Deriving Photometric Redshifts from Observed Flux Densities using Hierarchical Bayes and Machine Learning

Joshua S. Speagle[1,2⋆], Alexie Leauthaud[3,2], Kevin Bundy[3,4,2], Peter Capak[5], Jean Coupon[6], Daniel Eisenstein[1], Boris Leistedt[7], Daniel Masters[8], Daniel Mortlock[9,10], Hiranya Peiris[11,12]

[1] *Harvard University, 60 Garden St., MS 46, Cambridge, MA 02138, USA*
[2] *Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, Chiba, Japan*
[3] *University of California Santa Cruz, 1156 High St., Santa Cruz, CA 95064, USA*
[4] *University of California Observatories/Lick Observatory, 1156 High St., Santa Cruz, CA 95065, USA*
[5] *Spitzer Science Center, California Institute of Technology, Pasadena, CA 91125, USA*
[6] *University of Geneva ch. d'Ecogia 16, CH-1290 Versoix, Switzerland*
[7] *Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003, USA*
[8] *Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA*
[9] *Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London, SW7 2AZ, UK*
[10] *Department of Mathematics, Imperial College London, London, SW7 2AZ, UK*
[11] *Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK*
[12] *Oksar Klein Centre for Cosmoparticle Physics, Stockholm University, AlbaNova, SE-106 91 Stockholm, Sweden*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We combine Bayesian inference with machine learning to derive photometric redshifts in a robust, data-driven way. By conducting targeted likelihood fitting over observed photometric probability density functions (PDFs), we show how target objects can be sparsely projected onto an arbitrary photometric basis in the presence of observational errors, missing data, and selection effects. The joint posterior between the ensemble (population) and individual (object) distributions over this basis can then be explored using hierarchical Bayesian inference. Accurate photometric redshift PDFs (along with numerous other quantities) can subsequently be computed as a by-product. Our code (`FRANKEN-Z`), its extensions, and several interactive tutorials can be found online at `https://github.com/joshspeagle/frankenz`.

Key words: methods: statistical – techniques: photometric – galaxies: distances and redshifts

1 INTRODUCTION

In the ongoing era of "precision cosmology", large-scale extra-galactic surveys will collect photometric data to billions of galaxies from the optical to the near-infrared (NIR) in order to measure cosmological parameters and the growth of large-scale structure to (sub-)percent level accuracy. Many of these surveys (e.g., DES[1], KiDS[2], HSC[3], LSST[4], *Euclid*[5], *WFIRST*[6]) will rely on exquisite shear measurements (e.g., from galaxy-galaxy lensing and/or cosmic shear) to meet

[⋆] E-mail: jspeagle@cfa.harvard.edu

[1] The Dark Energy Survey (The Dark Energy Survey Collaboration 2005). See `darkenergysurvey.org`.

[2] Kilo-Degree Survey (de Jong et al. 2013). See `kids.strw.leidenuniv.nl`.

[3] The Hyper Suprime-Cam (HSC; Miyazaki et al. 2012) Subaru Strategic Program (SSP) survey (Miyazaki et al., in preparation). See `hsc.mtk.nao.ac.jp/ssp`.

[4] The Large Synoptic Survey Telescope (Ivezic et al. 2008). See `lsst.org`.

[5] Laureijs et al. (2011). See `euclid-ec.org`.

[6] The *Wide-Field Infrared Surey Telescope* (Spergel et al. 2015). See `wfirst.gsfc.nasa.gov`.

ambitious weak lensing-oriented science goals. Due to the sensitivity of these quantities on the inferred redshift (z) of (1) a given galaxy and (2) the overall redshift distribution across an ensemble of galaxies, meeting the science goals for these surveys heavily depends on measuring accurate, unbiased redshifts to a significant portion of galaxies in their photometric samples.

Due to the substantial challenge of obtaining high-confidence spectroscopic redshifts (spec-z's) for the majority of galaxies in the full photometric sample (Masters et al. 2015), these surveys will almost entirely rely on photometric redshifts (photo-z's) derived from galaxy broad-/narrow-band spectral energy distributions (SEDs). This leads to a unique set of challenges, as the sensitivity of shear measurement bias as a function of the assumed redshift requires not only accurate photo-z's on average but also accurate characterization of the associated uncertainties (i.e. the full probability density function; PDF) for each individual object.

Two main techniques have been used in the literature to derive $P(z|\hat{\mathbf{F}}_g, \hat{\mathbf{C}}_g)$ for a given set of observed fluxes $\hat{\mathbf{F}}_g$ and covariances $\hat{\mathbf{C}}_g$ of a given galaxy g. **Template fitting** relies on deriving a set of *forward mappings* and their likelihoods from a collection of "templates" (i.e. models) and associated (nuisance) modeling parameters to observed color space (e.g., Arnouts et al. 1999; Benítez 2000; Bolzonella et al. 2000; Ilbert et al. 2006; Cool et al. 2013; Johnson et al. 2013; Tanaka 2015). **Machine learning**, on the other hand, uses a collection of training data to derive the best *inverse mappings* from observed color space to redshift (e.g., Sheldon et al. 2012; Carrasco Kind & Brunner 2013; Hoyle 2016; Elliott et al. 2016; Almosallam et al. 2016).

While the "empirical" nature and computational speed of machine learning has made it an attractive research area over the past few years (see, e.g., Hildebrandt et al. 2010; Dahlen et al. 2013; Sánchez et al. 2014, for compilations of recent work), the increasingly sophisticated use of Bayesian methods has greatly improved the information content that can be extracted from photometric data using template fitting approaches (Benítez 2000; Moustakas et al. 2013). This includes advances in properly incorporating uncertainties from both the observed photometry and input templates, prior knowledge based on previous surveys, and the development of hierarchical models that can simultaneously model the joint distribution between individual galaxies and the overall population (Brammer et al. 2008; Tanaka 2015; Speagle & Eisenstein 2015a,b; Leistedt et al. 2016).

While Bayesian inference is powerful, to date applications have been limited exclusively to template-based methods. At a fundamental level, however, machine learning methods *are* Bayesian in nature[7], and should be interpreted as practical approximations to full Bayesian inference(Bishop 2006). From this perspective, conducting Bayesian inference over training data rather than an underlying set of models is not only possible, but natural. Indeed, from this perspective, the difference between pre-generated sets of model photometry and training data is mostly semantics — template-fitting methods assume a functional form for the prior/likelihood (to compute the posterior) while machine learning attempts to derive the posterior directly.

In this paper, we extend the hierarchical Bayesian formalism developed in Leistedt et al. (2016) to jointly explore the probabilistic association between individual objects and the overall populations of two arbitrary datasets spanning the same domain. This allows us to extend Bayesian analysis to the types of training data more typically used in machine learning approaches. By using machine learning to generate sparse approximations to the original likelihoods, our hybrid approach simultaneously leverages the speed and data-driven nature of machine learning while retaining the statistical rigor of traditional Bayesian analysis. This represents a significant departure from previous "hybrid" photo-z approaches, which have tended to focus on using Bayesian methods to combine photo-z PDFs *after* they have been computed using other methods (see, e.g., Carrasco Kind & Brunner 2014).

Our sparse approximations further allows us to keep track of when, where, and how *individual objects* from our training set contribute to individual and ensemble predictions. This not only allows us to decouple our original likelihoods from any posterior modifications we hope to compute later, but also enables detailed analysis of where the information content used to compute our photo-z PDFs is coming from. We exploit these feature in Speagle et al. (in prep.) to constrain possible uncertainties in photo-z estimation from heterogeneously populated training data using weak lensing-selected data from the HSC SSP Survey.

The outline of the paper is as follows. In §2, we outline the fundamental connection between model-fitting and machine-learning approaches to photometric inference, discuss its application to deriving redshifts, and provide an overview of our statistical framework. In §3, we discuss methods for computing likelihoods in the context of different informative metrics and missing data. In §4, we discuss how to modify our posteriors to account for selection effects. In §5, we describe how to conduct basic population inference using training data and its extensions to our full hierarchical Bayesian model. In §6, we outline how machine learning can be used to approximate the true distribution from a Bayesian perspective and outline our particular application to approximating our likelihoods and posteriors. In §7, we demonstrate that our model performs well on mock and real data. We discuss our results in §8, and conclude in §9.

Throughout this work, boldface ($\mathbf{x}$, $\boldsymbol{\theta}$) and braces ($\{x\}$, $\{\theta\}$) are used to represent collections of elements (i.e. vectors, matrices) while italics (x, θ) is used for singular variables and functions. All vectors used are *column vectors* (i.e. $n \times 1$ matrices) unless explicitly stated otherwise. Subscripts refer to the individual subset ($\mathbf{x}_i$) or element (θ_i) at that particular index (i). A summary of variables and some examples of notation are shown in Table **??**.

All of the tests included in this paper and our code can be found online at `https://github.com/joshspeagle/frankenz`.[8]

[7] This can be shown in great detail for linear regression (see, e.g., Hogg et al. 2010), which represents one of the simplest applications of "machine learning".

[8] Add Zenodo link when we freeze in the version.

2 BACKGROUND

2.1 Photometric Inference

At the most basic level, photo-z's rely on estimating is how "similar" two objects g and h are to each other, where $g \in \mathbf{g}$ is an object from sample $\mathbf{g} = \{\ldots, g, \ldots\}$ with some associated observed values $\hat{\mathbf{X}}_g$ and $h \in \mathbf{h}$ is an object from sample $\mathbf{h} = \{\ldots, h, \ldots\}$ with observed values $\hat{\mathbf{X}}_h$. In other words, given some hypothetical galaxy i with intrinsic values $\mathbf{X}_i$, we want to find out how likely we would observe both $\hat{\mathbf{X}}_g$ and $\hat{\mathbf{X}}_h$. Since we don't know what $\mathbf{X}_i$ is, however, we want to *marginalize* over these quantities to find the probability that h and g are consistent with being the same galaxy over all possible intrinsic values of our observables.

If we assume *Poisson independence* among all objects in $\mathbf{g}$ and $\mathbf{h}$, this can be written more formally as

$$\int P(\hat{\mathbf{X}}_g, \hat{\mathbf{X}}_h | \mathbf{X}_i) P(\mathbf{X}_i)\, d\mathbf{X}_i \equiv \int P(g, h | i') P(i')\, di' \tag{1}$$

$$= P(g, h) = P(g|h)P(h) = P(h|g)P(g) \quad . \tag{2}$$

where we have used our index variables (g, h, i) as a stand-in for their observed features $(\hat{\mathbf{X}}_g, \hat{\mathbf{X}}_h, \mathbf{X}_i)$ and the integral over i' is taken over all possible *realizations* of i (rather than all $i \in \mathbf{i}$). We will return to this formalism in §3.

In practice, $\mathbf{g}$ is often a set of *target* objects with *unknown* properties and $\mathbf{h}$ is a set of *training* objects with *known* properties. We are thus interested in $P(h|g)$, which tells us the probability that our observed object g is consistent with our training object h. Rearranging equation (2), we get *Bayes theorem*

$$P(h|g) = \frac{P(g|h)P(h)}{P(g)} \tag{3}$$

where $P(h|g)$ is the *posterior* of h given g, $P(g|h)$ is the *likelihood* of g given h, $P(h)$ is the *prior* for h, and

$$P(g) = \sum_{h \in \mathbf{h}} P(g|h)P(h) \tag{4}$$

is the *evidence* for g summed across all training objects $h \in \mathbf{h}$. Our chosen metric for computing the likelihood $P(g|h)$ is completely unspecified here, but is some function across a set of common observables between g and h related to equation (1).

From the model-fitting (i.e. template-fitting) perspective, this formalism is straightforward to interpret. Our set of "training data" (i.e. data where the properties of interest are known) $\mathbf{h} = \{\ldots, h(\boldsymbol{\theta}_h), \ldots\}$ is generated from an underlying set of models parameterized $\boldsymbol{\theta}$, our likelihood $P(g|h) = P(g|\boldsymbol{\theta}_h)$ compares the observables generated from our model with the actual observations, and our prior $P(h) = P(\boldsymbol{\theta}_h)$ is now defined over $\boldsymbol{\theta}$. We can either sample from the posterior $P(h|g)$ using "brute-force" tactics by fitting a set of models and their priors that were generated in advance (Ilbert et al. 2006; Moustakas et al. 2013), or by drawing samples directly from the posterior using, e.g., Markov Chain Monte Carlo (MCMC) methods (Acquaviva et al. 2011; Johnson et al. 2013; Speagle et al. 2016).

Most machine learning techniques do not assign weights to the objects in their training set, implicitly assuming that they are drawn from the posterior. Training a given machine learning algorithm on given dataset thus provides "direct access" to the posterior via the product $P(g|h)P(h) \propto P(h|g)$. In astronomy, this assumption is almost always violated in practice due to strong selection effects, which traditionally has required some combination of reweighting, resampling, and Monte Carlo (MC) methods to try and "massage" the training data $\mathbf{h}$ to better approximate to the target population $\mathbf{g}$ (see, e.g., Lima et al. 2008).

2.2 Inferring Redshift from Photometry

We are interested in inferring the redshift PDF $P(z|g)$ for our observed object g. Given our discrete set of training objects $\mathbf{h}$, this can be factored and written as

$$P(z|g) = \sum_{h \in \mathbf{h}} P(z, h|g) = \sum_{h \in \mathbf{h}} P(z|h)P(h|g)$$
$$= \sum_{h \in \mathbf{h}} P(z|h) \frac{P(g|h)P(h)}{P(g)} \quad , \tag{5}$$

which constitutes stacking a posterior-weighted set of underlying redshift kernels (i.e. probabilistic *labels*) associated with each training object.

One often overlooked subtlety in estimate photo-z PDFs is that $P(z|h)$—and by extension $P(z|g)$—is fundamentally a *mixed distribution* as a function of *object type* t_h. If h is a star, for instance, then we know that it's true redshift $z_{\rm true}$ is distributed as (i.e. is drawn from)

$$z_{\rm true} \sim P(z|t_h = {\rm star}) = \delta(z = 0) \quad , \tag{6}$$

which will always be located at the *discrete* value $z = 0$. If h is a galaxy, however, then

$$z_{\rm true} \sim P(z|t_h = {\rm gal}) \tag{7}$$

where $P(z|t_h = {\rm gal})$ is a *continuous* (unknown) distribution.

If we don't know the underlying t_h for a given h, then our redshift kernel $P(z|h)$ must be computed by *marginalizing* over t_h. Since t_h only encompasses two discrete classes, this is simply

$$P(z|h) = P(t_h = {\rm star})\delta(z = 0)$$
$$+ P(t_h = {\rm gal})P(z|h, t_h = {\rm gal}) \quad . \tag{8}$$

By extension, our redshift PDF for g can be written as

$$P(z|g) = \left[\sum_{h \in \mathbf{h}} P(t_h = {\rm star})P(h|g)\right] \delta(z = 0)$$
$$+ \sum_{h \in \mathbf{h}} P(t_h = {\rm gal})P(z|h, t_h = {\rm gal})P(h|g) \tag{9}$$
$$= P(t_g = {\rm star})\delta(z = 0) + P(z, t_g = {\rm gal}|h)$$

where $P(z, t_g = {\rm gal}|h) = P(t_g = {\rm gal})P(z|h, t_g = {\rm gal})$ is the redshift PDF for g assuming it's a galaxy.

In this paper, we will mostly focus on the more basic formalism outlined in equation (5) and assume that $t_h = t_g = {\rm gal}$ for all $g \in \mathbf{g}$ and $h \in \mathbf{h}$ for convenience. We note that our overall formalism can be easily extended to incorporate this additional subtlety, and discuss cases where explicitly modeling object types becomes more important in §8.

2.3 Hierarchical Inference

Our overarching motivation is to compute the **probabilistic association** between our target sample $\mathbf{g}$ and our training sample $\mathbf{h}$. These can be parameterized by a set of *population weights* $\mathbf{w}$ over $\mathbf{h}$, which have a corresponding PDF $P(\mathbf{w}|\mathbf{h},\mathbf{g},\mathbf{p_h})$ for a given prior $\mathbf{p_h} \equiv \{\dots,P(h),\dots\}$. The *maximum-likelihood* solution to this can be shown to be[9] the Bayesian evidence for training object h computed over all target objects g defined as

$$w = \sum_{g\in\mathbf{g}} P(h|g)P(g) = p_h \sum_{g\in\mathbf{g}} P(g|h) \quad . \tag{10}$$

This can be seen as the full Bayesian solution to the k-nearest neighbor-based reweighting scheme presented in Lima et al. (2008).

Ideally, however, we would like to estimate our population weights along with our collection of individual posteriors $P(\mathbf{h}|\mathbf{g})$. We are thus interested in sampling a *collection* of population weights $\{\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)}\}$ from the *joint distribution* $P(\mathbf{h},\mathbf{w}|\mathbf{g},\mathbf{p_h})$ using a **hierarchical Bayesian model**. Each set of weights then corresponds to a specific redshift population distribution and can be used to compute an updated set of posteriors, as outlined in §5.

An important feature of our model is that this probabilistic association is conducted entirely over the *observables* spanned by the discrete collection of training objects, without regards to any of the underlying (probabilistic) labels.[10] As a result, once we obtain a set of samples from the joint distribution over our training set $\mathbf{h}$, we can (in theory) "paint on" *any* underlying mapping we desire to project our samples onto the label(s) we are interested in. While in this paper we focus specifically on the $\mathbb{R}^p \to \mathbb{R}^1$ mapping from our p-dimensional space of observables into the 1-D space of redshift, in future work we hope to broaden this approach to multidimensional distributions encompassing ancillary physical parameters such as stellar mass.

3 PHOTOMETRIC LIKELIHOODS

We wish to compute the likelihood $P(g|h)$ between galaxies g and h given their *kernel density estimates* (KDEs) $K_g(\mathbf{F}|\hat{\mathbf{F}}_g)$ and $K_h(\mathbf{F}|\hat{\mathbf{F}}_h)$, where $\hat{\mathbf{F}}_g = \{\hat{F}_{g,1},\dots,\hat{F}_{g,n}\}$ and $\hat{\mathbf{F}}_h = \{\hat{F}_{h,1},\dots,\hat{F}_{h,n}\}$ are the corresponding set of n observables associated with galaxies h and g, respectively. These are most often a set of observed *flux densities* within a set of bands $\mathbf{b}$, as suggested by the notation, but can also include additional observables such as a galaxy's shape or size. We will henceforth refer to them as "flux densities" for simplicity.

As equation (1), our likelihood is the convolution of these two KDEs over all possible flux densities

$$\begin{aligned} P(g|h) &= \int P(\mathbf{F})P(g|h,\mathbf{F})\,d\mathbf{F} \\ &= \int P(\mathbf{F})K_g(\mathbf{F}|\hat{\mathbf{F}}_g)\,K_h(\mathbf{F}|\hat{\mathbf{F}}_h)\,d\mathbf{F} \end{aligned} \tag{11}$$

where $P(\mathbf{F})$ is a prior over the flux densities across $\mathbf{b}$.

3.1 Flux Densities

For most galaxies, we can approximate our flux density PDFs as a multivariate Normal (i.e. Gaussian) distribution such that

$$K_g(\mathbf{F}|\hat{\mathbf{F}}_g) = \mathcal{N}(\mathbf{F}|\hat{\mathbf{F}}_g,\hat{\mathbf{C}}_g) \equiv \frac{\exp\left[-\frac{1}{2}||\mathbf{F}-\hat{\mathbf{F}}||_{\hat{\mathbf{C}}^{-1}}\right]}{(2\pi)^{1/2}||\hat{\mathbf{C}}_g||^{1/2}} \tag{12}$$

where

$$||\mathbf{F}-\hat{\mathbf{F}}||_{\hat{\mathbf{C}}^{-1}} \equiv (\mathbf{F}-\hat{\mathbf{F}}_g)^T\hat{\mathbf{C}}_g^{-1}(\mathbf{F}-\hat{\mathbf{F}}_g) \tag{13}$$

is the *squared Mahalanobis distance* or *Mahalanobis norm* (i.e. the error-weighted distance) for vector $\mathbf{F}-\hat{\mathbf{F}}$, $\hat{\mathbf{C}}_g$ is the estimated *covariance matrix*, $\hat{\mathbf{C}}_g^{-1}$ is the corresponding *precision matrix*, T is the transpose operator, p is the total number of bands, and

$$||\hat{\mathbf{C}}|| \equiv ||\mathrm{vec}\left(\hat{\mathbf{C}}\right)|| = \mathrm{vec}\left(\hat{\mathbf{C}}\right)^T \mathrm{vec}\left(\hat{\mathbf{C}}\right) = \sum_{i,j}\hat{C}_{ij}^2 \tag{14}$$

is the Euclidean matrix norm where the vec$(\cdot)$ operator "vectorizes" an $n\times m$ matrix by stringing it out into an $nm\times 1$ column vector.

The product of two multivariate Normal distributions $\mathcal{N}(\mathbf{F}|\hat{\mathbf{F}}_g,\hat{\mathbf{C}}_g)$ and $\mathcal{N}(\mathbf{F}|\hat{\mathbf{F}}_h,\hat{\mathbf{C}}_h)$ is a scaled multivariate Normal of the form $S_{gh}\,\mathcal{N}(\mathbf{F}|\hat{\mathbf{F}}_{gh},\hat{\mathbf{C}}_{gh})$ where

$$S_{gh} = \mathcal{N}(\Delta\hat{\mathbf{F}}_{gh}|\mathbf{0},\hat{\mathbf{C}}_{g+h}) \tag{15}$$

$$\Delta\hat{\mathbf{F}}_{gh} \equiv \hat{\mathbf{F}}_g - \hat{\mathbf{F}}_h \tag{16}$$

$$\hat{\mathbf{C}}_{g+h} = \hat{\mathbf{C}}_g + \hat{\mathbf{C}}_h \tag{17}$$

$$\hat{\mathbf{F}}_{gh} = \hat{\mathbf{C}}_{gh}\left(\hat{\mathbf{C}}_g^{-1}\hat{\mathbf{F}}_g + \hat{\mathbf{C}}_h^{-1}\hat{\mathbf{F}}_h\right) \tag{18}$$

$$\hat{\mathbf{C}}_{gh} = \left(\hat{\mathbf{C}}_g^{-1} + \hat{\mathbf{C}}_h^{-1}\right)^{-1} \quad . \tag{19}$$

If we assume a uniform prior on our flux densities[11] $P(\mathbf{F}) = 1$, then our likelihood from equation (11) simplifies enormously, giving us

$$P(g|h) = S_{gh} = \mathcal{N}(\Delta\hat{\mathbf{F}}_{gh}|\mathbf{0},\hat{\mathbf{C}}_{g+h}) \quad . \tag{20}$$

The log-likelihood is then

$$\boxed{-2\ln P(g|h) = ||\Delta\hat{\mathbf{F}}_{gh}||_{\hat{\mathbf{C}}_{g+h}^{-1}} + \delta} \tag{21}$$

where

$$\delta \equiv N_b \ln 2\pi + \ln||\hat{\mathbf{C}}_{g+h}|| \quad , \tag{22}$$

which represents the squared Mahalanobis distance (plus a constant) between $\hat{\mathbf{F}}_g$ and $\hat{\mathbf{F}}_h$.

In the case where $\hat{\mathbf{C}}_g = \mathrm{diag}(\hat{\sigma}_{g,1}^2,\dots,\hat{\sigma}_{g,n}^2)$ and $\hat{\mathbf{C}}_h = \mathrm{diag}(\hat{\sigma}_{h,1}^2,\dots,\hat{\sigma}_{h,n}^2)$ are both diagonal matrices (i.e. $\hat{\mathbf{F}}_g$ and $\hat{\mathbf{F}}_h$ have fully independent errors), this reduces to

$$-2\ln P(g|h) = \sum_{b\in\mathbf{b}} \frac{\left(\hat{F}_{g,b}-\hat{F}_{h,b}\right)^2}{\hat{\sigma}_{g,b}^2+\hat{\sigma}_{h,b}^2} + \delta \quad . \tag{23}$$

[9] SHOULD I INCLUDE THIS PROOF? IT'S JUST GENERALIZING THE BINNED REPRESENTATION FROM LEISTEDT+16 THEN EVALUATING EVALUATING THE PEAK OF THE RESULTING DIRICHLET.

[10] Specific choices of priors might violate this assumption if they use ancillary information, but in general this is true.

[11] This is an uninformative and improper prior linear in $\mathbf{F}$. While the "correct" prior would likely be logarithmic in $\mathbf{F}$, assigning such a prior would render our resulting likelihood non-analytic.

3.2 Color Likelihoods

While (20) is analytic, straightforward to compute, and easily incorporates measurement errors from both g and h, the explicit dependence on an object's magnitude (i.e. observed flux density) requires a training set well-populated across all relevant magnitudes and colors.

As this requirement has historically not been available, most forward modeling-based photo-z approaches instead introduce an additional scale-factor α that modifies the relevant likelihood to instead depend on similar flux *ratios* (i.e. requiring objects to only have similar colors/SED shapes regardless of magnitude) to utilize a limited set of high-resolution bright templates to make predictions to faint objects. This color-based likelihood takes the form

$$\begin{aligned} P_C(g|h) &= \int P(\mathbf{F},\alpha)\mathcal{N}(\mathbf{F}|\hat{\mathbf{F}}_g,\hat{\mathbf{C}}_g)\mathcal{N}(\mathbf{F}|\alpha\hat{\mathbf{F}}_h,\alpha^2\hat{\mathbf{C}}_h)\,d\mathbf{F}d\alpha \\ &= \int P(\alpha)\mathcal{N}\left(\Delta\hat{\mathbf{F}}_{gh}(\alpha)|\mathbf{0},\hat{\mathbf{C}}_{g+h}(\alpha)\right)\,d\alpha \qquad (24) \end{aligned}$$

where

$$\Delta\hat{\mathbf{F}}_{gh}(\alpha) \equiv \hat{\mathbf{F}}_g - \alpha\hat{\mathbf{F}}_h \qquad (25)$$

$$\hat{\mathbf{C}}_{g+h}(\alpha) \equiv \hat{\mathbf{C}}_g + \alpha^2\hat{\mathbf{C}}_h \qquad (26)$$

and we have assumed the joint prior $P(\mathbf{F},\alpha) = P(\mathbf{F})P(\alpha)$ is separable and that $P(\mathbf{F})$ is again uniform.

As with the integral over $\mathbf{F}$, our integral over α will still converge under an (improper) uniform prior. In particular,

$$\lim_{\alpha\to\pm\infty} \propto 1/\alpha^2 = 0, \quad \lim_{\alpha\to 0} = \mathcal{N}(\hat{\mathbf{F}}_g|\mathbf{0},\hat{\mathbf{C}}_g) = \text{constant} \quad ,$$

which shows that our new integrand behaves like a standard Normal for small α but possesses inverse-quadratic rather than exponential tails. Unfortunately, however, this integral has no closed form solution and must instead be solved using numerical methods (see, e.g., Bertsimas et al. 2008).

To avoid this computational overhead, most previous approaches instead assume $\hat{\mathbf{C}}_h = 0$ (e.g., Ilbert et al. 2006; Brammer et al. 2008; Sawicki 2012; Speagle & Eisenstein 2015a,b). Our conditional log-likelihood $-2\ln P_C(g|h,\alpha)$ is then quadratic in α (i.e. a scaled Normal), with

$$\begin{aligned} -2\ln P_C(g|h,\alpha) &= ||\Delta\hat{\mathbf{F}}_{gh}(\alpha)||_{\hat{\mathbf{C}}_g^{-1}} + \delta_c \qquad (27) \\ &= \frac{\alpha_{\rm ML}^2}{\gamma^2}(\alpha'-1)^2 - 2\ln P_C(g|h,\alpha_{\rm ML}), \qquad (28) \end{aligned}$$

where $\alpha' \equiv \alpha/\alpha_{\rm ML}$ is the ML-normalized scalefactor and $\gamma/\alpha_{\rm ML}$ is the associated *shapefactor* (i.e. the "standard deviation"). Solving for $\alpha_{\rm ML}$ and γ, we get

$$\alpha_{\rm ML} = \frac{\gamma^2}{2}\left(\hat{\mathbf{F}}_g^T\hat{\mathbf{C}}_g^{-1}\hat{\mathbf{F}}_h + \hat{\mathbf{F}}_h^T\hat{\mathbf{C}}_g^{-1}\hat{\mathbf{F}}_g\right) \qquad (29)$$

$$\gamma = \left(\hat{\mathbf{F}}_h^T\hat{\mathbf{C}}_g^{-1}\hat{\mathbf{F}}_h^T\right)^{-1/2} . \qquad (30)$$

Marginalizing over α', our color-based likelihood[12] is then

$$P_C(g|h) = \mathcal{N}(\Delta\hat{\mathbf{F}}_{gh}(\alpha_{\rm ML})|\mathbf{0},\hat{\mathbf{C}}_g)\left(\frac{2\pi\gamma^2}{\alpha_{\rm ML}^2}\right)^{\frac{1}{2}} \qquad (31)$$

with log-likelihood

$$\boxed{-2\ln P_C(g|h) = ||\Delta\hat{\mathbf{F}}_{gh}(\alpha_{\rm ML})||_{\hat{\mathbf{C}}_g^{-1}} + \delta_C} \qquad (32)$$

where

$$\delta_C \equiv (N_b - 1)\ln 2\pi + \ln||\hat{\mathbf{C}}_g|| + 2\ln\frac{\alpha_{\rm ML}}{\gamma} \quad . \qquad (33)$$

For independent errors, this reduces to

$$-2\ln P_C(g|h) = \sum_{b\in\mathbf{b}}\frac{\left(\hat{F}_{g,b} - \alpha_{\rm ML}\hat{F}_{h,b}\right)^2}{\hat{\sigma}_{g,b}^2} + \delta_C \quad , \qquad (34)$$

which resembles the more familiar "χ^2 statistic" (plus a constant) often cited in photo-z model-fitting papers.

3.3 Dealing with Missing Data

In practice, our unknown galaxy g and training galaxy h might not be observed in the exact same bands. More formally, this implies that the overlapping bands $\mathbf{b}_{gh}$ between $\mathbf{b}_g$, the bands g is observed in, and $\mathbf{b}_h$, the bands h is observed in, are a subset of $\mathbf{b}$ such that $\mathbf{b}_{gh} \equiv \mathbf{b}_g \cap \mathbf{b}_h \subseteq \mathbf{b}$.

Because our likelihoods can only be computed in overlapping bands, this is equivalent to introducing a $p \times 1$ *band mask* vector $\boldsymbol{\beta}(g|h)$ over our p bands that evaluates to one for $b \in \mathbf{b}_{gh}$ and zero otherwise. Assuming our observed flux densities and associated covariances are always defined over all bands and defining

$$\underline{\hat{\mathbf{C}}}_g^{-1} \equiv \mathbf{B}(g|h) \bullet \hat{\mathbf{C}}_g^{-1} \qquad (35)$$

to be the masked precision matrix given

$$\mathbf{B}(g|h) \equiv [\boldsymbol{\beta}(g|h)]\,[\boldsymbol{\beta}(g|h)]^T \quad , \qquad (36)$$

our original likelihoods can be rewritten as

$$\begin{aligned} -2\ln P(g|h) &= ||\Delta\hat{\mathbf{F}}_{gh}||_{\underline{\hat{\mathbf{C}}}_{g+h}^{-1}} + \underline{\delta} \qquad (37) \\ -2\ln P_C(g|h) &= ||\Delta\hat{\mathbf{F}}_{gh}(\alpha_{\rm ML})||_{\underline{\hat{\mathbf{C}}}_g^{-1}} + \underline{\delta}_C \qquad (38) \end{aligned}$$

where $\underline{\delta}$ and $\underline{\delta}_C$ have also absorbed changes from the inclusion of the band mask.

The changing number of bands introduces several complications into our likelihood calculation. First, it harshly penalizes adding additional dimensions, as both the Mahalanobis distance (the main portion of our likelihood computation) and the normalization (from our multivariate Gaussian) increase with the number of dimensions. This means that, *for the exact same training/target galaxy pair*, our original likelihoods would tend to favor a solution with fewer bands (more masking) than one with more bands (less masking). Intuitively, we expect these should give similarly good fits, so this is clearly undesirable behavior. By extension, this makes it difficult to properly compare likelihoods between training/target galaxy pairs with varying numbers of bands, which are important for properly computing posteriors and population weights.

To correct for this, we need to examine how our likelihood changes as the number of available bands (i.e. the

[12] Our choice to marginalize over α' instead of α, leading to an additional $\alpha_{\rm ML}^2$ term in our likelihood, is to normalize our results to the ML-scale such that objects with similar colors but different magnitudes (and thus different scalefactors α) are considered more or less equally likely.

dimensionality of the fit) changes. In particular, for $\hat{\mathbf{X}}$ distributed as (i.e. drawn from) a d-dimensional multivariate Normal with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, the square Mahalanobis distance between $\hat{\mathbf{X}}$ and $\mathbf{0}$ is distributed as a *non-central chi-square* $\chi^2_p(\lambda = ||\boldsymbol{\mu}||_{\mathbf{C}^{-1}})$ with p degrees of freedom and non-centrality parameter λ. More formally,

$$\hat{\mathbf{X}} \sim \mathcal{N}_p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{C}) \Rightarrow \hat{\mathbf{X}}^T \mathbf{C}^{-1} \hat{\mathbf{X}} \sim \chi^2_p(\lambda = ||\boldsymbol{\mu}||_{\mathbf{C}^{-1}}) \quad . \tag{39}$$

Going back to equation (15), we see that our expression for S_{gh} is simply the likelihood of $\Delta\hat{\mathbf{F}}_{gh}$ assuming $\Delta\hat{\mathbf{F}}_{gh} \sim \mathcal{N}(\Delta\mathbf{F}_{gh}|\mathbf{0}, \hat{\mathbf{C}}_{h+g})$. This implies that

$$||\Delta\hat{\mathbf{F}}_{gh}||_{\underline{\hat{\mathbf{C}}}^{-1}_{g+h}} \sim \chi^2_p(\lambda) \tag{40}$$

$$||\Delta\hat{\mathbf{F}}_{gh}(\alpha_{\rm ML})||_{\underline{\hat{\mathbf{C}}}^{-1}_{g}} \sim \chi^2_{p-1}(\lambda_C) \tag{41}$$

where $p = \sum_{b \in \mathbf{b}} \beta_b(g|h)$ is the number of bands used in the fit (i.e. the dimensionality) and the non-centrality parameters λ and λ_C are unknown.[13] Our original likelihoods are thus distributed as

$$-2 \ln P(g|h) \sim \chi^2_p(\lambda_p + \delta_p) \tag{42}$$

$$-2 \ln P_C(g|h) \sim \chi^2_{p-1}(\lambda_{C,p-1} + \delta_{C,p-1}) \tag{43}$$

where we have added in p subscripts to indicate all of our non-centrality parameters are also (monotonically increasing) functions of the dimensionality p.

To more clearly illustrate the behavior of our likelihoods as a function of dimensionality, we note that the expectation value of our likelihood is

$$E\left[-2 \ln P(g|h)\right] = p + (\lambda_p + \delta_p) \quad , \tag{44}$$

which scales as the number of bands in our fit p both directly (via p) and indirectly (via λ_p and δ_p). Again, this is clearly undesirable since we would like for our likelihoods to be (approximately) band-invariant rather than growing super-linearly in p.

To define "band-invariant" likelihoods, we want to shift our perspective from an absolute to a relative goodness-of-fit metric. In other words, rather than comparing the likelihood of g given h absent any context, we want to ask how much *better* h is relative to the assumption that h is actually a Monte Carlo realization g' of g. This modified log-likelihood (ratio) is then distributed as the convolution of two independent non-central chi-square distributions

$$-2 \ln P(g|h) \sim \chi^2_p(\lambda_p + \delta_p) - \chi^2_p(\delta_p) \tag{45}$$

with expectation value $E\left[-2 \ln P(g|h)\right] = \lambda_p$. This new likelihood is thus an unbiased probe of the true non-centrality (i.e. "mis-centering") of our fit, and for a given λ_p and δ_p we can directly compare our results across bands.

Unfortunately, λ_p is unknown (by construction). While we could theoretically compute

$$P(g|h) = \int P(\lambda_p) P(g|h, \delta_p, \lambda_p)\, d\lambda_p \tag{46}$$

by numerically marginalizing over λ_p, this would break our original analytic formalism. Implementing a first-order correction, however, is extremely straightforward since $E\left[\chi^2_p(\delta_p)\right] = p + \delta_p$. Our new approximate **band-invariant likelihoods** then are

$$\boxed{-2 \ln P(g|h) \equiv ||\Delta\hat{\mathbf{F}}_{gh}||_{\underline{\hat{\mathbf{C}}}^{-1}_{g+h}} - p} \tag{47}$$

in flux density and

$$\boxed{-2 \ln P_C(g|h) \equiv ||\Delta\hat{\mathbf{F}}_{gh}(\alpha_{\rm ML})||_{\underline{\hat{\mathbf{C}}}^{-1}_{g}} - (p-1)} \tag{48}$$

in color.

Our choice to only implement a first-order correction means that although the likelihoods above are band-invariant *on average*, their higher-order moments are not. We find the possible biases this assumption introduces to be negligible during our tests (see §7), but note that this might not hold true in general.

4 SELECTION EFFECTS

In general, we have often *selected* our sample $\mathbf{g}$ through a series of implicit and explicit *selection functions*, which may be different from those used to select our training sample h (cf. Daylan et al. 2016). One concrete instance of this is for applications to weak lensing science for current/upcoming surveys such as HSC and *Euclid*, where (1) the training sample $\mathbf{h}$ is observed at deeper depths than the overall survey and (2) the galaxies $\mathbf{g}$ used for weak lensing-oriented science are often selected to have significantly higher signal-to-noise than the general survey.

Following Leistedt et al. (2016) we can define an $N_g \times 1$ *sample selection mask* vector $\mathbf{s}(\boldsymbol{\mathcal{S}})$ over $\mathbf{g}$ as a function of the selection criteria $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_1, \dots, \mathcal{S}_n\}$ used to construct $\mathbf{g}$. This is analagous to our band mask defined in §3.3, where $s_g = 1$ for $g \in \mathbf{g}$ selected and 0 otherwise. Note that, by design, $\mathbf{s}(\boldsymbol{\mathcal{S}}) = \mathbf{1}$.

Conditioning on $\boldsymbol{\mathcal{S}}$, we are now interested in estimating the posterior conditioned on $s_g = 1$, $P(h|g, s_g = 1, \boldsymbol{\mathcal{S}})$, which is the probability that g is consistent with h given that g has been selected in our sample. Our target posterior is then

$$\boxed{P(h|g, s_g = 1, \boldsymbol{\mathcal{S}}) = \frac{P(s_g = 1|h, \boldsymbol{\mathcal{S}}) P(h|g)}{P(s_g = 1|\boldsymbol{\mathcal{S}})}} \quad , \tag{49}$$

where

$$P(s_g = 1|\boldsymbol{\mathcal{S}}) = \sum_{h \in \mathbf{h}} P(s_g = 1|h, \boldsymbol{\mathcal{S}}) P(h|g) \tag{50}$$

and our *selection probability* $0 \leq P(s_g = 1|h, \boldsymbol{\mathcal{S}}) \leq 1$ is defined as

$$P(s_g = 1|h, \boldsymbol{\mathcal{S}}) = \frac{\int P(s_g = 1|g', \boldsymbol{\mathcal{S}}) P(g'|h)\, dg'}{\int P(g'|h)\, dg'} \quad , \tag{51}$$

where the integral over g' is taken over all possible realizations of g. Note that this integral often cannot be solved analytically and must be computed using numerical methods.

We emphasize that this derivation is completely agnostic to our selection criteria $\boldsymbol{\mathcal{S}}$ and the corresponding selection probability $P(s_g = 1|g', \boldsymbol{\mathcal{S}})$. In particular, there are no requirements that

- $P(s_g = 1|g', \boldsymbol{\mathcal{S}})$ and/or $P(s_g = 1|h, \boldsymbol{\mathcal{S}})$ are analytic,
- $P(s_g = 1|g', \boldsymbol{\mathcal{S}})$ is separable, and/or

[13] The change from d to $d - 1$ comes from removing a degree of freedom when solving for $\alpha_{\rm ML}$.

- $P(s_g = 1|g', \boldsymbol{\mathcal{S}})$ is a strict binary cut.[14]

These specific cases, however, often do arise in practice, and properly exploiting them can dramatically simplify the process of computing our selection probability. We outline several examples using flux density cuts in §4.1 below. For completeness, we also outline a more general case involving color cuts in §4.2.

4.1 Flux Density Cuts

4.1.1 Magnitude Cuts

The simplest types of selection effects are those that operate directly on our flux densities (i.e. our observables). These are most often imposed as a *magnitude cut* corresponding to a specific flux density cutoff $\mathbf{F}_{\rm cut}$, and can be parameterized by

$$P(s_g = 1|g', \mathcal{S}_{\rm mag}) = \mathcal{H}(\hat{\mathbf{F}}'_g - \mathbf{F}_{\rm cut}) \tag{52}$$

where

$$\mathcal{H}(\mathbf{x}) \equiv \prod_{b \in \mathbf{b}} \mathcal{H}(x_b) \tag{53}$$

is the "vectorized" *Heavyside function*, which evaluates to 1 for $\mathbf{x} \geq 0$ element-wise and 0 otherwise. Our selection probability then becomes

$$P(s_g = 1|h, \mathcal{S}_{\rm mag}) = \frac{\int_{\mathbf{F} > \mathbf{F}_{\rm cut}} P(g'|h)\, dg'}{\int P(g'|h)\, dg'} \quad . \tag{54}$$

If we assume we are working with the likelihoods defined via equations (47) or (48), then $P(g'|h)$ simply represents a scaled multivariate Gaussian. Our selection probability then can be seen as the complement of the cumulative distribution function (CDF) of this multivariate Gaussian (i.e. the chance of *not* having our galaxy below our magnitude cut). If we assume that our covariance matrix is diagonal, the complement of the multivariate CDF factors into the product of the complements of the CDFs of univariate Gaussians. These can be parameterized by

$$\begin{aligned}\Phi_c(x|\mu, \sigma^2) &\equiv 1 - \Phi(x|\mu, \sigma^2) \equiv \int_x^{+\infty} \mathcal{N}(x|\mu, \sigma^2) \\ &= 1 - \frac{1}{2}\,\mathrm{erfc}\left(\frac{\mu - x}{\sqrt{2}\sigma}\right) \quad ,\end{aligned} \tag{55}$$

where erfc is the *complementary error function*.

Assuming our likelihoods are defined directly from our flux densities as equation (47), our final expression for the selection probability introduced by $\mathcal{S}_{\rm mag}$ is

$$P(s_g = 1|h, \mathcal{S}_{\rm mag}) = \prod_{b \in \mathbf{b}} \Phi_c\left(F_{{\rm cut},b} | \hat{F}_{h,b}, \hat{\sigma}^2_{g,b} + \hat{\sigma}^2_{h,b}\right) \quad . \tag{56}$$

4.1.2 Signal-to-Noise Cuts

With this function as a baseline, we can easily construct the selection probabilities for a signal-to-noise (S/N) cut. Note that a S/N cut is functionally equivalent to a flux density cut for each individual object, but where the cut is a function of the measurement error. For a S/N threshold X, this gives

$$P(s_g = 1|h, \mathcal{S}_{\rm S/N}) = \prod_{b \in \mathbf{b}} \Phi_c\left(X\hat{\sigma}_{g,b} | \hat{F}_{h,b}, \hat{\sigma}^2_{g,b} + \hat{\sigma}^2_{h,b}\right) \quad . \tag{57}$$

By building on these baseline cases with judicious application of logical statements and the *inclusion-exclusion principle*, more complex cuts can be constructed. For instance, photometric extraction thresholds are typically a logical combination of individual S/N cuts, where photometry for a galaxy is extracted if it is detected above a S/N threshold X in $\geq Y$ bands. In the case $Y = 1$, we note that the probability of detecting a given galaxy with S/N $\geq X$ in *more than* one band is the complement of the probability of observing with S/N $\geq X$ in *no* bands. This gives

$$P(s_g = 1|h, \mathcal{S}_{\rm det}) = 1 - \prod_{b \in \mathbf{b}} \Phi\left(X\hat{\sigma}_{g,b} | \hat{F}_{h,b}, \hat{\sigma}^2_{g,b} + \hat{\sigma}^2_{h,b}\right) \quad . \tag{58}$$

Note that our various selection cuts contained within $\boldsymbol{\mathcal{S}}$ might or might not be fully independent of each other. For instance, if we impose magnitude cuts in two *separate* bands b_1 and b_2, our corresponding selection probabilities are independent and $P(s_g = 1|h, \boldsymbol{\mathcal{S}}) = P(s_g = 1|h, \mathcal{S}_{b_1}) \times P(s_g = 1|h, \mathcal{S}_{b_2})$. However, if we are instead imposing a magnitude and S/N cut in the *same* band b, then our flux density threshold used to compute $P(s_g = 1|h, \boldsymbol{\mathcal{S}})$ instead becomes $F_{\rm eff,b} = \max\{F_{\rm cut,b}, X\hat{\sigma}_{g,b}\}$.

If we are concerned with keeping our selection probability in a (pseudo-)analytic form, thinking carefully about these issues is critical to ensuring our final result is consistent with our original set of selection criteria. As discussed in the next section, however, there always isn't a simple functional form for $P(s_g|h, \mathcal{S})$, at which point these issues become less critical to resolve because we are forced to evaluate our integrals numerically regardless.

4.2 Color Cuts

Outside of magnitude cuts, a common way to select interesting (sub)samples of objects from parent survey populations involves imposing a series of color cuts, which can be interpreted as cuts across a given (unique) collection of *flux density ratios*

$$\hat{\mathbf{R}}_g \equiv \left\{\hat{F}_{g,i}/\hat{F}_{g,j} \;\forall\; j > i\right\} \quad . \tag{59}$$

Our selection function can again be modeled as a Heavyside function evaluated element-wise across $\hat{\mathbf{R}}_g$ such that

$$P(s_g = 1|g', \mathcal{S}_{\rm color}) = \mathcal{H}(\hat{\mathbf{R}}'_g - \mathbf{R}_{\rm cut}), \tag{60}$$

where $\hat{\mathbf{R}}'_g$ is generated from $\hat{\mathbf{F}}'_g$. Unfortunately, there is no analytic form for the corresponding integral over $\mathbf{R}$, requiring us to resort to numerical integration techniques to evaluate $P(s_g = 1|h, \mathcal{S}_{\rm color})$.

[14] A good example of a non-binary selection effect might be where our observables serve only as a (probabilistic) proxy for a cut is being applied via a set of variables we do not (currently) have access to.

The most straightforward numerical integration method here is *importance sampling*. For a series of n Monte Carlo realizations $\left\{\hat{\mathbf{F}}_g^{(1)}, \dots, \hat{\mathbf{F}}_g^{(n)}\right\}$ of the flux density of galaxy g labeled as

$$\left\{g^{(1)}, \dots, g^{(n)}\right\} \sim Q(g|h) \tag{61}$$

drawn from a prior distribution $Q(g|h)$ designed to approximate $P(g|h)$, we can approximate our original integral as

$$\int P(s_g = 1|g', \boldsymbol{\mathcal{S}}) P(g'|h)\, dg'$$
$$\approx \frac{1}{n} \sum_{i=1}^{n} w(g^{(i)}|h) P(s_g = 1|g^{(i)}, \boldsymbol{\mathcal{S}}) \tag{62}$$

where

$$w(g^{(i)}|h) \equiv \frac{P(g^{(i)}|h)}{Q(g^{(i)}|h)} \tag{63}$$

is the associated *importance weight* for the ith sample.

For our flux density and color-based likelihoods defined in equations (47) and (48), we can sample directly from $P(g'|h)$. Our importance weights then become unity such that $w(g^{(1)}|h) = \ldots = w(g^{(N)}|h) = 1$, and our numerical integration becomes a simple counting exercise of how frequently our samples $g^{(i)}$ meet our selection criteria. It is not always possible to sample directly from more complex likelihoods such as, e.g., equation (24), however, and so we keep the full derivation above for generality.

5 HIERARCHICAL INFERENCE

We now want to broaden our individual posteriors from §4 to incorporate population information using a **hierarchical Bayesian model**. In §5.1, we outline a simple case of how to conduct population inference in 1-D given a sample of known redshifts. In §5.2, we broaden our framework to p-dimensional inference over photometric observables including selection effects. In §5.3, we describe our full hierarchical model.

ADD OVERVIEW FIGURE FOR FRAMEWORK+IMPLEMENTATION.

5.1 Population Inference with Redshifts (1-D)

Let's first assume that instead of photometry, for each unknown galaxy $g \in \mathbf{g}$ and training galaxy $h \in \mathbf{h}$ we have a 1-D redshift PDF characterized by $K_g(z|\hat{z}_g)$ and $K_h(z|\hat{z}_h)$, respectively, where $\hat{z}_i$ is the "observed" redshift. For a set of *population weights* $\mathbf{w}$ our redshift *population distribution* can be expressed as

$$P(z|\{\mathbf{h}, \mathbf{w}\}) = \sum_{h \in \mathbf{h}} w_h K_h(z|\hat{z}_h) \Bigg/ \sum_{h \in \mathbf{h}} w_h \tag{64}$$

where $\mathbf{w}$ is an implicit function of (at the minimum) $\mathbf{g}$ and $\mathbf{h}$.

One possible choice for w_h is the Bayesian evidence $\mathcal{Z}_h$ for h computed over all $g \in \mathbf{g}$, defined as

$$\mathcal{Z}_h \equiv \sum_{g \in \mathbf{g}} P(h|g) P'(g) \quad . \tag{65}$$

In this symmetric representation of our original posteriors defined over $\mathbf{h}$, $P'(g)$ no longer represents the evidence from our original fit $P(g)$ but rather a *prior weight* for object g. Defining these prior weights over $\mathbf{g}$ as $\mathbf{p} = \{\ldots, P'(g), \ldots\}$ and our original set of priors over $\mathbf{h}$ as $\mathbf{q} = \{\ldots, P(h), \ldots\}$, our evidence becomes

$$\mathcal{Z}_h \equiv \sum_{g \in \mathbf{g}} p_g \times \frac{q_h P(g|h)}{\sum_{h \in \mathbf{h}} q_h P(g|h)} \tag{66}$$

where our likelihood is now

$$P(g|h) = \int P(z) K_g(z|\hat{z}_g) K_h(z|\hat{z}_h)\, dz \quad . \tag{67}$$

This is a specific case of our generalized likelihood outlined in equation (11) evaluated over a given redshift prior $P(z)$.

This formulation of the Bayesian evidence constitutes a fully probabilistic generalization of the projected number densities typically used in the literature. We can recover the histogrammed redshift number densities used more commonly in the literature by making the following assumptions:

(i) the redshift PDFs $K_g(z|\hat{z}_g) = \delta(z - \hat{z}_g)$ for $g \in \mathbf{g}$ are perfectly measured they are effectively delta functions,
(ii) our prior weights $\mathbf{p}$ over $\mathbf{g}$ are uniform,
(iii) the redshift PDFs $K_h(z|\hat{z}_h) = \Theta\left(z - \hat{z}_h^-\right) \Theta\left(\hat{z}_h^+ - z\right)$ for $h \in \mathbf{h}$ are a series of histogram bins with bin centers $\hat{z}_h$ and widths $\Delta\hat{z}_h$, where $\hat{z}_h^\pm \equiv (\hat{z}_h \pm \Delta\hat{z}_h)/2$,
(iv) our prior $\mathbf{q}$ over $\mathbf{h}$ is uniform, and
(v) our redshift prior $P(z)$ is uniform.

Our evidence then becomes

$$\mathcal{Z}_h = \sum_{g \in \mathbf{g}} \Theta\left(\hat{z}_g - \hat{z}_h^-\right) \Theta\left(\hat{z}_h^+ - \hat{z}_g\right) \quad , \tag{68}$$

which is now equivalent to counting up the number of objects within each $\hat{z}_h^- \leq z < \hat{z}_h^+$ bin.

Note that our redshift kernels $K_g(z|\hat{z}_g)$ defined over $\mathbf{g}$ don't have to strictly be observed PDFs, but can be seen instead as a choice of "smoothing scale(s)" similar to the width of bins in a histogram. A common choice is a Normal distribution where

$$K_g(z|\hat{z}_g) = \mathcal{N}(z|\hat{z}_g, \sigma_g^2), \;\; \sigma_g^2 = (\Delta z)^2 + \hat{\sigma}_g^2 \tag{69}$$

for $\hat{\sigma}_g$ the associated measurement error of $\hat{z}_g$ and Δz a global a smoothing scale. This incorporates known measurement uncertainty while suppressing sampling variation at scales smaller than Δz in order to generate a smoothly varying redshift density estimate.

Since we've now shown that the Bayesian evidence defined in equation (66) is simply the generalization of the typical number density projected over a histogram basis, following (Leistedt et al. 2016) we know that the full posterior distribution for our population weights $P(\mathbf{w}|\{\mathbf{g}, \mathbf{p}\}, \{\mathbf{h}, \mathbf{q}\})$ can be written as a Dirichlet distribution over our $\boldsymbol{\mathcal{Z}} \equiv \{\ldots, \mathcal{Z}_h, \ldots\}$ such that

$$P(\mathbf{w}|\{\mathbf{g}, \mathbf{p}\}, \{\mathbf{h}, \mathbf{q}\}) = \mathrm{Dir}(\mathbf{w}|\boldsymbol{\mathcal{Z}} + 1) \equiv \frac{1}{\mathrm{B}(\boldsymbol{\mathcal{Z}})} \prod_{h \in \mathbf{h}} w_h^{\mathcal{Z}_h} \tag{70}$$

where

$$\mathrm{B}(\boldsymbol{\mathcal{Z}}) = \frac{\prod_{h \in \mathbf{h}} \Gamma(\mathcal{Z}_h)}{\Gamma\left(\sum_{h \in \mathbf{h}} \mathcal{Z}_h\right)} \tag{71}$$

is the multivariate Beta function and $\Gamma(\cdot)$ is the Gamma function.

To understand this result more intuitively, note that the Dirichlet distribution is the conjugate prior of the multinomial distribution (i.e. the product of a multinomial and a Dirichlet is Dirichlet). In other words, if we "pick" a random galaxy $g \in \mathbf{g}$ and want to guess which particular $h \in \mathbf{h}$ out of N_h possible options it is associated with, where each h has a corresponding probability proportional to $\mathcal{Z}_h$, then the resulting distribution for the fractional counts over $\mathbf{h}$ will be Dirichlet.

The joint distribution $P(z, \mathbf{w}|\{\mathbf{g},\mathbf{p}\},\{\mathbf{h},\mathbf{q}\})$ between z and $\mathbf{w}$ conditioned on $\{\mathbf{g},\mathbf{p}\}$ and $\{\mathbf{h},\mathbf{q},\mathbf{w}$ is then simply

$$\begin{aligned} P(z,\mathbf{w}|\{\mathbf{g},\mathbf{p}\},\{\mathbf{h},\mathbf{q}\}) &= P(z|\{\mathbf{g},\mathbf{p}\},\{\mathbf{h},\mathbf{q},\mathbf{w}\}) \\ &\quad \times P(\mathbf{w}|\{\mathbf{g},\mathbf{p}\},\{\mathbf{h},\mathbf{q}\}) \\ \Rightarrow P(z,\mathbf{w}|\{\mathbf{h},\boldsymbol{\mathcal{Z}}\}) &= P(z|\{\mathbf{h},\mathbf{w}\})P(\mathbf{w}|\boldsymbol{\mathcal{Z}}) \quad . \end{aligned} \tag{72}$$

We can't probe this distribution directly, but can sample from it in a straightforward manner by first drawing a set of population weights $\mathbf{w}' \sim P(\mathbf{w}|\boldsymbol{\mathcal{Z}})$ and then computing $P(z|\{\mathbf{h},\mathbf{w}'\})$. Since the Dirichlet distribution is symmetric with respect to $\boldsymbol{\mathcal{Z}}$, marginalizing over $\mathbf{w}$ is identical to taking the maximum-likelihood (ML) estimate $P(z|\{\mathbf{h},\mathbf{w}=\boldsymbol{\mathcal{Z}}\})$.

5.2 Population Inference with Photometry (p-D)

While §5.1 is an instructive ideal 1-D case, in practice $\hat{\mathbf{z}}_g$ is unknown. Instead, we must infer $P(z|\{\mathbf{g},\mathbf{p},\mathbf{s}\},\{\mathbf{h},\mathbf{q},\mathbf{w}\},\boldsymbol{\mathcal{S}})$ using a set of p observed flux density PDFs $K_g(\mathbf{F}|\hat{\mathbf{F}}_g)$ for $g \in \mathbf{g}$ and a corresponding set of PDFs in flux density $K_h(\mathbf{F}|\hat{\mathbf{F}}_h)$ and redshift $K_h(z|\hat{z}_h)$ for $h \in \mathbf{h}$.

Our redshift population distribution can again be expressed as a linear combination of the redshift kernels defined over our training set following equation (64). However, the evidence

$$\begin{aligned} \mathcal{Z}_h &= \sum_{g\in\mathbf{g}} P(h|g, s_g = 1, \boldsymbol{\mathcal{S}})P(g) \\ &= \sum_{g\in\mathbf{g}} p_g \times \frac{q_h P(s_g=1|h,\boldsymbol{\mathcal{S}})P(g|h)}{\sum_{h\in\mathbf{h}} q_h P(s_g=1|h,\boldsymbol{\mathcal{S}})P(g|h)} \end{aligned} \tag{73}$$

now incorporates selection effects $\boldsymbol{\mathcal{S}}$ and $P(g|h)$ is now defined in terms of the observed flux densities. The photometric likelihood $P(g|h)$ can be computed following equations (47) or (48) while our selection probability $P(s_g|h,\boldsymbol{\mathcal{S}})$ can be computed following equation (51). As before, the full posterior distribution for our population weights follow a Dirichlet distribution over $\boldsymbol{\mathcal{Z}}$ and we can sample from it by drawing $\mathbf{w}' \sim P(\mathbf{w}|\boldsymbol{\mathcal{Z}})$ and then computing $P(z|\{\mathbf{h},\mathbf{w}'\})$.

While in the 1-D redshift case we demonstrated the freedom to specify our basis over $\mathbf{h}$ as a histogram in z, binned representations become progressively more intractable as the dimensionality p increases. Instead, it becomes more useful to consider $\mathbf{h}$ more generally as a set of training objects with an associated set of multivariate Normal flux density kernels $K_h(\mathbf{F}|\hat{\mathbf{F}}_h)$ following equation (12). Unlike for a p-D histogram, where the number of bins (and hence evaluations) for each object scales as $\mathcal{O}(n^p)$, the number of evaluations required over such a discrete basis remains unchanged (N_h) as a function of p.

5.3 Hierarchical Inference with Photometry

In §5.1 and §5.2, we focused exclusively on population inference conditioned on $\{\mathbf{g},\mathbf{s},\mathbf{p}\}$ and $\{\mathbf{h},\mathbf{q}\}$ under the influence of selection effects $\boldsymbol{\mathcal{S}}$. We showed that this reduces down to inference over our Bayesian evidences $\boldsymbol{\mathcal{Z}}$ and that sampling from the joint distribution $P(z,\mathbf{w}|\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{S}})$ is functionally equivalent to mapping samples drawn from $P(\mathbf{w_h}|\boldsymbol{\mathcal{Z}})$ to redshift via $P(z|\{\mathbf{h},\mathbf{w}\})$.

In most cases, we are interested in not only inferring the general population distribution for all objects but also redshift PDFs for individual objects. We are thus interested in probing the full joint distribution

$$P(z,\mathbf{w},\boldsymbol{\mathcal{Z}}|\mathbf{h}) = P(z|\{\mathbf{h},\mathbf{w}\})P(\mathbf{w},\boldsymbol{\mathcal{Z}}) \quad . \tag{74}$$

Since $P(z|\{\mathbf{h},\mathbf{w}\})$ is a deterministic function, we are really interested in directly sampling from the joint distribution $P(\mathbf{w},\boldsymbol{\mathcal{Z}})$ over $\mathbf{h}$. While sampling from the joint distribution directly using MCMC methods with rejection sampling-based updates (e.g., standard Metropolis-Hastings updates; Metropolis et al. 1953; Hastings 1970) is difficult, the conditional distributions can be easily sampled using Gibbs updates (Geman & Geman 1984). In that case, our chain draws samples by looping over

$$\begin{aligned} &\mathbf{w}^{(i)},\boldsymbol{\mathcal{Z}}^{(i)} \sim P(\mathbf{w},\boldsymbol{\mathcal{Z}}) \\ &\qquad \Rightarrow \begin{cases} \mathbf{w}^{(i)} \sim P(\mathbf{w}|\boldsymbol{\mathcal{Z}}^{(i-1)}) \\ \boldsymbol{\mathcal{Z}}^{(i)} \sim P(\boldsymbol{\mathcal{Z}}|\mathbf{w}^{(i)}) \end{cases} . \end{aligned} \tag{75}$$

As above, our population weights $\mathbf{w_h}^{(i)}$ can be sampled from a Dirichlet distribution as

$$\mathbf{w}^{(i)} \sim P(\mathbf{w}|\boldsymbol{\mathcal{Z}}^{(i-1)}) = \mathrm{Dir}(\mathbf{w}|\boldsymbol{\mathcal{Z}}^{(i-1)}+1) \quad .$$

Our "population-weighted" evidence $\boldsymbol{\mathcal{Z}}^{(i-1)}$ over $\mathbf{h}$ is actually a deterministic function of $\mathbf{w_h}$, where

$$\mathcal{Z}_h^{(i)} = \sum_{g\in\mathbf{g}} p_g \times \frac{w_h^{(i)} q_h P(s_g=1|h,\boldsymbol{\mathcal{S}})P(g|h)}{\sum_{h\in\mathbf{h}} w_h^{(i)} q_h P(s_g=1|h,\boldsymbol{\mathcal{S}})P(g|h)} \quad . \tag{76}$$

Under this definition, we can examine how our previous population-oriented models outlined in §5.1 and §5.2 fit into our hierarchical model. There, we originally estimate our evidences $\boldsymbol{\mathcal{Z}}$ assuming $\mathbf{w}$ is uniform, and then subsequently use $\boldsymbol{\mathcal{Z}}$ to estimate $\mathbf{w}$. Our population models thus represent the first iteration sampling from our hierarchical model as outlined in equation (75).

In addition, our population-oriented sampling scheme gives us individual redshift posteriors "for free". For a sample of population weights $\mathbf{w}^{(i)}$, our redshift PDF for galaxy $g \in \mathbf{g}$ is simply

$$\begin{aligned} P(z|g,\mathbf{w}^{(i)},\boldsymbol{\mathcal{S}}) &= \sum_{h\in\mathbf{h}} P(z|h)P(h|g,s_g=1,w_h^{(i)},\boldsymbol{\mathcal{S}}) \\ &= \sum_{h\in\mathbf{h}} K_h(z|\hat{z}_h) \times \frac{w_h^{(i)} q_h P(s_g=1|h,\boldsymbol{\mathcal{S}})P(g|h)}{\sum_{h\in\mathbf{h}} w_h^{(i)} q_h P(s_g=1|h,\boldsymbol{\mathcal{S}})P(g|h)} \quad . \end{aligned} \tag{77}$$

This represents a weighted sum of the redshift kernels $K_h(z|\hat{z}_h)$ over $\mathbf{h}$, where our weights represent the product of our prior p_h, population weight $w_h^{(i)}$, likelihood $P(g|h)$, and selection probability $P(h,s_g|g,\boldsymbol{\mathcal{S}})$.

As our population case, this represents a (weighted) mapping from flux density to redshift using a collection of posteriors $P(h, s_g|g, w_h, \boldsymbol{\mathcal{S}})$ conditioned on $\mathbf{w_h}$ over $\mathbf{h}$ given g. It is straightforward to show that the combined set $\{\mathbf{w}^{(1)}, P(h|g, s_g = 1, w_h^{(1)}, \boldsymbol{\mathcal{S}}), \dots, \mathbf{w}^{(n)}, P(h|g, s_g = 1, w_h^{(n)}, \boldsymbol{\mathcal{S}}),\}$ of population weights and associated posteriors are equivalent to sampling from the relevant joint distribution for each individual object $g \in \mathbf{g}$.

ADD PLATE DIAGRAM FIGURE OUTLINING HIERARCHICAL MODEL.

6 SPARSE LIKELIHOODS WITH MACHINE LEARNING

To properly evaluate $\boldsymbol{\mathcal{Z}}$ for a given sample $\mathbf{w}^{(i)}$, we have to (re-)evaluate our posteriors over all $g \in \mathbf{g}$ and $h \in \mathbf{h}$. While by construction we want to compute our posteriors $P(h|g, s_g = 1, \boldsymbol{\mathcal{S}})$ for every target galaxy g, properly evaluating this over every training galaxy h (and over every posterior sample $\mathbf{w}^{(i)}$) is clearly computationally limiting in practice. Furthermore, this brute-force approach penalizes us for utilizing larger, well-sampled training sets by scaling as $\mathcal{O}(N_h)$.

What we would like to do is compute approximate $\overline{P}(g|h)$, sparse $\underline{P}(g|h)$, or approximate *and* sparse $\overline{\underline{P}}(g|h)$ representations of the likelihood $P(g|h)$ that can be computed in significantly less than $\mathcal{O}(N_h)$ time.

Using a typical machine learning algorithm $\mathcal{M}$, for a target galaxy g we can compute an $N_h \times 1$ *neighbor selection vector* $\boldsymbol{\omega}(g|\mathcal{M})$ over our N_h training galaxies for a target galaxy g. Like with our sample selection vector $\mathbf{s}$, $\omega_h = 1$ indicates galaxy h has been selected as a neighbor of galaxy g and $\omega_h = 0$ indicates otherwise.

Over this (relatively) small subset of neighbors, most machine learning algorithms approximate the likelihood as $\overline{P}(g|h)$ using, e.g., a uniform distribution. This gives us

$$\overline{\underline{P}}(g|h, \mathcal{M}) = \overline{P}(g|h)\omega_h(g|\mathcal{M}) \quad , \tag{78}$$

which evaluates to 0 for all non-neighboring objects and $\overline{P}(g|h)$ for neighboring ones. Using an approximation $\overline{P}(h)$ for the prior (most often uniform) and defining

$$\mathbf{H}_g \equiv \{h : \omega_h(g|\mathcal{M}) > 0\} \tag{79}$$

to be our subset of neighbors, our redshift prediction can then be approximated as

$$P(z|g) \approx \sum_{H \in \mathbf{H}_g} K_H(z|\hat{z}_H) \times \frac{\overline{P}(g|H)\overline{P}(H)}{\sum_{H \in \mathbf{H}_g} \overline{P}(g|H)\overline{P}(H)} \quad . \tag{80}$$

Note that we have ignored selection effects here, which are rarely accounted for in most machine learning methods.

6.1 Individual Example: Decision Trees

Before outlining our specific approach, we first want to gain intuition of how some common machine learning methods operate under our framework. Here we focus on the use of *decision trees* (Breiman et al. 1984), which have proved popular for photo-z estimation (Gerdes et al. 2010; Carrasco Kind & Brunner 2013; Hoyle et al. 2015).

A decision tree works as follows:

- the set of neighbors defined by $\boldsymbol{\omega}(g|\mathcal{M})$ are selected by successively partitioning the data space in a way that minimizes the variance in the associated label(s) (here redshift) until some pre-determined stopping criterion is reached,
- the likelihood $\overline{P}(g|h)$ is approximated as uniform over the selected objects, and
- the training data is assumed to be sampled directly from the underlying distribution such that $\overline{P}(h)$ is uniform across $\mathbf{h}$.

Redshift KDEs for each training galaxy are assumed to be delta functions (i.e. point estimates) such that $K_h(z|\hat{z}_h) = \delta(z = \hat{z}_h)$. Defining $N_{\mathbf{H}_g} = \sum_{h \in \mathbf{h}} \omega_h(g|\mathcal{M})$ to be the number of neighboring objects, a decision tree then produces the following redshift estimate

$$P(z|g, \mathcal{M}) \approx \frac{1}{N_{\mathbf{H}_g}} \sum_{h \in \mathbf{h}} \delta(z = \hat{z}_h)\omega_h(g|\mathcal{M}) \tag{81}$$

for each galaxy g. This is equivalent to selecting an unweighted collection of the neighboring $\hat{z}_h$'s.

While this expression represents the underlying redshift PDF fundamentally generated from a decision tree, in practice most "out of the box" solvers (such as from, e.g., `scikit-learn`; Pedregosa et al. 2011) do not return this object-level information. Instead, they return the average

$$z_{\rm avg}(g|\mathcal{M}) = \int z P(z|g, \mathcal{M}) dz \approx \frac{1}{N_{\mathbf{H}}} \sum_{H \in \mathbf{H}_g} \hat{z}_H \quad . \tag{82}$$

6.2 Ensemble Extension: Random Forests

Often, an *ensemble* of estimators is used to generate an improved estimate. Our above formulation for a single decision tree can be easily generalized to an ensemble of decision trees (i.e. a *random forest*; Breiman 1996, 2001). In that case, inference over our ensemble of individual learners $\boldsymbol{\mathcal{M}} = \{\dots, \mathcal{M}, \dots\}$ trained over some combination of Monte Carlo, bootstrap resampling, and/or dimensional subsampling of $\mathbf{h}$ can locate the same neighbor multiple times.

Under this condition, our neighborhood selection vector now becomes a set of *neighborhood weights* that incorporate measurement error and sample variance that can be interpreted as a data-driven "poor man's likelihood". Our ensemble redshift estimate thus becomes

$$z_{\rm avg}(g|\boldsymbol{\mathcal{M}}) \approx \frac{1}{N_{\mathbf{H}_g}} \sum_{H \in \mathbf{H}_g} \omega_H(g|\boldsymbol{\mathcal{M}}) \times \hat{z}_H \tag{83}$$

where $N_{\mathbf{H}_g}$ is now the *effective* number of neighboring objects.

6.3 Ensemble Application: *k*-*d* Trees

While we are interested in generating sparse approximations to our posteriors similar to the example outlined above, we actually have access to much more information than is typically available. In particular, we can specify almost all parts of our inference problem explicitly in the context of a hierarchical Bayesian model. Our goal is thus to use machine learning in a targeted way to generate a sparse and/or approximate representation of our posterior while keeping as much of the Bayesian formalism as possible.

More explicitly, defining our $N_h \times 1$ *likelihood vector* $\boldsymbol{\mathcal{L}}_g$, *selection probability vector* $\boldsymbol{\rho}_g$, and *posterior vector* $\boldsymbol{\Psi}_g$ as

$$\boldsymbol{\mathcal{L}}_g \equiv \{\dots, P(g|h), \dots : h \in \mathbf{h}\} \tag{84}$$

$$\boldsymbol{\rho}_g \equiv \{\dots, P(s_g = 1|h, \boldsymbol{\mathcal{S}}), \dots : h \in \mathbf{h}\} \tag{85}$$

$$\boldsymbol{\Psi}_g \equiv \{\dots, P(h|g, s_g = 1, w_h, \boldsymbol{\mathcal{S}}), \dots : h \in \mathbf{h}\} \tag{86}$$

to be the set of elements for a given row g over all columns $h \in \mathbf{h}$, we can write our posterior samples $\boldsymbol{\Psi}_g^{(i)}$ for g given $\mathbf{w}^{(i)}$ as

$$\boldsymbol{\Psi}_g^{(i)} = \frac{\mathbf{w}^{(i)} \bullet \mathbf{q} \bullet \boldsymbol{\rho}_g \bullet \boldsymbol{\mathcal{L}}_g}{\langle \mathbf{w}^{(i)}, \mathbf{q} \bullet \boldsymbol{\rho}_g \bullet \boldsymbol{\mathcal{L}}_g \rangle} \tag{87}$$

where $\langle \mathbf{w}^{(i)}, \mathbf{q} \bullet \boldsymbol{\rho}_g \bullet \boldsymbol{\mathcal{L}}_g \rangle$ is the *inner product* (also called dot product) between $\mathbf{w}^{(i)}$ and $(\mathbf{q} \bullet \boldsymbol{\rho}_g \bullet \boldsymbol{\mathcal{L}}_g)$ and $\mathbf{q}$ is again our prior vector over $\mathbf{h}$. In this form, it is straightforward to see that a sparse estimate of any individual component of our expression ($\mathbf{w}^{(i)}$, $\mathbf{q}$, $\boldsymbol{\rho}_g$ or $\boldsymbol{\mathcal{L}}_g$) will automatically translate to a sparse estimate of our posterior $\boldsymbol{\Psi}_g^{(i)}$.

We choose to use machine learning to impose a sparse *but exact* representation of our likelihood over $\mathbf{H}_g$ as

$$\underline{\boldsymbol{\mathcal{L}}}_{\mathbf{H}_g} \equiv \{\dots, P(g|H), \dots : H \in \mathbf{H}_g\} \quad , \tag{88}$$

where $\mathbf{H}_g$ again includes all objects with non-zero neighborhood weights $w_h(g|\boldsymbol{\mathcal{M}}) > 0$. Our posterior then immediately becomes

$$\underline{\boldsymbol{\Psi}}_{\mathbf{H}_g}^{(i)} = \frac{\underline{\mathbf{w}}_{\mathbf{H}_g}^{(i)} \bullet \underline{\mathbf{q}}_{\mathbf{H}_g} \bullet \underline{\boldsymbol{\rho}}_{\mathbf{H}_g} \bullet \underline{\boldsymbol{\mathcal{L}}}_{\mathbf{H}_g}}{\left\langle \underline{\mathbf{w}}_{\mathbf{H}_g}^{(i)}, \underline{\mathbf{q}}_{\mathbf{H}_g} \bullet \underline{\boldsymbol{\rho}}_{\mathbf{H}_g} \bullet \underline{\boldsymbol{\mathcal{L}}}_{\mathbf{H}_g} \right\rangle} \quad , \tag{89}$$

which only have to be evaluated over $N_{\mathbf{H}_g}$ objects instead of N_h.

To ensure our set of neighbors leads to an unbiased approximation, we utilize an ensemble of $N_{\rm trees}$ *k-d* trees (Bentley 1975) to avoid utilizing any label (redshift) information when selecting our neighbors. Each tree in our ensemble is trained on a Monte Carlo realization $\mathbf{h}'$ of $\mathbf{h}$ and the nearest $N_{\rm nghbr}$ neighbors of a Monte Carlo realization g' of g are selected. By construction, this generates an unbiased collection of $N_{\mathbf{H}_g} \leq N_{\rm trees} \times N_{\rm nghbr} \ll N_h$ unique neighbors that properly incorporates measurement errors contributed from both $\mathbf{h}$ and g.

For $N_h \gg 2^p$ given p bands (i.e. with $\mathbf{h}$ large and p small-to-moderate), a *k-d* tree can locate a set of neighbors in $\mathcal{O}(\log N_h)$ time. This represents a significant speed up if $N_{\mathbf{H}_g} \ll N_h$ since the majority of our original model is spent computing likelihoods and posteriors.

The drawback to this computationally expedient approximation is that it degrades the predictions of any individual object: while the approximation is unbiased *on average* and works well for objects above moderate S/N, for a low-S/N object a sparse approximation utilizing a small number of Monte Carlo realizations (via $N_{\rm trees}$) is almost certainly not adequate. In addition, we're more likely to undersample the tails of the PDF (both in redshift and flux density) for any individual object and thus underestimate (or miss entirely) rare but certain "catastrophic failures" that are bound to occur in large samples.

7 TESTS

A bunch of dN/dz tests, possibly supplemented with some photo-z accuracy tests (but that's not what we're aiming for).

7.1 Mocks

We generate a bunch of cool mocks because we are cool people. These have HSC specs for photometry, COSMOS templates (no dust), BPZ priors. Added in data censoring and magnitude bias in training data.

7.2 SDSS

Real tests with SDSS? here

8 DISCUSSION

Filler text. Filler text.

9 CONCLUSION

Filler text. Filler text.

ACKNOWLEDGEMENTS

JSS would like to thank Tansu Daylan, Ben Hoyle, Song Huang, Harshil Kamdar, Jae Hyeon Lee, and Alex I. Malz for insightful discussions. JSS is grateful for financial support from the CREST program, which is funded by the Japan Science and Technology (JST) Agency. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. XXX. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

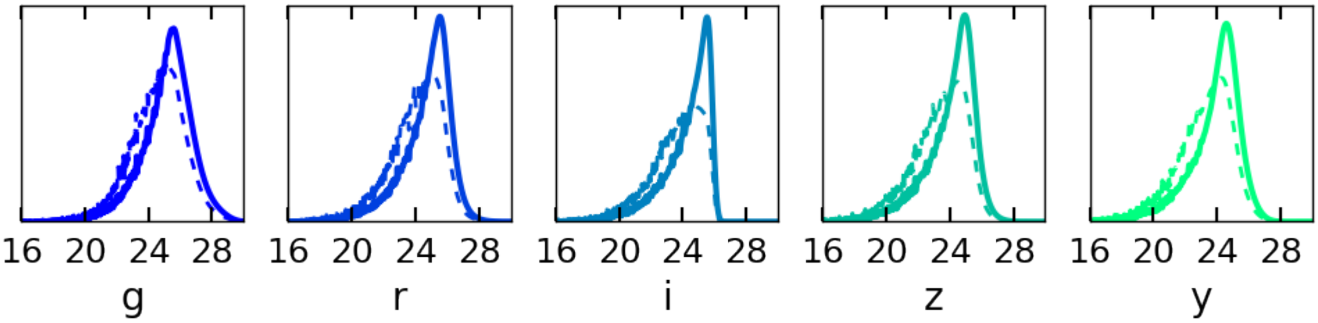

Figure 1. Normalized asinh magnitude distribution of selected galaxies. Training set (dashed), testing set (solid).

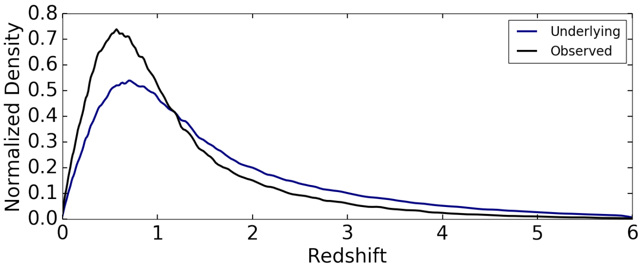

Figure 2. Redshift distribution of the parent sample (black) and observed sample (blue).

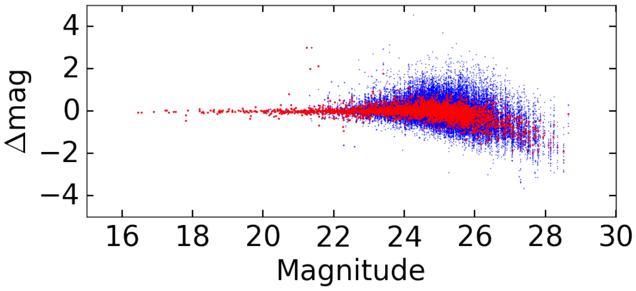

Figure 3. Imputed photometry (converted to asinh magnitudes). Random draws are shown in blue, median quantities in red.

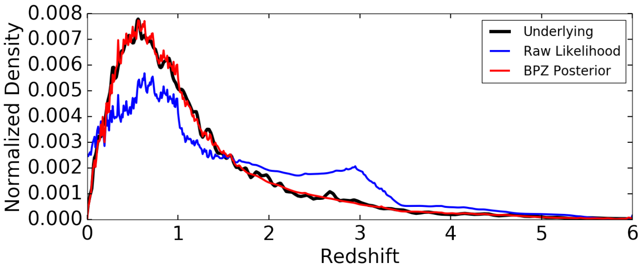

Figure 4. Redshift distribution derived through template-fitting (noiseless model photometry). The raw likelihoods are shown in blue, the BPZ-computed posteriors (no mismatch) is in red.

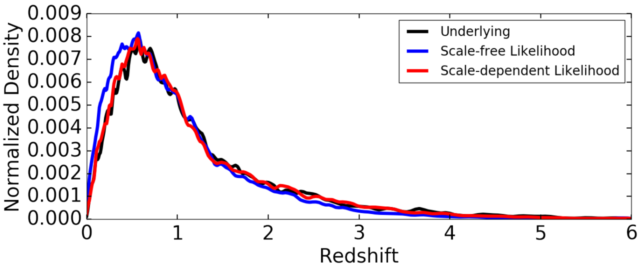

Figure 5. Redshift distribution derived from noisy training data assuming a scale-free likelihood (typically used; blue) and a scale-dependent likelihood (red).

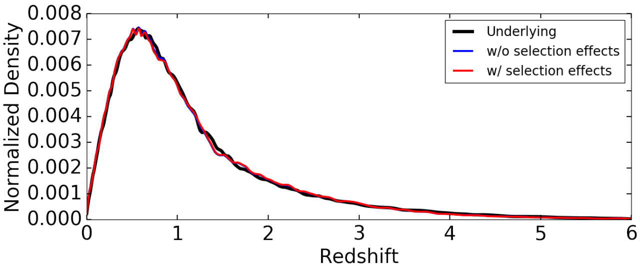

Figure 6. Redshift distributions derived from FRANKEN-Z with (red) and without (blue) corrections for selection effects. Effects are minor because the training/testing sets are selected from the same parent sample with similar noise properties. Can elaborate on later.

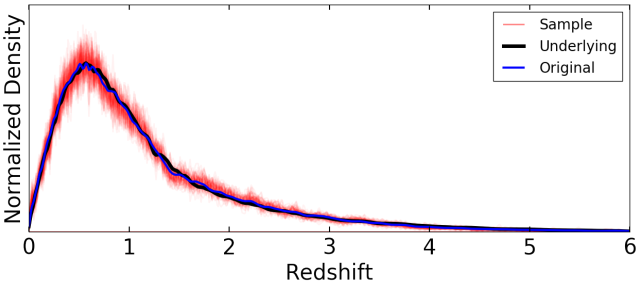

Figure 7. Redshift distributions derived from FRANKEN-Z after including Hierarchical Bayesian modeling over the training set. Samples from the distribution are shown in red (original in blue).

References

Acquaviva V., Gawiser E., Guaita L., 2011, ApJ, 737, 47

Almosallam I. A., Jarvis M. J., Roberts S. J., 2016, MNRAS, 462, 726

Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540

Benítez N., 2000, ApJ, 536, 571

Bentley J. L., 1975, Commun. ACM, 18, 509

Bertsimas D., Doan X. V., Lasserre J., 2008, Operations Research Letters, 36, 205

Bishop C. M., 2006, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA

Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, 476

Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503

Breiman L., 1996, Machine Learning, 24, 123

Breiman L., 2001, Machine Learning, 45, 5

Breiman L., Friedman J., Stone C. J., Olshen R. A., 1984, Classification and regression trees. CRC press

Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483

Carrasco Kind M., Brunner R. J., 2014, MNRAS, 442, 3380

Cool R. J., et al., 2013, ApJ, 767, 118

Dahlen T., et al., 2013, ApJ, 775, 93

Daylan T., Portillo S. K. N., Finkbeiner D. P., 2016, preprint, (arXiv:1607.04637)

Elliott J., de Souza R. S., Krone-Martins A., Cameron E., Ishida E. E. O., Hilbe J., 2016, The Universe of Digital Sky Surveys, 42, 91

Geman S., Geman D., 1984, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 721

Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823

Hastings W., 1970, Biometrika, 57, 97

Hildebrandt H., et al., 2010, A&A, 523, A31

Hogg D. W., Bovy J., Lang D., 2010, preprint, (arXiv:1008.4686)

Hoyle B., 2016, Astronomy and Computing, 16, 34

Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015, MNRAS, 449, 1275

Ilbert O., et al., 2006, A&A, 457, 841

Ivezic Z., et al., 2008, preprint, (arXiv:0805.2366)

Johnson S. P., Wilson G. W., Tang Y., Scott K. S., 2013, MNRAS, 436, 2535

Laureijs R., et al., 2011, preprint, (arXiv:1110.3193)

Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258

Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118

Masters D., et al., 2015, ApJ, 813, 53

Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, J. Chem. Phys., 21, 1087

Miyazaki S., et al., 2012, in Ground-based and Airborne Instrumentation for Astronomy IV. p. 84460Z, doi:10.1117/12.926844

Moustakas J., et al., 2013, ApJ, 767, 50

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Sánchez C., et al., 2014, MNRAS, 445, 1482

Sawicki M., 2012, PASP, 124, 1208

Sheldon E. S., Cunha C. E., Mandelbaum R., Brinkmann J., Weaver B. A., 2012, ApJS, 201, 32

Speagle J. S., Eisenstein D. J., 2015a, preprint, (arXiv:1510.08073)

Speagle J. S., Eisenstein D. J., 2015b, preprint, (arXiv:1510.08080)

Speagle J. S., Capak P. L., Eisenstein D. J., Masters D. C., Steinhardt C. L., 2016, MNRAS, 461, 3432

Spergel D., et al., 2015, preprint, (arXiv:1503.03757)

Tanaka M., 2015, ApJ, 801, 20

The Dark Energy Survey Collaboration 2005, ArXiv Astrophysics e-prints,

de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, Experimental Astronomy, 35, 25

This paper has been typeset from a TEX/LATEX file prepared by the author.