabilistic graphical model presented in Sec. 2.1.

In addition to the choices and assumptions underlying the probabilistic graphical model, the implementation of `CHIPPR` makes choices and assumptions of its own.

1. `CHIPPR` currently only accepts photo-z PDFs and produces $n(z)$ samples of a single format, that of the piecewise constant parametrization, also referred to as a binned histogram parametrization and a sum of top hat functions.

3. VALIDATION ON SIMPLE MOCK DATA

We demonstrate the superiority of `CHIPPR` over alternative approaches in a number of compelling test cases on mock data. Each experiment is characterized by a single change to a fiducial case in order to isolate the influence of systematic effects known to be relevant to photo-z estimation and propagation in analysis.

3.1. *Mock Data & Metrics*

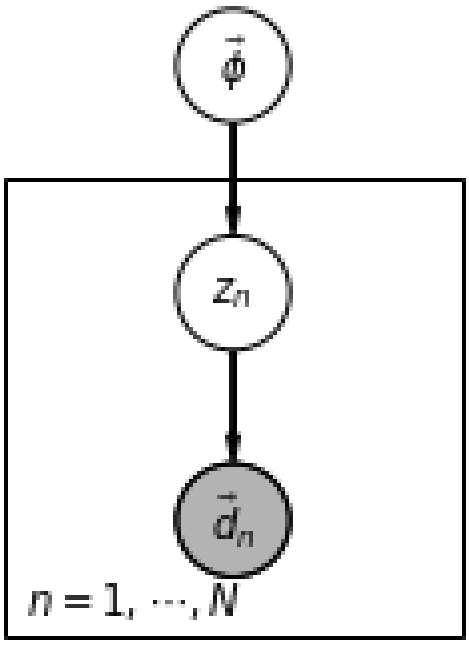

Figure 1. This directed acyclic graph corresponds to a PGM for a hierarchical inference of $p(\vec{\theta}|\{\vec{d}_i\})$. In this graph, all random variables are shown in circles, with observed random variables shown in shaded circles. Relationships between variables are indicated by arrows from parameters to the variables distributed according to functions of them. 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 mock data in these tests consists of photo-z interim posteriors rather than photometric data because the various existing methods for deriving photo-z interim posteriors do not in general yield results that are consistent with one another, indicating that their systematics are not well-understood. Because the mock data is independent of any choice of photo-z PDF production method, we not only ensure that our photo-z interim priors are perfectly understood but also deter readers from assuming that `CHIPPR` has any preference over the method by which photo-z interim posteriors are derived from photometric data.

3.1.1. *Mock Data*

The mock data used here are produced by the following steps.

1. Choose a true $n(z)$ that is a continuous function with known parameters $\vec{\theta}$.
2. Sample the true $n(z)$ to create a catalog of N true redshifts z_i.
3. Choose a true intrinsic scatter σ, and sample Gaussians $p(z_i'|z_i, \sigma) = \mathcal{N}(z_i, \sigma)$ to get "observed" redshifts z_i' defining Gaussian likelihoods $p(\vec{d}_i|z_i', \sigma) = \mathcal{N}(z_i', \sigma)$.
4. Choose a parametrization and the parameters $\vec{\phi}^*$ of the interim prior.
5. Create interim posteriors $p(z_i|\vec{d}_i, \vec{\phi}^*) = p(\vec{d}_i|z_i)\ p(z_i|\vec{\phi}^*)$ in this parametrization.

In all of the following validation tests, we use 10 bins and $N = 10,000$ galaxies. In Secs. 3.2 and 3.3, we use a flat interim prior. This method for deriving mock data is referred to as the fiducial case, and variations on it will refer directly to the steps that are altered.

3.1.2. *$n(z)$ Accuracy Metric*

The Kullback-Leibler divergence is our primary measure of the accuracy of estimators of $n(z)$ in cases of mock data with known true redshifts.

Review precision and bias from `kld.ipynb` and interpret in terms of a % difference.

3.2. *Underlying $n(z)$ Effects*

Existing $n(z)$ estimators are systematically smoother than the true $n(z)$. Here we show that the traditional estimators perform better when the true $n(z)$ is weakly featured than when the true $n(z)$ is strongly featured by experimenting with Step 1 of Sec. 3.1.1. The implication of this issue is quite serious; the consistently smooth, unimodal $n(z)$ estimates appearing in the literature could result from much more featured true redshift distributions, and there would be no way to catch this error without using a fully probabilistic method.

3.2.1. *Featureless $n(z)$*

In this test, we choose a weakly featured true $n(z)$ of Fig. **??** with a smooth, unimodal shape, based on the interim prior used for the SDSS DR8 photo-z PDFs. [check and include citation!]

3.2.2. *Featured $n(z)$*

In this test, we choose the true $n(z)$ of Fig. 3 with nontrivial structure.

This featured true $n(z)$ will henceforth be referred to as the fiducial $n(z)$

3.3. *Emulated Data Quality Effects*

In the following test cases, we vary the properties of the mock photo-z likelihoods in an effort to emulate known systematics in photo-z estimation. These tests vary Step 3 of Sec. 3.1.1.

3.3.1. *Intrinsic Scatter*

One major concern about photo-zs is the intrinsic scatter of point estimators, including those derived from photo-z PDFs, that is observed to varying extents with every existing photo-z algorithm and illustrated in Fig.

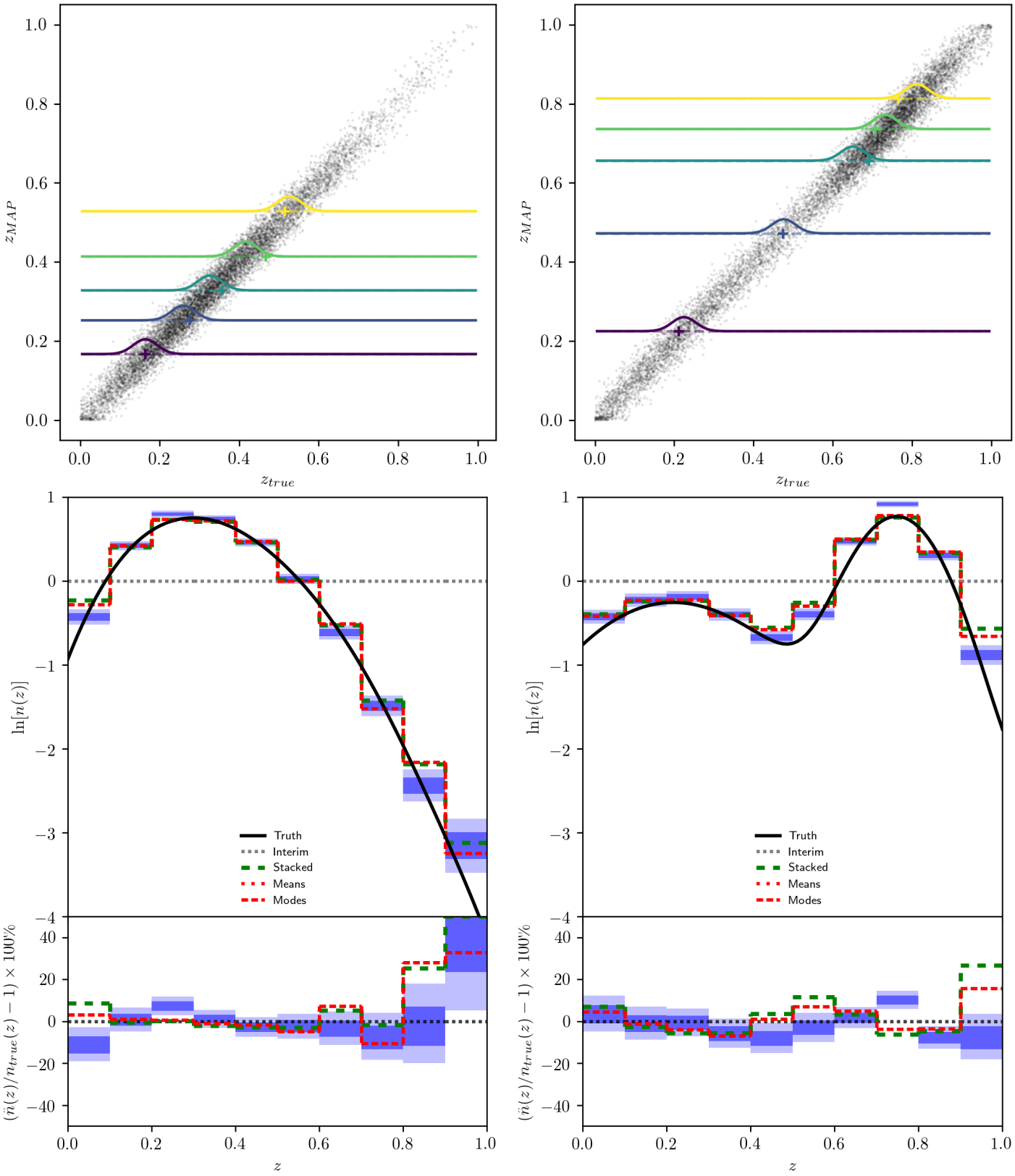

Figure 2. All estimators perform well when the true $n(z)$ is well behaved, exhibiting significant deviation only when $n(z)$ is very small, as the sample size of true redshifts in that range will be small. Top panel: The traditional MAP reduction of a photo-z PDF against the true redshifts with a few rescaled photo-z interim posteriors are overplotted in solid lines, with a dotted line indicating zero probability. Bottom panel: Various estimators of $\ln[n(z)]$, the interim prior, and the true $\ln[n(z)]$ as a continuous function and under a binned parametrization.

Figure 3. [This plot doesn't really show what I want because the intrinsic scatter is too low! I used $\sigma = 0.03$ because that's what has been quoted as what LSST, etc. needs, but that's not what comes out of photo-z estimation methods before they impose aggressive cuts. . .] Top panel: The traditional MAP reduction of a photo-z PDF against the true redshifts with a few rescaled photo-z interim posteriors are overplotted in solid lines, with a dotted line indicating zero probability. Bottom panel: Various estimators of $\ln[n(z)]$, the interim prior, and the true $\ln[n(z)]$ as a continuous function and under a binned parametrization.

4. To emulate intrinsic scatter, we modify the fiducial case to simply broaden the single Gaussian component of the likelihood. To enforce self-consistency, the mean is a random variable drawn from a Gaussian distribution with the newly increased variance.

As the intrinsic scatter increases, the discrepancy between estimators increases.

3.3.2. *Template-like Catastrophic Outliers*

In addition to intrinsic scatter, photo-z methods employing template fitting tend to produce catastrophic outliers that are distributed to be broad in z_{spec} and narrow in z_{phot}, as in Fig. 5. The systematic behind these catastrophic outliers may be described as an attractor in the space of z_{phot}; some galaxies at a range of z_{spec} map onto a single z_{phot} (with some scatter) if their true SED does not have sufficiently strong features (as is the case for blue galaxies), leading galaxies of that SED type at many z_{spec} to have similar colors.

3.3.3. *Training-like Catastrophic Outliers*

Data driven photo-z methods tend to suffer from a different form of catastrophic outliers that are distributed to be narrow in z_{spec} and broad in z_{phot}, as in Fig. 6. The systematic behind these catastrophic outliers may be described as an attractor in the space of z_{spec}; some galaxies near a particular z_{spec} map to a range of z_{phot} if the training set galaxies at that z_{spec} have inconsistent z_{phot}, as might occur if their SED's features fall between photometric filters.

3.4. *Emulated Interim Prior Effects*

The interim prior encapsulates the the relationship between observed photometry and redshift information upon which a photo-z estimate is based. Interim priors are in general not identical to the true $n(z)$ we wish to estimate; if they were, we would not need any data! For template fitting photo-z methods, the interim prior is usually an input chosen by the researcher. However, for machine learning methods, the interim prior is some function of the training set data that in many cases may be influenced by random numbers and is rarely output with the redshift estimates. Interim priors for template fitting methods tend to have incomplete coverage in the space of true photometry, because they are limited by the choice of the library of SEDs. Interim priors for machine learning methods tend to have incomplete coverage in the space of redshifts, because there are fewer galaxies with spectroscopically confirmed redshifts at high redshifts than low redshifts.

Existing $n(z)$ estimation routines will always produce a biased estimator when the interim prior is not equal to the true $n(z)$. We demonstrate here that regardless of the appropriateness of the interim prior as an approximation to the true $n(z)$, `CHIPPR` is not affected by the choice of the interim prior so long as it has nontrivial coverage in the space of redshift. These tests modify Step 4 of Sec. 3.1.1.

3.4.1. *Template-like Interim Prior*

An interim prior based on a template library may be a sum of smooth functions representing $n(z)$ for each SED type in the library. Template libraries do not include every possible galaxy SED, and the $n(z)$ used for each SED

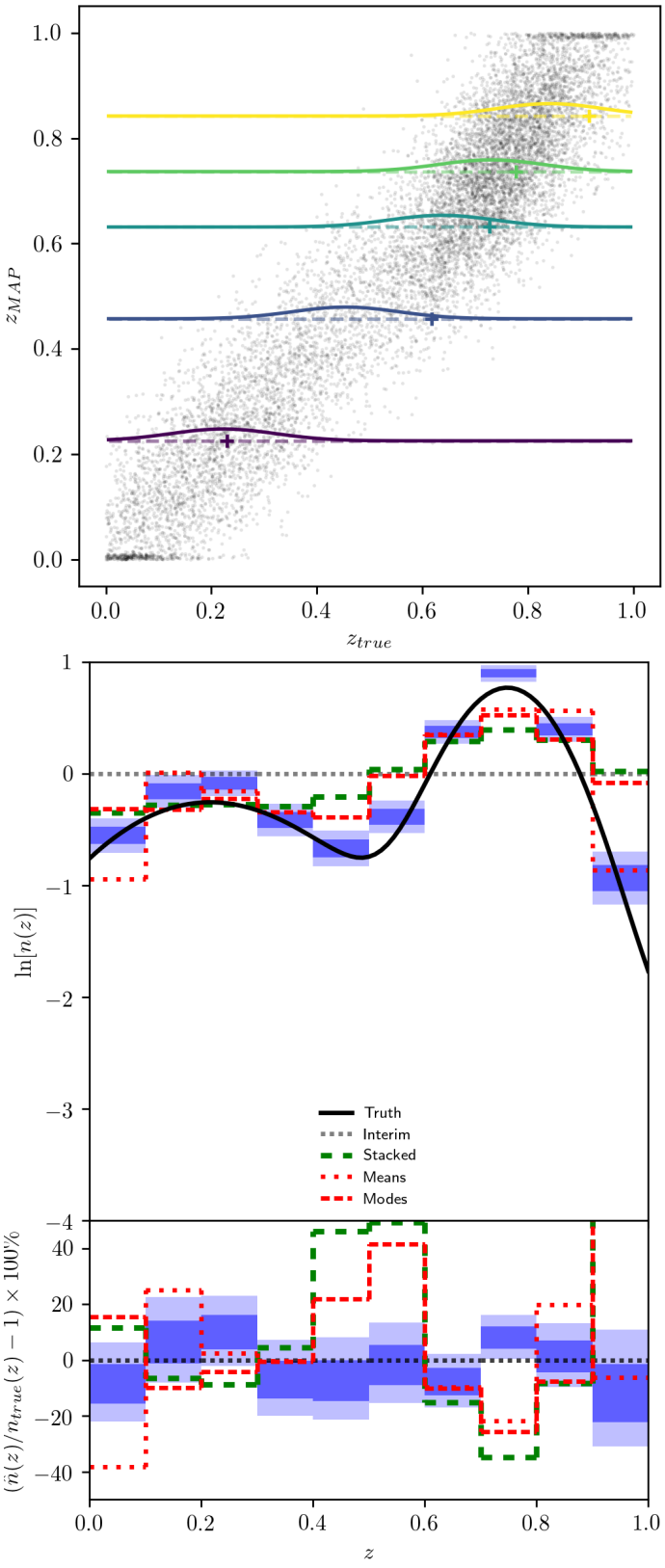

Figure 4. As the intrinsic scatter increases, the discrepancy between estimators increases. In particular, the stacked estimator and marginalized point estimators predict $\ln[n(z)]$ to be smoother than the truth, while the [This would be a lot more compelling with more galaxies. Also, the weird edge effects in the top panel are real, because the point estimator is the MAP, not the center of the Gaussian likelihood, and there's no requirement that the mean of the likelihood be within the true redshift range.] Top panel: The traditional MAP reduction of a photo-z PDF against the true redshifts with a few rescaled photo-z interim posteriors are overplotted in solid lines, with a dotted line indicating zero probability. Bottom panel: Various estimators of $\ln[n(z)]$, the interim prior, and the true $\ln[n(z)]$ as a continuous function and under a binned parametrization.

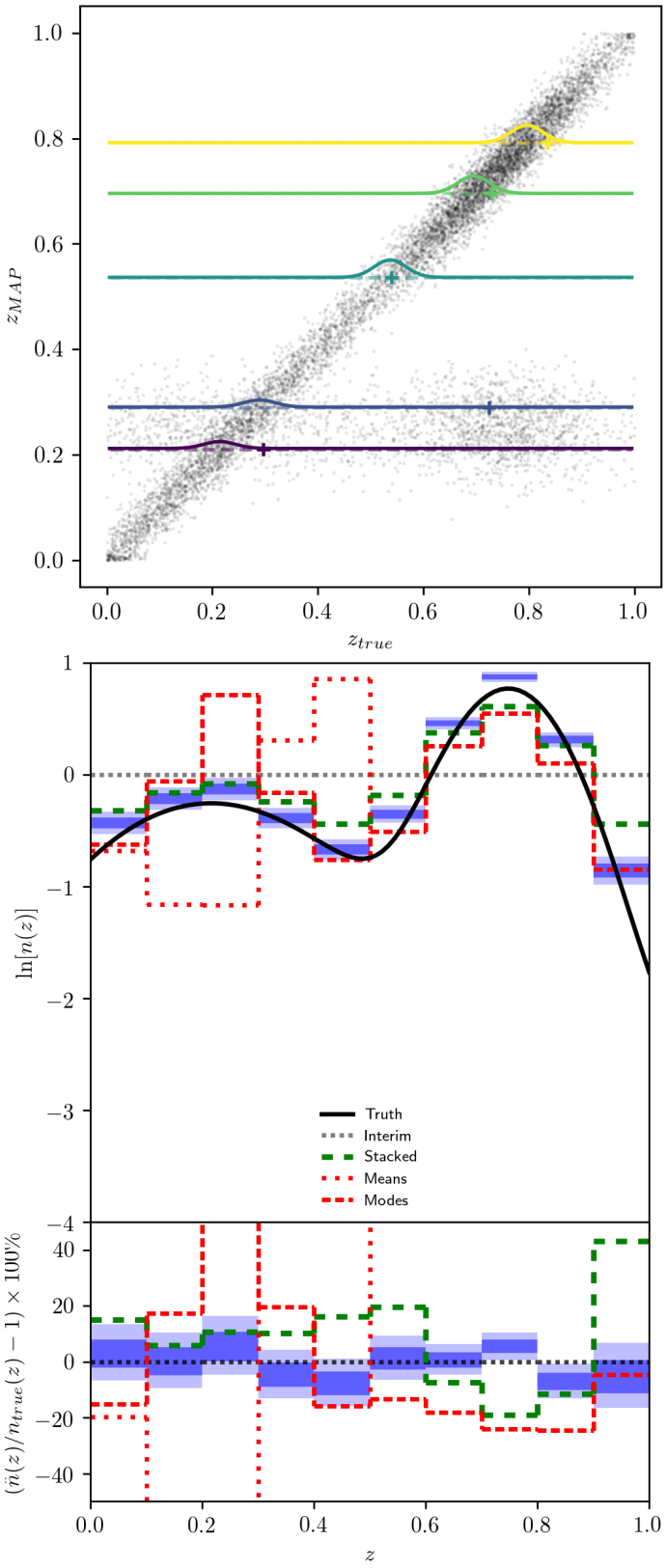

Figure 5. Top panel: The traditional MAP reduction of a photo-z PDF against the true redshifts with a few rescaled photo-z interim posteriors are overplotted in solid lines, with a dotted line indicating zero probability. Bottom panel: Various estimators of $\ln[n(z)]$, the interim prior, and the true $\ln[n(z)]$ as a continuous function and under a binned parametrization.

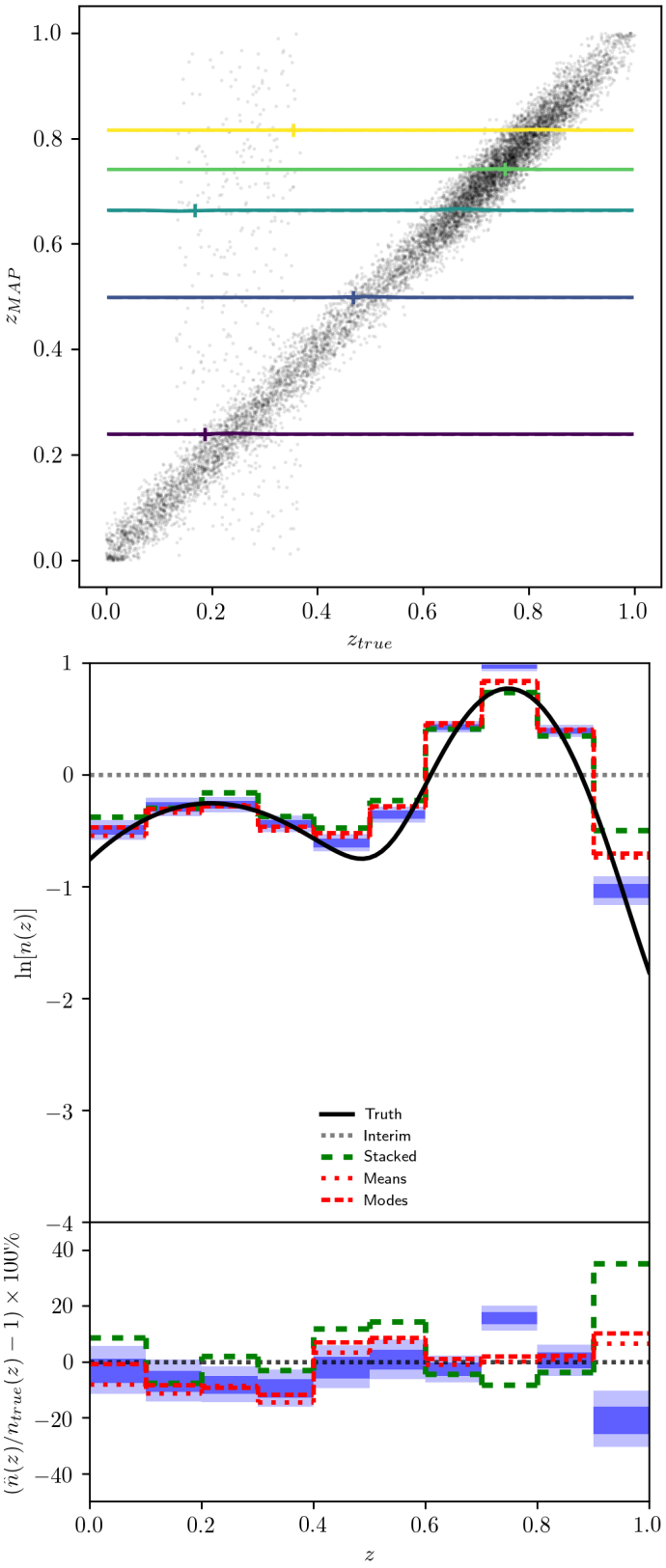

Figure 6. [This has proven to be the most challenging test case to implement, and there's clearly still a bug in the function that makes the likelihoods.]

type may not be accurate. The interim prior shown in Fig. 7 is an emulation of an interim prior corresponding to a template library of this type.

3.4.2. *Training-like Interim Prior*

An interim prior based on a training set may be biased toward low redshifts due to the dearth of distant galaxies with spectroscopic redshifts. The interim prior shown in Fig. 8 is an emulation of an interim prior corresponding to a training set biased in this way. We chose an interpolation of the interim prior used for the SDSS DR8 photo-z PDFs.

4. APPLICATION TO REALISTIC MOCK DATA

To show how the choice of $n(z)$ estimator propagates to cosmological constraints, we apply `CHIPPR` to a data from a realistic cosmological simulation (probably `Buzzard`) with photo-z PDFs produced by a popular method (probably `BPZ`).

4.1. *Mock Data & Metrics*

As in Sec. 3.1, the mock data takes the form of photo-z interim posteriors, but that is where the similarity ends. These photo-z interim posteriors are derived from the photometry resulting from the `Buzzard` simulation by way of `BPZ`. Because the simulation begins with setting true values of the cosmological parameters, we can propagate the different estimators of $n(z)$ through a forecasting code (which one?) to generate error ellipses on the cosmological parameters.

4.1.1. *Mock Data*

We summarize the details of the `Buzzard` simulation here.

4.1.2. *Cosmological Constraint Metric*

Because the mock data is associated with true values of the cosmological parameters, we may compare the quality of cosmological constraints.

Check Dark Energy Task Force metrics on error ellipses, multi-dimensional KLD, centroid offset, etc.

4.2. *Results*

5. DISCUSSION

Thee results of Fig 9 have significant implications for the developing data analysis pipelines of next-generation telescope surveys. However, the method presented here has its own limitations, which are reiterated to discourage the community from applying this work inappropriately.

We intend to pursue a number of extensions of the work presented in this paper in future work.

6. CONCLUSION

We now summarize answers to the questions posed in the introduction:

- Existing $n(z)$ estimation methods produce biased estimators that propagate to inaccuracies in characterizing the cosmological parameters.

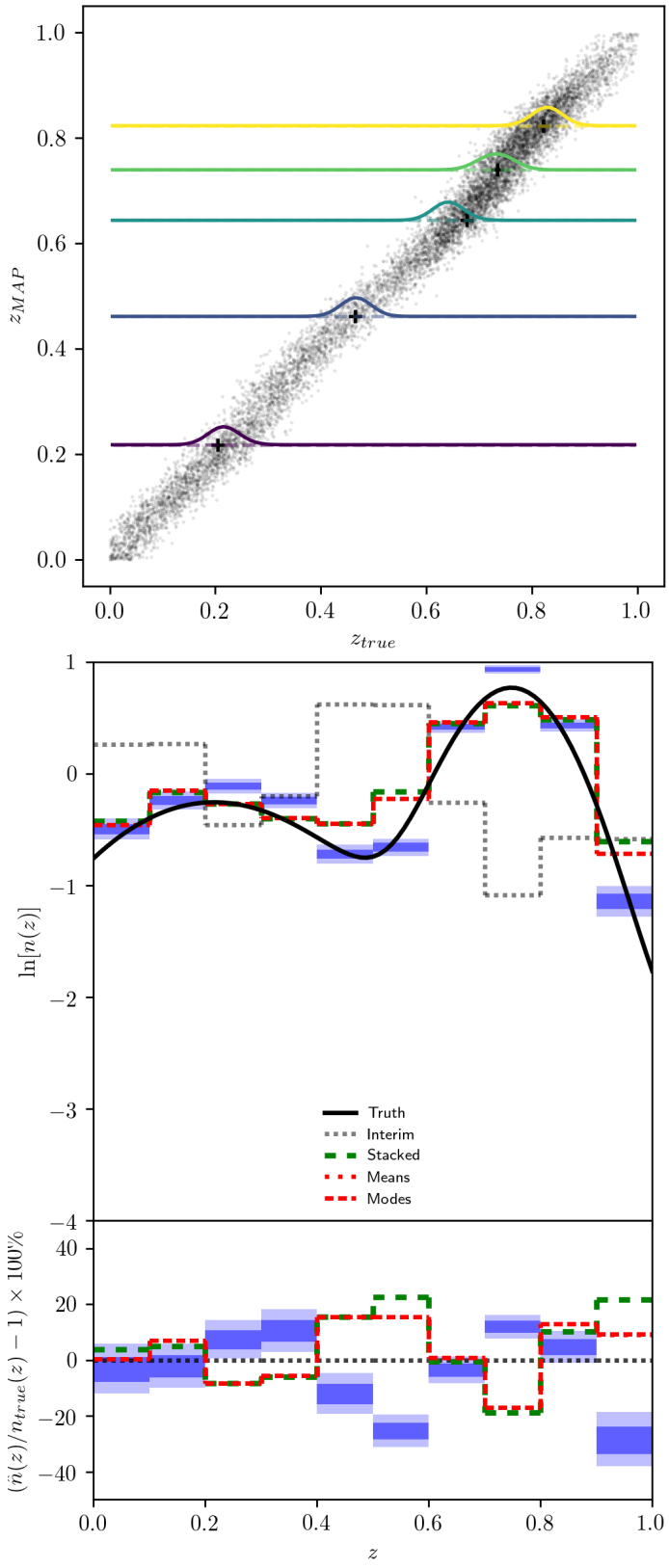

Figure 7. [The case of a multimodal interim prior was a very compelling test in the previous version but somehow isn't anymore.]

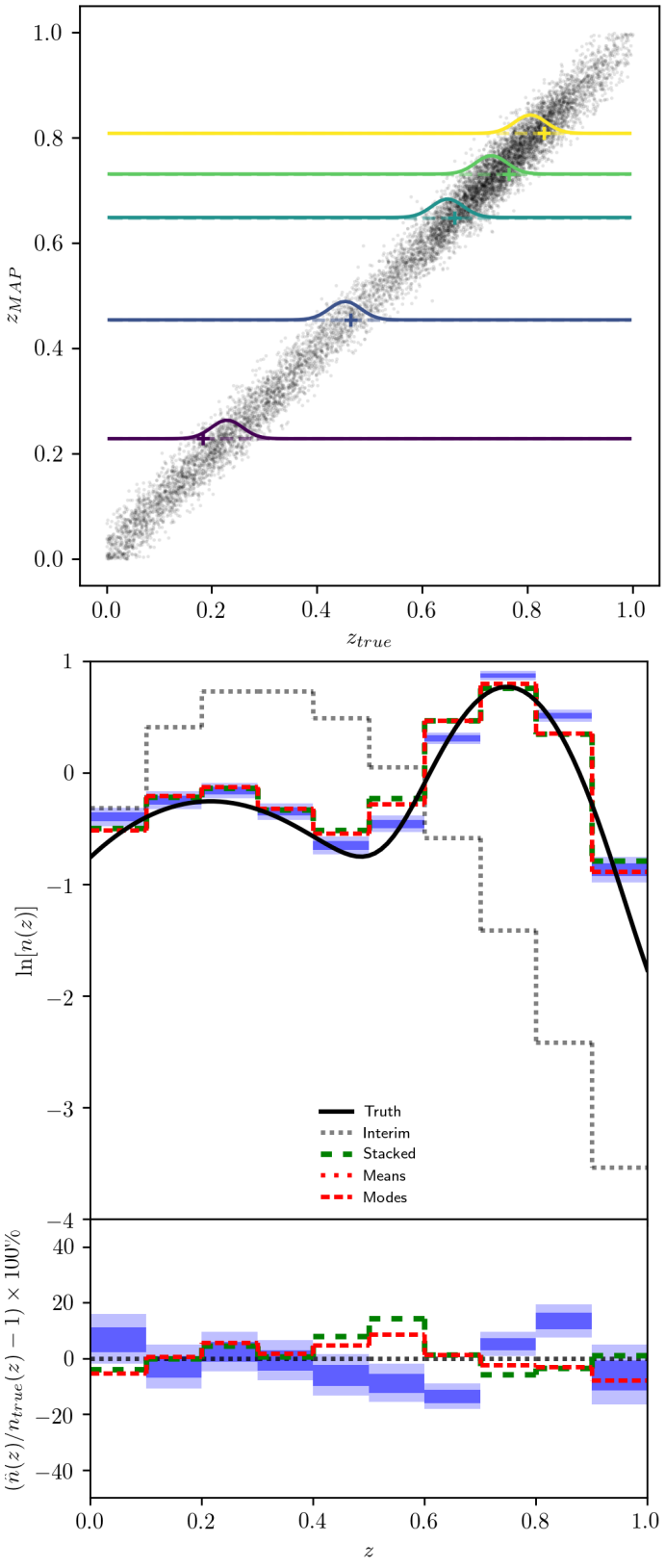

Figure 8. [The case of a low-z favoring interim prior was a very compelling test in the previous version but somehow isn't anymore.]

Figure 9. [moneyplot of error ellipses resulting from different $n(z)$ estimators]

- Photo-z PDFs are probabilistic data products so must be handled in a mathematically consistent manner such as the probabilistic graphical model outlined in this paper.
- In addition to coming with its own error distribution, the $n(z)$ estimator produced by `CHIPPR` is quantifiably more accurate than established estimators.
- Propagation of the `CHIPPR` result leads to a quantifiable improvement in the constraints on cosmological parameters.

In conclusion, we discourage the community from continuing to use the stacked estimator and reductions of photo-z PDFs to redshift point estimates in obtaining estimators of $n(z)$. `CHIPPR` is freely available to the community for incorporation into evolving data analysis pipelines.

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