Draft version July 7, 2017
Preprint typeset using LaTeX style emulateapj v. 12/16/11

SUPERNOVA COSMOLOGY INFERENCE WITH PROBABILISTIC PHOTOMETRIC REDSHIFTS

Alex Malz[1], Tina Peters[2], Humna Awan, Anita Bahmanyar[2], Lluis Galbany, Renee Hlozek[2], Boris Leistedt[1], and Kara Ponder

Draft version July 7, 2017

ABSTRACT

The BEAMS framework employs probabilistic supernova classifications to estimate the Hubble parameter that quantifies the relationship between distance and redshift over cosmic time. This work extends BEAMS to replace high-confidence spectroscopic redshifts with probabilistic photometric redshifts, enabling inference of the Hubble parameter as a function of two probabilistic variables. By combining posterior probabilities of supernova type and posterior probabilities of host galaxy redshift, we infer a posterior probability distribution over the redshift-dependent Hubble parameter. This work also produces the `scipprcode` that can be used to infer cosmological parameters from probabilistic supernova fit parameters.

1. INTRODUCTION

Kunz *et al.* (2007); Kelly *et al.* (2008); Hlozek *et al.* (2012)

- LSST era of photometric-only lightcurves and unreliable redshift point estimates
- BEAMS and type probabilities
- photo-z PDFs

The problem at hand is to infer the cosmological parameters and redshift-dependent supernova type proportions from purely photometric data in the form of supernova lightcurves and host galaxy photometry.

2. METHOD

This work demonstrates a principled approach to inferring cosmological parameters and redshift-dependent supernova type proportions from purely photometric lightcurves of supernovae and host galaxy photometry. First, we outline the notation used throughout this paper.

Let us consider observations of N supernovae n, each with a photometric lightcurve $\underline{\ell}_n$ and host galaxy photometry $\vec{f}_n$ comprised of fluxes, magnitudes, or colors. When we conduct a photometric survey, not all supernovae in the universe will make it into our sample due to selection effects; we cannot include observations we did not make, nor do we include those with low confidence or insufficiently complete data. These choices are quantified by nuisance parameters $\vec{\alpha}$ and $\vec{\beta}$ restricting the supernova lightcurves and host galaxy photometry respectively.

We believe that every supernova has several intrinsic parameters that are not directly observable: a type t_n, a redshift z_n, and a distance modulus μ_n. Traditional methods estimate the values of these parameters from observed lightcurves and host galaxy photometry by assuming that the lightcurves and host galaxy photometry are random variables drawn from distributions that are functions of these unobservable parameters.

The unobservable parameters, however, are themselves random variables, with type being discrete and redshift and distance modulus being continuous, drawn from distributions that are functions of some hyperparameters. When we use types, redshifts, and distance moduli to constrain the cosmological parameters and/or redshift-dependent type proportions, we make this very assumption. As in traditional approaches, we seek to constrain the cosmological parameters comprising $\vec{\Omega}$, among them the Hubble parameter H_0 and matter density Ω_0, and the redshift-dependent type proportions parametrized by $\underline{\phi}$ under some parametrization that need not be specified at this time. The redshift-dependent type proportions can be thought of as probabilities over type and redshift $p(t, z|\underline{\phi})$.

2.1. *Model*

We introduce a hierarchical Bayesian model for inferring the hyperparameters without estimating the unobservable parameters. The directed acyclic graph of Fig. 1 shows the relationships between the variables discussed above. We assume that the parameters associated with supernova n and its host galaxy are statistically independent of the parameters associated with supernova n' and its host galaxy.

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] Dunlap Institute & Department of Astronomy and Astrophysics, University of Toronto, 50 St George Street, Toronto, ON M5S 3H4 Canada

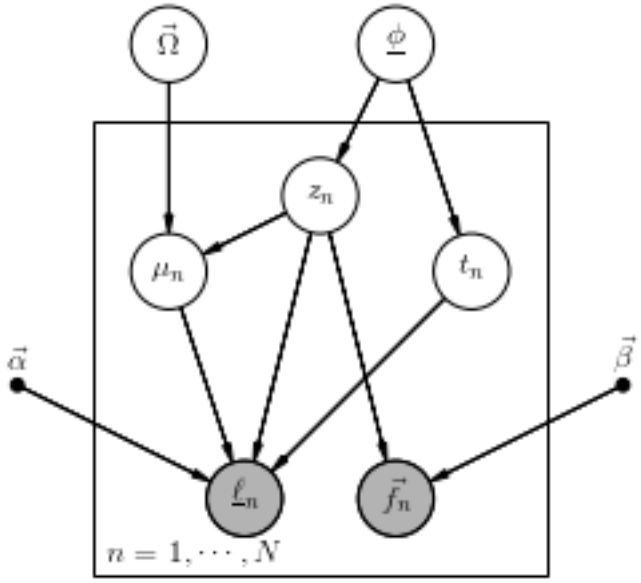

FIG. 1.— This directed acyclic graph corresponds to a probabilistic graphical model for our hierarchical inference of the cosmological parameters and redshift-dependent type proportion parameters. In this graph, all random variables are shown in circles, with observed variables shown in shaded circles. The box indicates that there are N copies of the relationships between boxed parameters, each statistically independent of all others. The hyperparameters we would like to infer are the cosmological parameters in $\vec{\Omega}$ and the redshift-dependent type proportion parameters comprising $\underline{\phi}$. Drawn from functions of these hyperparameters are the distance moduli $\{\mu_n\}_N$, redshifts $\{z_n\}_N$, and supernova types $\{t_n\}_N$. Here, we observe host galaxy colors $\{\vec{f}_n\}_N$ and multi-band supernova lightcurves $\{\underline{\ell}_n\}_N$, shown in shaded circles. The solid dots indicate known constants that factor into the model; $\vec{\alpha}$ represents the parameters defining a selection function in the space of observed lightcurves, and $\vec{\beta}$ includes the parameters defining a selection function in the space of host galaxy photometry. The arrows encode the relationships between variables, going from parameters defining probability distributions to variables drawn from those probability distributions.

The motivation for this approach is the existence of photo-z PDFs and the anticipation of PDFs over lightcurve fit parameters. Photo-z PDFs are posterior probability distributions; we observe the host galaxy photometry $\vec{f}_n$ and learn something about the host galaxy redshift z_n. The process by which we derive a relationship between the observed and unobserved parameters imprints its biases in the form of an interim prior that defines a global probability distribution over redshifts parametrized by $\vec{\theta}$. The selection function also biases the posterior probability in the form of an interim prior parametrized by $\vec{\beta}$. Thus photo-z PDF is an interim posterior probability distribution $p(z_n|\vec{f}_n, \vec{\theta}, \vec{\beta})$.

We anticipate the production of lightcurve parameter PDFs, which will also be interim posterior probability distributions, but in a higher dimensional space. As in the case of photo-z PDFs, the observed multi-band lightcurves $\{\vec{\ell}_n\}_N$ inform us about the latent parameters of supernova types $\{t_n\}_N$, redshifts $\{z_n\}_N$, and distance moduli $\{\mu_n\}_N$. The interim prior will be a probability distribution over t, z, and μ with known parameters comprising $\underline{\xi}$. We will also have a lightcurve selection function of parameters $\vec{\alpha}$ that is another distribution over this three-dimensional space. The overall interim posterior of the supernova lightcurve is then $p(t_n, z_n, \mu_n|\underline{\ell}_n, \underline{\xi}, \vec{\alpha})$.

(@aimalz Discuss interim priors and selection functions in more detail here, or somewhere else?)

The goal here is to constrain the posterior distribution $p(\vec{\Omega}, \underline{\phi}|\{\underline{\ell}_n\}_N, \{\vec{f}_n\}_N)$ of the physically important hyperparameters in $\vec{\Omega}$ and $\underline{\phi}$ given a catalog of N interim posterior distributions $p(t_n, z_n, \mu_n|\underline{\ell}_n, \underline{\xi}, \vec{\alpha})$ and another N interim posterior distributions $p(z_n|\vec{f}_n, \vec{\theta}, \vec{\beta})$. We will derive an expression for the posterior distribution over hyperparameters that we want to find in terms of the interim posteriors over latent parameters that are available to us.

We begin by expanding this in terms of Bayes' Rule.

$$p(\vec{\Omega}, \underline{\phi}|\{\underline{\ell}_n\}_N, \{\vec{f}_n\}_N) \propto p(\vec{\Omega}, \underline{\phi})\ p(\{\underline{\ell}_n\}_N, \{\vec{f}_n\}_N|\vec{\Omega}, \underline{\phi}) \tag{1}$$

In plain English we are saying that the posterior probability of a certain cosmology and redshift-dependent type proportions given some set of observed supernova lightcurves and host galaxy photometry is proportional to the product of a prior belief about the cosomological parameters and redshift-dependent type proportions and the likelihood of the lightcurves and host galaxy photometry given the cosmology and redshift-dependent type proportions.

Next, we invoke the statistical independence of the supernova parameters and observations; each set of $(t_{n'}, z_{n'}, \mu_{n'}, \underline{\ell}_{n'}, \vec{f}_{n'})$

is independent from all other parameters in $\bigcup_n (t_{n\neq n'}, z_{n\neq n'}, \mu_{n\neq n'}, \underline{\ell}_{n\neq n'}, \vec{f}_{n\neq n'})$.

$$p(\{\underline{\ell}_n\}_N, \{\vec{f}_n\}_N|\vec{\Omega}, \underline{\phi}) = \prod_n^N p(\underline{\ell}_n, \vec{f}_n|\vec{\Omega}, \underline{\phi}) \tag{2}$$

This assumption is necessary for us to easily combine the contributions of individual supernovae to the likelihood of the whole survey of supernovae. Here we are saying the the likelihood of the set of lightcurves and fluxes is simply the product of each of the individual likelihoods.

Next, we invoke marginalization over the latent variables, as an alternative to point estimation thereof.

$$p(\underline{\ell}_n, \vec{f}_n|\vec{\Omega}, \underline{\phi}) = \iiint p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n)\ p(\mu_n, z_n, t_n|\vec{\Omega}, \underline{\phi})\ d\mu_n\ dz_n\ dt_n \tag{3}$$

This is how the unobservable variables enter the inference. We note that Eq. 3 calls for likelihoods $\{p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n)\}_N$ of a supernova lightcurve and host galaxy photometry given some particular values of distance modulus, redshift, and type, but what we have are interim posteriors $\{p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})\}_N$, which are probabilities of distance modulus, redshift, and type given set values for the supernova lightcurve, host galaxy photometry, and priors from the data analysis procedure and survey program.

We thus must transform our expression to be in terms of quantities we actually have. We do this by multiplying the likelihood by an inspired factor of unity, in terms of the interim posteriors.

$$p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n) = p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n)\ \frac{p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})}{p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})} \tag{4}$$

We expand the denominator of that factor of unity according to Bayes' Rule.

$$p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n) = p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n)\ \frac{p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})\ p(\underline{\ell}_n, \vec{f}_n|\vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})}{p(\mu_n, z_n, t_n|\vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})\ p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})} \tag{5}$$

By the independence of the hierarchical levels in the probabilistic graphical model, we may split up the most daunting term in the above expression as

$$p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta}) = p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n)\ p(\underline{\ell}_n, \vec{f}_n|\vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta}). \tag{6}$$

Noting the presence of $p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n)$ and $p(\underline{\ell}_n, \vec{f}_n|\vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})$ in both the numerator and denominator for $p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n)$, we cancel the like terms to express the individual likelihoods in terms of known quantities, obtaining

$$p(\underline{\ell}_n, \vec{f}_n|\mu_n, z_n, t_n) = \frac{p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})}{p(\mu_n, z_n, t_n|\vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})}. \tag{7}$$

We are now ready to plug the individual likelihoods into Eq. 3, leading to

$$p(\underline{\ell}_n, \vec{f}_n|\vec{\theta}, \underline{\phi}) = \iiint p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}^*, \underline{\phi}^*)\ \frac{p(\mu_n, z_n, t_n|\vec{\theta}, \underline{\phi})}{p(\mu_n, z_n, t_n|\vec{\theta}^*, \underline{\phi}^*)}\ d\mu_n\ dz_n\ dt_n \tag{8}$$

Finally, we plug Eq. 8 back into Eq. 2.

$$p(\{\underline{\ell}_n, \vec{f}_n\}_N|\vec{\Omega}, \underline{\phi}) = \prod_n^N \iiint p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})\ \frac{p(\mu_n, z_n, t_n|\vec{\Omega}, \underline{\phi})}{p(\mu_n, z_n, t_n|\vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})}\ d\mu_n\ dz_n\ dt_n \tag{9}$$

And finally, we plug the product back into Eq. 1 to obtain

$$p(\vec{\Omega}, \underline{\phi}|\{\underline{\ell}_n, \vec{f}_n\}_N) \propto p(\vec{\Omega}, \underline{\phi})\ \prod_n^N \iiint p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})\ \frac{p(\mu_n, z_n, t_n|\vec{\Omega}, \underline{\phi})}{p(\mu_n, z_n, t_n|\vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})}\ d\mu_n\ dz_n\ dt_n, \tag{10}$$

the posterior on cosmological parameters and redshift-dependent type proportions that we wish to sample.

It is very important to document the assumptions we make with this model:

1. The expression of Eq. 10 is only as correct as the probabilistic graphical model of Fig. 1 is complete.
2. We must choose a prior probability distribution $p(\vec{Omega}, \underline{\phi})$.
3. The interim prior parameters $\vec{\theta}$ and $\underline{\xi}$ are known and shared among all supernovae n.

4. The selection function parameters $\vec{\beta}$ and $\underline{\alpha}$ are known and shared among all supernovae n.

5. All latent and observed parameters associated with the supernovae and their host galaxies are statistically independent from the latent and observed parameters of all other supernovae and their host galaxies.

6. The interim posterior distributions $\{p(z_n|\vec{f}_n, \vec{\theta}, \vec{\beta})\}_N$ and $\{p(t_n, z_n, \mu_n|\underline{\ell}_n, \underline{\xi}, \vec{\alpha})\}_N$ are accurate.

There are a few caveats to these assumptions.

Item 1 is present with any approach, as no model can include every aspect of the physics – to make any problem tractable, we must make simplifications. The hierarchical Bayesian approach, however, easily accommodates complications so is extensible to more refined models.

Skeptics of Bayesian statistics often cite item 2 as a major weakness of this approach. It is true that choosing a prior distribution can impart a bias on the resulting inference, but it can be an advantage when the data quality is poor and there are already trustworthy constraints on the parameters in question. We will nonetheless try to minimize the information imparted to the posterior by the prior.

It is noted that this model may be adapted to the cases when Items 3 and 4 are violated, which will be the case if the interim posteriors are not derived by a single method or if the photometry comes from a combination of surveys.

Item 5 is never truly valid in that all data observed by a single instrument are correlated, for example, but if the data primarily inform us about the phenomenon in question and we believe the principle of the universality of physics, it is a safe assumption; furthermore, virtually no statistical analysis would be possible without it, so all alternative methods already make this assumption.

It may seem to go without saying, but Item 6 can be difficult to guarantee. As will be discussed further in Sec. 3, there is as yet no established method for producing the interim posterior distributions, and validating any such method's accuracy will be a challenge, as it has proven to be for photo-z PDFs.

We implement the model of Sec. 2.1 in the form of the Supernova Cosmology Inference with Probabilistic Photometric Redshifts code `scippr`, which is a `Python` code freely available to the community.

3. MOCK DATA

The mock data used to validate this method is unusual in that joint interim posteriors over supernova type, redshift, and distance modulus have never before been made. This paper does not approach the problem of how to produce this data product because doing so would require us to make many more assumptions that could limit the impact of this work. As was mentioned in Sec. 2.1, it is nontrivial to confirm the accuracy of any probabilistic data analysis method. Doing so would require painstaking simulations of lightcurves and host galaxy photometry as well as a choice of lightcurve fitting scheme, and the method presented here would then depend on these being correct. If the mock data consists of interim posteriors, we can verify their self-consistency with our knowledge of the selection functions and interim priors by way of a rigorous forward model. For these reasons, we will assume that the data product in hand is a catalog of three-dimensional interim posteriors $\{p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\phi}, \vec{\alpha}, \vec{\beta})\}_N$.

In this section, we outline the forward model by which the mock interim posteriors are generated. The following subsections explain the choices made in our validation tests. From this point on, we will only work with log-probability distributions

3.1. *The true hyperparameters and parameters*

We begin by choosing parametrizations for the hyperparameters. We must have some functions of the hyperparameters producing a distribution from which the supernova type, redshift, and distance modulus are drawn.

For the redshift-dependent type proportions, we choose $p(t, z|\underline{\phi})$ to be a two-dimensional piecewise constant function over T types and Z redshift bins, making $\underline{\phi}$ a $T \times Z$ array with values ϕ_i equal to the probability of a supernova having the given type with a redshift in a given bin. A normalization condition is enforced such that summing over the discrete variable of supernova type and integrating over all redshift bins yields a value of unity.

To demonstrate that `scippr`is an extension of BEAMS, we consider $T = 3$, with $\tau \in \{Ia, Ibc, II\}$. Under this parametrization, we use realistic values for the elements of $\underline{\phi}'$ derived by convolving the delay time distribution (DTD) and the star formation history. This procedure is outlined in Secs. 3.1.1, 3.1.2, and 3.1.3. As an overview, we set the relative rates of SN Ia and Core Collapse to be 25% and 75%, respectively, at z = 0. **(@tinapeters How were these numbers modified for three SN populations?)** We use the DTD for SN Ia from Graur et al (2013) and the DTD for SN II from Zapartas et al (2017), and the Cosmic Star Formation Rate from Behroozi et al (2013). **(@tinapeters Put these references into the .bib file.)** The redshift-dependent supernova type proportions derived in this way are shown in Fig. 2. **(@tinapeters What is the native format of these curves, i.e. a continuous function evaluated on a grid or something else? I wrote about a piecewise constant function, but it can and should be whatever's consistent with the actual functions you calculated.)** $N = 10^4$ pairs of (t'_n, z'_n) are drawn from this distribution.

For the parametrization of the cosmological parameters, we assume a ΛCDM cosmology so $\vec{\Omega}$ is comprised of H_0, Ω_m, etc. We choose the true cosmological parameters comprising $\vec{\Omega}'$ to be the maximum likelihood point estimates observed by Planck: $H_0 = 67.9\ Mpc/km/s$ and $\Omega_m = 0.307$. **(@aimalz fix these constants here and in code)**

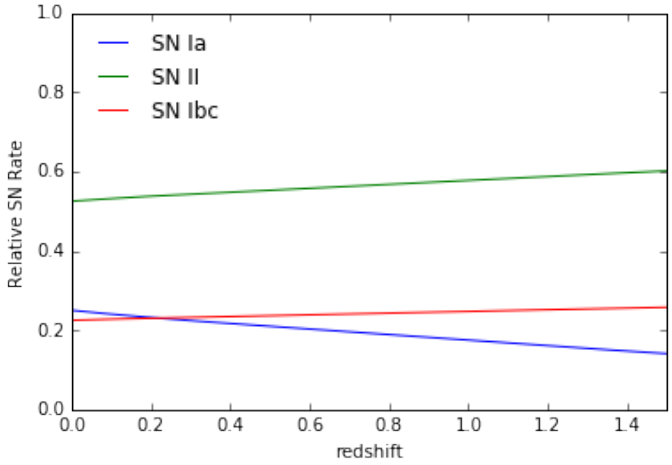

FIG. 2.— The relative supernova rates as a function of redshift. Sums to one at every redshift (but should actually integrate to one over type and redshift). **(@aimalz Also plot the "observed" version of this for our sample)**

The function that produces μ from a given z_n is

$$\mu = 5\log\left[(1+z)\frac{1}{10\ pc}\int_0^z \frac{dz'}{\sqrt{\Omega_M(1+z)^3+\Omega_k(1+z)^2+\Omega_\Lambda}}\right]. \tag{11}$$

This means that once the pair (t'_n, z'_n) is drawn from $p(t, z|\underline{\phi}')$, $p(\mu_n|\vec{\Omega}', z')$ is a delta function centered at μ'_n following Eq. 11. We have now set the true values of the hyperparameters and parameters

3.1.1. *Supernova Type Ia Rate with Redshift*

3.1.2. *Supernova Type Ibc Rate with Redshift*

3.1.3. *Supernova Type II Rate with Redshift*

To determine the rate of SNII per unit comoving volume we will basically apply the approach by Forster et al. 2006/Strogler et al. 2004 adapted to SNe II following Botticella et al. 2012.

The rate of SNe II per unit time per unit comoving volume (R_{II}) is given by the star formation rate (SFR) per unit time per unit comoving volume convolved with the number of stars crearted that will explode as SNe II.

$$R_{II} = K_{II} \times SFR \tag{12}$$

First, we must calculate the fraction of stars that explode as SN II:

$$K_{II} = \frac{\int_{m_{l,II}}^{m_{u,II}} \phi(m)\mathrm{d}m}{\int_{m_l}^{m_u} m\phi(m)\mathrm{d}m} \tag{13}$$

When we integrate from a minimum mass, $m_{l,II}$, of 8 to a maximum mass, $m_{u,II}$, to 25 and assuming a Salpeter et al. (1955) IMF we get a rate of 6.027^{-3}.

We will use three different models of how the SFR evolutions with redshift: Cole et al. (2001), Horiuchi et al. (2011), and Madau et al. (2014)

Next we will take the spectral model of Dessart et al. (2013) which is anchored to SN II 1999em, as a reference. We simulate its r-band light curve at different redshifts and convolve it with the LSST filters to see when it falls below the detection limit.

For this model SN II, we compare the magnitude as a function of time and redshift, $m(t, z)$, with the limiting magnitude, m_{lim}, of the LSST camera to obtain the probability of detecting (the peak of) a SN at a given redshift, $\Delta_t(z)$. Then we estimate the number of detected SNe II per unit of redshift, dN/dz, using Equations 2 and 5 of Forster et al. (2006).

Finally, we will use a a sample of 10,000 SNe, simulated using MCMC on real parameters of CSP SNe II. We will put each at 100 random redshifts from 0.01 to 1.20, so we will have 1,000,000 SNe. Now we consider the magnitude at the end of the plateau phase, instead of the peak magnitude. This is because in SN II cosmology, we do not standardize the magnitude at peak, but either the magnitude at around the center of the plateau (for spectroscopic methods), or the length/brightness-decline of the plateau (for photometric methods).

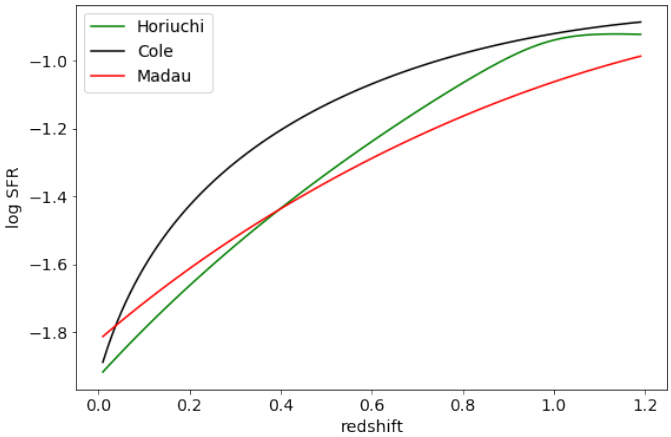

FIG. 3.— SFR evolutions with redshift: Cole et al. (2001), Horiuchi et al. (2011), and Madau et al. (2014).

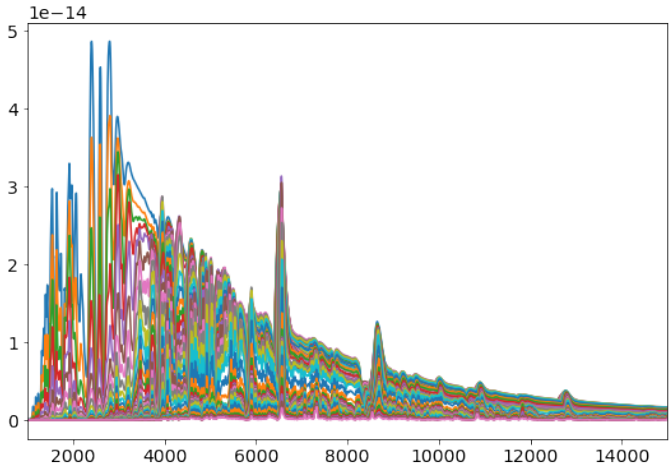

FIG. 4.— SN II spectral model of Dessart et al. (2013): $M_{99em} = -16.6$; $M_{05J} = -17.2$

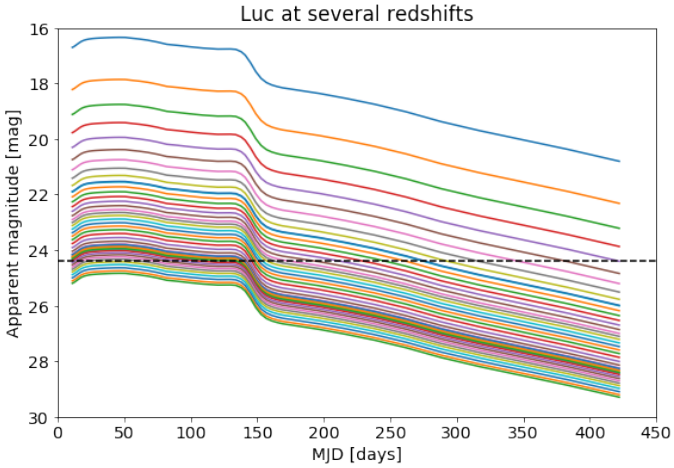

FIG. 5.— The r-band lightcurve of spectral model of SN II 1999em shifted to a range of redshifts. The horizontal dashed line indicates the limiting magnitude of LSST in the r-band in a single visit.

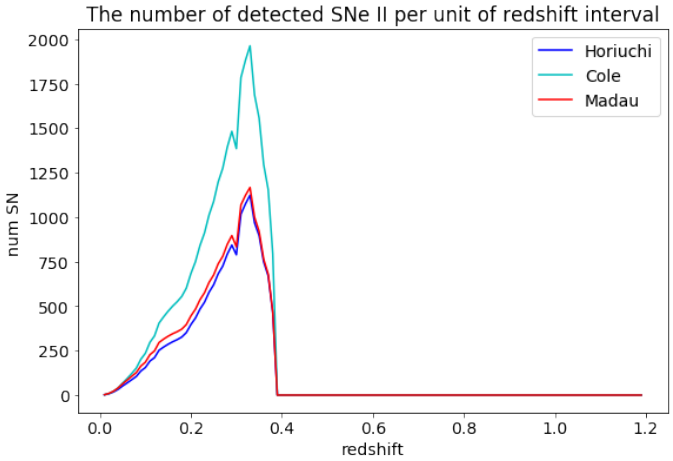

FIG. 6.— The number of SN II detected, for each of the SFR models, as a function of redshift. The total number of SN II for Horiuchi et al. 16985, Cole et al. 29207, and Madau et al. 18322.

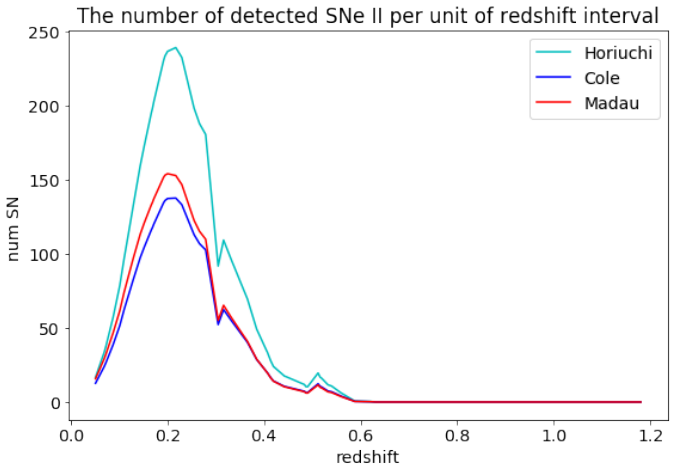

FIG. 7.— The number of SN II detected, for each of the SFR models, as a function of redshift. The total number of SN II for Horiuchi et al. 16985, Cole et al. 29207, and Madau et al. 18322.

We use the apparent magnitudes at the end of the plateau phase and k-correct the magnitudes. Then we measure $\Delta_t(z)$, selecting different magnitude limit depending on the filter we use for observations. Finally, we compare the number of supernovae as a function of redshift for the three different SFR models.

3.2. *Constructing PDFs*

The catalog of trios of true latent variables must be transformed into three-dimensional probability distribution functions (PDFs) over these three variables. To improve the efficiency of `scippr`, we use a binned parametrization for the PDFs $p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n, \vec{\theta}, \underline{\xi}, \vec{\alpha}, \vec{\beta})$ so that the mathematical manipulations of the PDFs are simple operations on arrays. This choice is made at the level of writing code, not something intrinsic to the model. For these tests, we use $J = 20$ equally spaced bins in redshift and $K = 20$ equally spaced bins in distance modulus.

We construct the PDFs from simple components described in the following sections, guided by the graphical model of Fig. 1. We will be assuming separability of the data

3.2.1. *Creating posteriors*

The first components we must construct are the posterior distributions $p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n)$. We note the statistical independence of $\underline{\ell}_n$ and $\vec{f}_n$, as there are no direct connections between them in Fig. 1. Furthermore, while the supernova lightcurve is a function of all three latent variables, the host galaxy photometry is independent of the distance modulus and supernova type according to the model of Fig. 1. (Though there is some evidence against this assumption, the details of any potential relationship have not yet been determined well enough to model, so we do not include this effect at this time.) Because of these instances of independence, we may decompose the posterior as

$$p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n) = p(\mu_n, z_n, t_n|\underline{\ell}_n)\ p(z_n|\vec{f}_n). \tag{14}$$

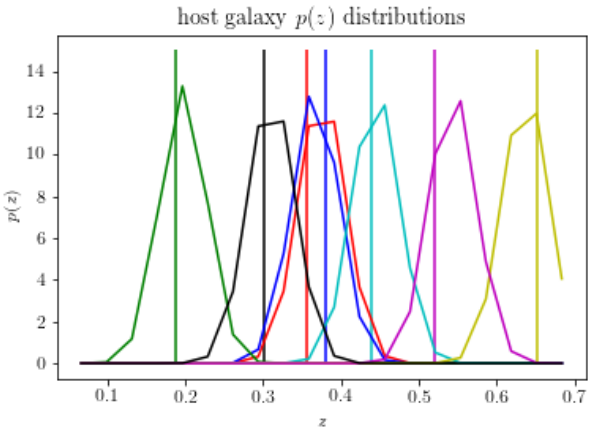

FIG. 8.— These are examples of mock photo-z posteriors. **(@aimalz Change these to plot as piecewise constant.)**

We construct these terms without simulating supernova lightcurves nor host galaxy photometry. The true supernova types, redshifts, and distance moduli serve as a proxy for the information that would be carried by the observed data; in other words, we are saying

$$p(\mu_n, z_n, t_n|\underline{\ell}_n, \vec{f}_n) = p(\mu_n, z_n, t_n|\mu'_n, z'_n, t'_n)\ p(z_n|z'_n). \tag{15}$$

We thus emulate the desired posteriors based on our understanding of the forward model of point estimates of these quantities.

We first approach the second term, the photo-z posterior. The piecewise constant parametrization is flexible enough to accommodate many shapes for the redshift posterior, but for now, we assume it is a binned Gaussian distribution $\mathcal{N}(z''^f_n, \sigma_f^2)$ with a variance σ_f^2 shared by the entire dataset and a mean of z''^f_n drawn from a Gaussian distribution $\mathcal{N}(z'_n, \sigma_f^2)$ of the same variance with a mean of z'_n, the true redshift. Examples of these photo-z posteriors are shown in Fig. 8.

The first term is more challenging due to its higher dimensionality. We are modeling the expected product of a probabilistic lightcurve fitter that gives a joint probability distribution over supernova type, redshift, and distance modulus. As this is an emulation procedure, rather than a simulation, we construct the desired quantity using an understanding of how a lightcurve fitter works. Depending on the type of supernova, a different fitting function is used, so typically a classifier is run on the lightcurves first, and then that information is fed into a lightcurve fitter. This is equivalent to

$$p(\mu_n, z_n, t_n|\mu'_n, z'_n, t'_n) = p(t_n|t'_n)\ p(\mu_n, z_n|\mu'_n, z'_n, t'_n, t_n). \tag{16}$$

The quantity $p(t|t'_n)$ is a vector of length T in which each cell is the probability of a supernova's classified type given its true type, evaluated at the true type t'_n of supernova n. The $T \times T$ matrix defined by $p(t|t')$ is called the confusion matrix $\underline{C}$ of a classifier, whose diagonal elements give the fraction of supernovae classified as their true type. **(We need to choose a confusion matrix typical of a modern classifier and use that in the tests.)**

We motivate the mechanism producing $p(\mu_n, z_n|\mu'_n, z'_n, t'_n, t_n)$. We currently assume that this function is separable into a redshift-dependent component $p(z_n|z'_n)$ and a redshift-independent component $p(\mu_n|\mu'_n, t'_n, t_n)$. **(We can and should revise this assumption. @reneehlozek Where can we find information upon which to base a better emulation model?)** For these, we use the same model as for the redshift posteriors based on host galaxy photometry, with a variance σ_ℓ^2 and mean z''^ℓ_n, yielding a vector of length J.

When constructing a traditional Hubble diagram, we only attempt to estimate a distance modulus for type Ia supernovae, leaving the distance modulus of non-standardizable types unconstrained. This choice is based on the classified type t_n rather than the true type. When there is a misclassification, i.e. $t_n \neq t'_n$, a type Ia supernova might be given an unconstrained distance modulus and a core-collapse supernova would be assigned systematically incorrect distance modulus. In particular, type II contaminants tend to have a high scatter around a constant distance modulus and type Ibc contaminants tend to be systematically biased to lower distance moduli. **(We need to find one of those figures showing these functions!)** Based on these observations, we establish functions $p(\mu_n|\mu'_n, t'_n, t_n)$ corresponding to all elements of the confusion matrix and evaluate them at the known values of μ'_n and t'_n and each possible value of t_n, yielding a $T \times K$ matrix for each supernova in the sample. The functions used in this emulation

procedure are as follows

$$p(\mu_n|\mu'_n, t'_n = Ia, t_n = Ia) = \mathcal{N}(\mu''_n, \sigma^2_{Ia}) \; with \; \mu''_n \sim \mathcal{N}(\mu'_n, \sigma^2_{Ia}) \tag{17}$$

$$p(\mu_n|\mu'_n, t'_n = Ibc, t_n = Ia) = \mathcal{N}(\mu''_n, \sigma^2_{Ibc}) \; with \; \mu''_n \sim \mathcal{N}(\mu'_n - c_{Ibc}, \sigma^2_{Ibc}) \tag{18}$$

$$p(\mu_n|\mu'_n, t'_n = II, t_n = Ia) = \mathcal{N}(\mu''_n, \sigma^2_{II}) \; with \; \mu''_n \sim \mathcal{N}(c_{II}, \sigma^2_{II}) \tag{19}$$

$$p(\mu_n|\mu'_n, t'_n, t_n \neq Ia) = U(\mu_{min}, \mu_{max}), \tag{20}$$

where c_{Ibc} and c_{II} are constants shared among all supernovae.

3.2.2. *Incorporating the selection function*

So far, we have not considered the effect of selection functions in the space of observed supernova lightcurves and host galaxy photometry, though they are obviously very important. We assume that the survey producing the catalog of three-dimensional posteriors knows its selection functions in the space of data, which correspond to $p(\underline{\ell}|\vec{\alpha})$ and $p(\vec{f}|\vec{\beta})$. Given models for the relationships between data and latent variables $p(\mu, z, t|\underline{\ell})$ and $p(z|\vec{f})$, we could marginalize over the space of all possible data to obtain $p(\mu, z, t|\vec{\alpha})$ and $p(z|\vec{\beta})$, which would then simply be multiplied to get $p(\mu, z, t|\vec{\alpha}, \vec{\beta})$.

Because we do not simulate the supernova lightcurves nor host galaxy photometry, this approach is not available to us. Once again, we will emulate it, this time using summary statistics of mock data from realistic simulations. We can calculate the recovery rate in the space of latent variables given different selection function parameters by actually imposing those selection functions on realistic simulations for which the true distribution in the space of μ, z, and t is known.

In the simple case of host galaxy photometry, we assume the cuts in magnitude and signal-to-noise ratio expected of LSST and apply these to the photometry of the Buzzard catalog with a known redshift distribution. By simply taking the ratio of recovered galaxies to simulated galaxies within each of the J redshift bins. Note that we have already assumed no correlation between host galaxy properties and supernova type, but if there were a simulation that accurately included this effect, we could obtain a selection function in the space of t and z based on that. **(We need to actually do this with realistic cuts and mock data.)** This procedure gives us $p(z|\vec{\beta})$.

In the three-dimensional case of supernova lightcurves, there is not a mock dataset with information on the true number of supernovae at each μ, z, and t. **(Is this the nominal reason why we have to take the ratio of what's recovered with the WFD selection cuts to what's recovered with the DDF selection cuts? I don't understand how we have either of those numbers if that data is not available. Is it because the recovery rate requires choosing a classifier and lightcurve fitter?)** Instead of using the ratio of the number of supernovae recovered over the number of supernovae simulated as a function of their latent variables, we use the ratio of the recovery rate for the selection function in question over the recovery rate for the most generous possible selection function. For LSST, these correspond to the Wide Fast Deep (WFD) and Deep Drilling Fields (DDF) selection functions. This procedure gives us $p(\mu, z, t|\vec{\alpha})$.

3.2.3. *Making interim posteriors*

To make interim posteriors, we introduce the interim priors of Eq. **??**. We choose $\vec{\theta}^*$ and $\underline{\phi}^*$ [specify what a good, general choice would be]. Because the parametrization hasn't changed, the term $p(z_\zeta, t_\tau|\underline{\phi}^*)$ takes the simple form of $\phi^*_{\tau\zeta}$. Similarly, $p(\mu_\nu|z_\zeta, \vec{\theta}^*)$ will be a delta function $\delta_{f_{\vec{\theta}^*}(z_\zeta)}(\mu_\nu)$.

Finally, in Eq 21, we put these pieces together to express the form of the individual interim posteriors of the form of Eq. **??**.

$$p_n(\mu_\nu, z_\zeta, t_\tau|\underline{\ell}_n, \vec{m}_n, \vec{\theta}^*, \underline{\phi}^*) = KC_{\tau'_n\tau}\mathcal{N}_{(\hat{z}^\ell, \hat{\mu}^\ell), \underline{\Sigma}_n}(z_\zeta, \mu_\nu)\mathcal{N}_{\hat{z}^m_n, \sigma^2_n}(z_\zeta)\phi^*_{\zeta\tau}\delta_{f_{\vec{\theta}^*}(z_\zeta)}(\mu_\nu) \tag{21}$$

The constant of proportionality K here will be set such that $\sum_\nu^D \sum_\zeta^Z \sum_\tau^T p_n(\mu_\nu, z_\zeta, t_\tau|\underline{\ell}_n, \vec{m}_n, \vec{\theta}^*, \underline{\phi}^*) = 1$.

4. VALIDATION

5. DISCUSSION

6. CONCLUSION

REFERENCES

M. Kunz, B. A. Bassett, and R. A. Hlozek, Physical Review D **75**, 103508 (2007).

B. C. Kelly, X. Fan, and M. Vestergaard, The Astrophysical Journal **682**, 874 (2008).

R. Hlozek, M. Kunz, B. Bassett, M. Smith, J. Newling, M. Varughese, R. Kessler, J. P. Bernstein, H. Campbell, B. Dilday, B. Falck, J. Frieman, S. Kuhlmann, Hubert Lampeitl, J. Marriner, R. C. Nichol, A. G. Riess, M. Sako, and D. P. Schneider, The Astrophysical Journal **752**, 79 (2012)