Draft version March 10, 2015
Preprint typeset using LaTeX style emulateapj v. 12/16/11

# PROBABILISTIC REDSHIFT DISTRIBUTION: MINIMAL APPROACH

A.I. Malz[1]

*Draft version March 10, 2015*

## ABSTRACT

This paper answers the question of how one would calculate the redshift distribution function $\mathcal{N}(z)$ from a set of likelihood functions for the photometric redshifts of individual galaxies.

*Subject headings:* photo-z

## 1. INTRODUCTION

We would like to learn the redshift distribution function $\mathcal{N}(z)$ for a set of $J$ galaxies $j$. We assume each galaxy has a known likelihood function $p(\vec{d_j}|z)$ for the observed data $\vec{d_j}$ (a set of magnitudes in each of several filters) over redshift $z$. The full dataset of $\{\vec{d_j}\}_{j=1,...,J}$ will be denoted as $\underline{D}$. The redshift distribution function may be expressed as Eq. 1.

$$p(z|\mathcal{N}) = \mathcal{N}(z) \equiv \frac{\mathcal{N}|_z}{\int \mathcal{N}|_z dz} \tag{1}$$

The likelihood function for the redshift distribution function is given in Eq. 2, where the likelihoods for each galaxy's redshift are considered to be independent.

$$p(\underline{D}|\mathcal{N}) = \prod_{j=1}^{J} \int p(\vec{d_j}|z)\ p(z|\mathcal{N}) dz \tag{2}$$

By Bayes' Rule, we may find the desired posterior according to Eq. 3. We want the posterior for direct comparison with prior work in the literature, but the likelihood of Eq. 2 is in fact preferable.

$$p(\mathcal{N}|\underline{D}) = \frac{p(\underline{D}|\mathcal{N})p(\mathcal{N})}{p(\underline{D})} \tag{3}$$

It is generally considered difficult to calculate the posterior $p(\mathcal{N}|\underline{D})$ directly due to not knowing $p(\underline{D})$. Instead, we may sample the desired distribution using Monte Carlo-Markov chain (MCMC) methods.

## 2. METHOD

### 2.1. *Probabilistic Model*

We consider the $K = 35$ redshift bins $B_k = [z^B_{k-1}, z^B_k]$ for which Sheldon, et al. (2011) calculated posteriors for the redshift of each galaxy based on observations of the apparent magnitude in the five photmetric filters of SDSS. We parametrize $\mathcal{N}(z)$ as a vector of histogram heights $\vec{\mathcal{N}}$ representing the probability that a galaxy's redshift lies within the corresponding bin. Eq. 1 becomes Eq. 4, although we actually work with the log of $\vec{\mathcal{N}}$ throughout.

$$p(B_k|\vec{\mathcal{N}}) = \mathcal{N}_k \tag{4}$$

### 2.2. *Fake Data*

In order to simulate data, we must select a true redshift function $\vec{\mathcal{N}}^0$ from some prior evaluated at the center of each bin $z_k = (z^B_{k+1} - z^B_k)/2$. In this case, we choose the prior $p(\vec{\mathcal{N}})$ to be a multivariate Gaussian distribution satisfying Eq. 5. This covariance structure is chosen to enforce continuity of the histogram heights, to better approximate a potentially realistic $\mathcal{N}(z)$. IT SHOULD BE NOTED THAT SMOOTHNESS IMPROVES IF A MORE BROADLY CORRELATED COVARIANCE MATRIX IS USED.

$$p(\ln[\vec{\mathcal{N}}]) = N(\ln[\vec{\mu}], \underline{\Sigma}) \tag{5}$$

$$\mu_k = \frac{z_k^{z^B_K} \exp[-z_k * z^B_1]}{\sum_{k=0}^{K} z_k^{z^B_K} \exp[-z_k * z^B_1]}$$

aimalz@nyu.edu
[1] CCPP

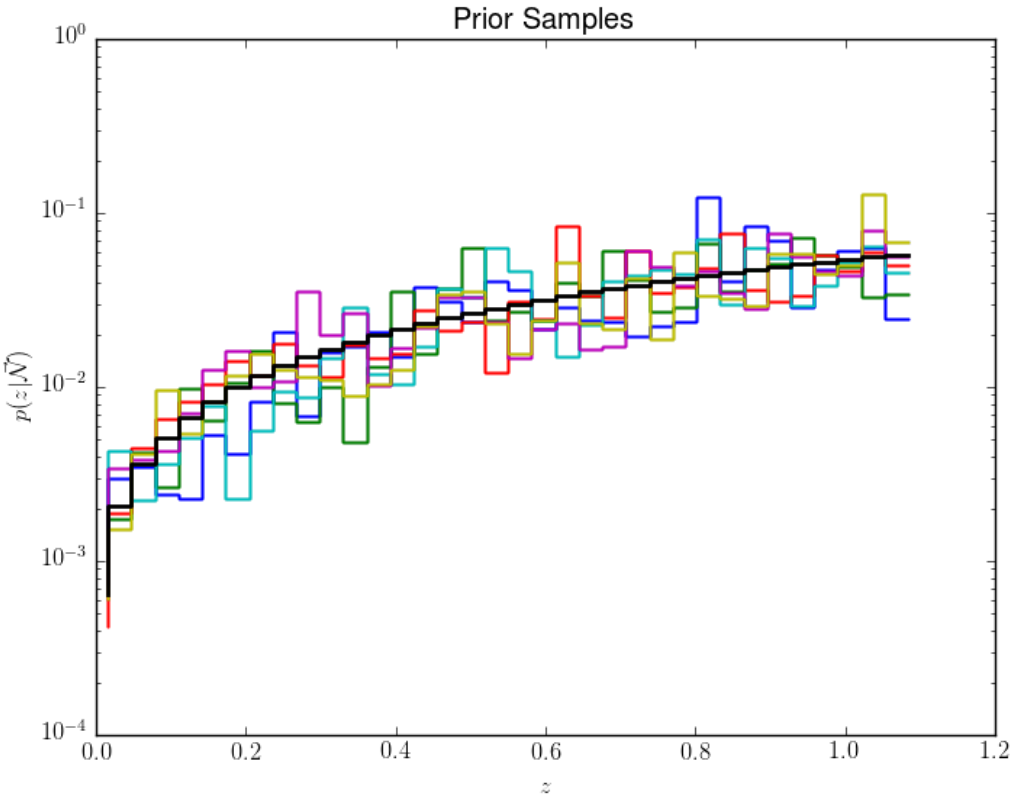


FIG. 1.— Several random samples of $\vec{\mathcal{N}}$ from the prior distribution of Eq. 5 are shown here. The mean of the prior is shown in black.

$$\Sigma_{ij} = \begin{cases} i = j & K^{-1/2} \\ |i-j| = 1 & K^{-1} \\ else & 0 \end{cases}$$

Some examples of samples from the prior are shown in Fig. 2.2.

Next, we generate galaxy redshift likelihood functions as follows. We first assign a bin number $b_j = k$ from $k = 1, \ldots, K$ to each of $J = 1000$ galaxies by randomly sampling the $K$ bins with weights given by the prior as $p(b_j) \sim \mathcal{N}_k^0$. We then assign each galaxy a true redshift $z_j$ chosen uniformly from within the bin $B_{b_j}$ to which it was assigned. Here it is convenient to define $\beta$ as the mean redshift per bin.

The true redshift of each galaxy is shifted by a random error according to Eq. 6 to simulate inaccuracy in measurements, yielding a shifted redshift $z'_j$ for each galaxy.

$$\begin{aligned} z'_j &= z_j + e_j \\ p(e_j) &= N(0, \beta^2) \\ \beta &= \frac{\sum_{b_j=1}^{K} z_{b_j} - z_{b_j-1}}{K} \end{aligned} \tag{6}$$

The discretized likelihood function $p(\vec{d_j}|B_k)$ for each galaxy is taken to be Gaussian with a mean equal to the shifted redshift and a standard deviation proportional to the square of the true redshift to simulate the fact that uncertainty increases with redshift. Thus the observed probability that galaxy $j$ has a redshift in bin $k$ is given by Eq. 7. As a final step, all likelihoods are normalized such that they sum to unity. Fig. 2.2 shows a few examples of simulated likelihoods.

$$\mathcal{L}_j^k \equiv p(\vec{d_j}|B_k) = \int_{z_{k-1}}^{z_k} \frac{1}{\sqrt{2\pi(\sqrt{\beta}z_j^2)^2}} \exp\left[-\frac{(z'_j - \tilde{z})^2}{2(\sqrt{\beta}z_j^2)^2}\right] d\tilde{z} \tag{7}$$

There are some other physically motivated modifications of this process that would be valuable to explore in future

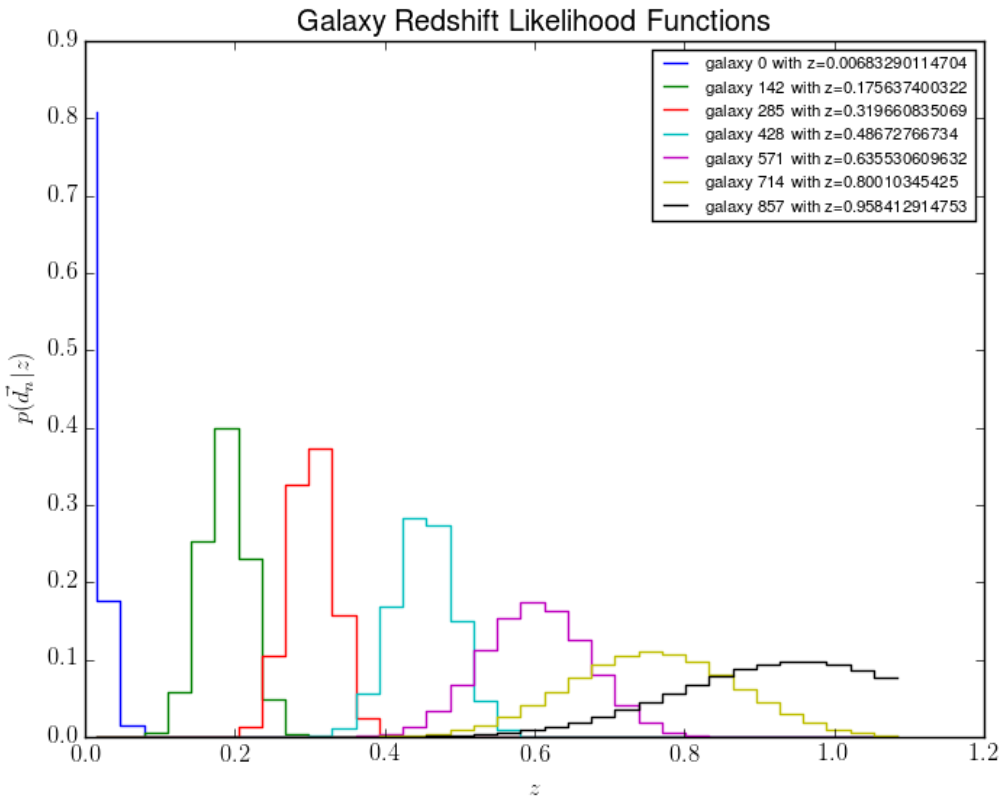


FIG. 2.— Several random redshift likelihood functions are shown here. Note that the width of the Gaussian increases with redshift.

work. In particular, a multimodal distribution replacing the Gaussian shape of the likelihoods would better simulate the degeneracies present in photometric redshift estimates.

## 3. RESULTS

One instance of $\vec{\mathcal{N}}$ was selected at random as a true value. Samples of true galaxy redshifts were taken from that distribution and transformed into likelihood functions according to the procedure of Sec. 2.2. Fig. 3 shows histograms of the fake data generated. It is worth noting that any method aiming to calculate the posterior can do no better than the "observed redshifts" that go into generating the data.

The Metropolis-Hastings algorithm is applied to sample the posterior. The procedure is initialized with the log of the average probability (i.e. a flat distribution), which shall be denoted $\ln \vec{\mathcal{N}} \equiv \ln p(z|\vec{\mathcal{N}})$. The log of the numerator of the posterior of Eq. 3 is calculated according to Eq. 8 and is denoted as $\ln \tilde{p}(\vec{\mathcal{N}}|\underline{D})$. The following pseudo-code outlines the algorithm.

1. Randomly sample $\ln \tilde{p}(\vec{\mathcal{N}}|\underline{D})$ to generate $\ln \vec{\mathcal{N}}'$.
2. Calculate the log of the numerator of the posterior as in Eq. 8 to produce $\ln \tilde{p}(\vec{\mathcal{N}}'|\underline{D})$.
3. Calculate $r = \ln \tilde{p}(\vec{\mathcal{N}}'|\underline{D}) - \ln \tilde{p}(\vec{\mathcal{N}}|\underline{D})$.
4. If $r \geq 0$, set and record $\ln \vec{\mathcal{N}} = \ln \vec{\mathcal{N}}'$; recalculate $\ln \tilde{p}(\vec{\mathcal{N}}|\underline{D})$.
   If $r < 0$, select a random number $n$ from the uniform distribution between 0 and 1.
   (a) If $n < e^r$, set and record $\vec{\mathcal{N}} = \vec{\mathcal{N}}'$; recalculate $\ln \tilde{p}(\vec{\mathcal{N}}|\underline{D})$.
5. Check if the threshold has been achieved; if not, return to Step 1.

$$\ln \tilde{p}(\vec{\mathcal{N}}'|\underline{D}) \equiv \ln p(\vec{\mathcal{N}}) + \sum_{j=1}^{J} \ln \left[ \sum_{k=1}^{K} p(\vec{d}_j|z_k)\ p(z_k|\vec{\mathcal{N}}) \right] \propto \ln p(\vec{\mathcal{N}}|\underline{D}) \tag{8}$$

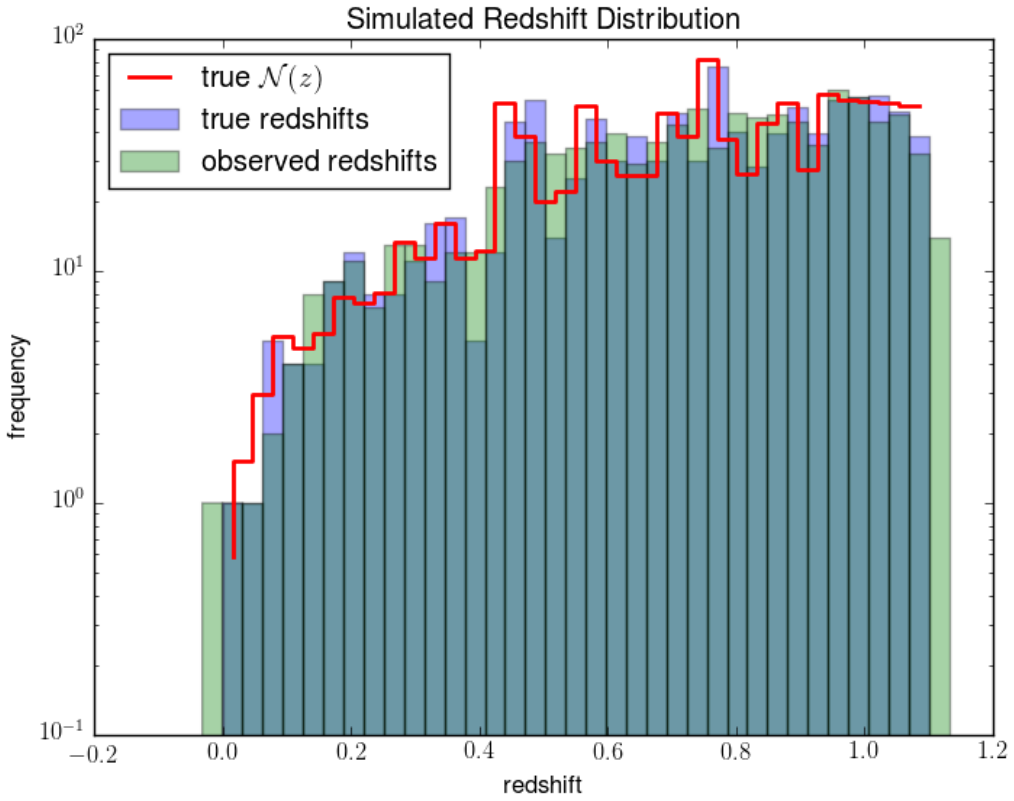


FIG. 3.— The true $\vec{\mathcal{N}}$ chosen for this test is plotted here, along with histograms of the true redshifts sampled from it and the shifted redshifts resulting from the fake data generation procedure.

Here, the threshold was an arbitrary $R = 10000$ iterations of the algorithm. All accepted proposals from one instance of the code are shown in Fig. **??**. The acceptance fraction was $\sim 0.1\%$ for this and other runs. Since 10000 iterations likely doesn't even get through the "burn-in" period of the algorithm, this acceptance fraction is not surprising! If I were convinced it were otherwise valid, I would run it until some convergence criterion were achieved. However, since that might take quite some time, I will conservatively refrain from doing so.

## 4. DISCUSSION

It is desirable to compare this result to what would have been obtained by the method of Sheldon, et al. (2011), which directly calculates the posterior for the entire dataset using the posteriors for each galaxy according to Eq. 9.

$$p(B_k|\vec{\mathcal{N}}) = \sum_{n=1}^{N} p(B_k|\vec{d}_j) \tag{9}$$

To do this, I calculate the posteriors $p(z|\vec{d}_j)$ for each galaxy using Eq. 10, the product of the estimate of $\vec{\mathcal{N}}$ and the likelihood for each galaxy. This is done for all accepted values of $\vec{\mathcal{N}}$.

$$p(B_k|\vec{d}_j) = p(\vec{d}_j|B_k)p(\vec{\mathcal{N}}|\underline{D}) \tag{10}$$

Fig. **??** compares the result of summing the posteriors as in Eq. 9 with the result of the MCMC solutions of Eq. 3. The method of Sheldon, et al. (2011) underestimates the probability of observing low redshifts. As one would expect, the MCMC estimate irreversibly loses some substructure because of the shifting error added to the simulated data.

## REFERENCES

Benitez, N., ApJ 536:571-583, 2000 June 20
Cunha, C.E., Lima, M., Oyaizu, H., Frieman, J., Lin, H., arxiv:0810.2991
Fadely, R., Hogg, D.W., Willman, B., arxiv:1206.4306
Foreman-Mackey, D., Hogg, D.W., and Morton, T.D., arxiv:1406.3020
Hogg, D.W., arxiv:astro-ph/9905116
Hogg, D.W., Myers, A.D., Bovy, J., arxiv:1008.4146
Hogg, D.W., arxiv:1205.4446
Lima, M., Cunha, C.E., Oyaizu, H., Frieman, J., Lin, H., Sheldon, E.S., MNRAS, 390, 118

Sheldon, E.S., Cunha, C., Mandelbaum, R., Brinkmann, J., Weaver, B.A., arxiv:1109.5192

FILL IN MORE OF THESE!