Draft version April 23, 2015
Preprint typeset using LaTeX style emulateapj v. 12/16/11

# PROBABILISTIC REDSHIFT DISTRIBUTION: MINIMAL APPROACH

A.I. Malz[1]

*Draft version April 23, 2015*

## ABSTRACT

This paper outlines a method for calculating the redshift distribution function $N(z)$ from a set of posteriors for the photometric redshifts of individual galaxies.

*Subject headings:* photo-z

## 1. INTRODUCTION

The redshift distribution function $N(z)$ indicates the expected number of galaxies at each redshift.

## 2. METHOD

### 2.1. *Probabilistic Model*

We would like to learn the redshift distribution function $N(z)$ describing the number of galaxies at each redshift. Redshifts are parameters we would like to estimate, and the parameters $N$ determining $N(z)$ may be considered hyperparameters that determine the probability of a redshift occurring. The redshift distribution function may be expressed as Eq. **??**.

$$p(z|N) = \frac{N(z)}{\int N(z)\ dz} = \frac{N(z)}{E[N]} \tag{1}$$

Let us consider a galaxy $j$ whose data $\vec{d_j}$ is a set of magnitudes in each of several filters and seek the likelihood of observing the data under the model $N(z)$. We may express this likelihood $p(\vec{d_j}|N)$ in terms of the likelihood $p(\vec{d_j}|z)$ as in Eq. **??**. However, it is generally considered impossible to calculate $p(\vec{d_j}|z)$ from $\vec{d_j}$. It is easy to find the posterior probability $p(\vec{d_j}|z, N_0)$ of a galaxy's photometry given its redshift and an assumed redshift distribution function $N_0(z)$ serving as an uninformative prior. This has been done by **?**, for example. We may use this posterior to find the likelihood function $p(\vec{d_j}|N)$ for the observed data $\vec{d_j}$ under an arbitrary redshift distribution function $N(z)$, as in Eq. **??**. (**?**)

$$\begin{aligned} p(\vec{d_j}|N) &= \int\ p(\vec{d_j}|z)\ p(z|N)\ dz \qquad (2)\\ &= \int\ p(\vec{d_j}|z)\ p(z|N)\ \frac{p(z|\vec{d_j}, N_0)}{p(z|\vec{d_j}, N_0)}\ dz \\ &= \int\ p(\vec{d_j}|z)\ p(z|N)\ \frac{p(z|\vec{d_j}, N_0)\ p(\vec{d_j}|N_0)}{p(\vec{d_j}|z)\ p(z|N_0)}\ dz \\ &= \int\ p(z|N)\ p(z|\vec{d_j}, N_0)\ \frac{p(\vec{d_j}|N_0)}{p(z|N_0)}\ dz \\ &\propto \int\ p(z|\vec{d_j}, N_0)\ \frac{p(z|N)}{p(z|N_0)}\ dz \qquad (3) \end{aligned}$$

Now let us consider an ensemble of $J$ galaxies. The full dataset of $\{\vec{d_j}\}_{j=1,...,J}$ will be denoted as $\underline{D}$. The likelihood function for the redshift distribution function is given in Eq. **??**, where the likelihoods for each galaxy's redshift are assumed to be independent.

$$p(\underline{D}|N) = \prod_{j=1}^{J}\ p(\vec{d_j}|N) \tag{4}$$

aimalz@nyu.edu
[1] CCPP

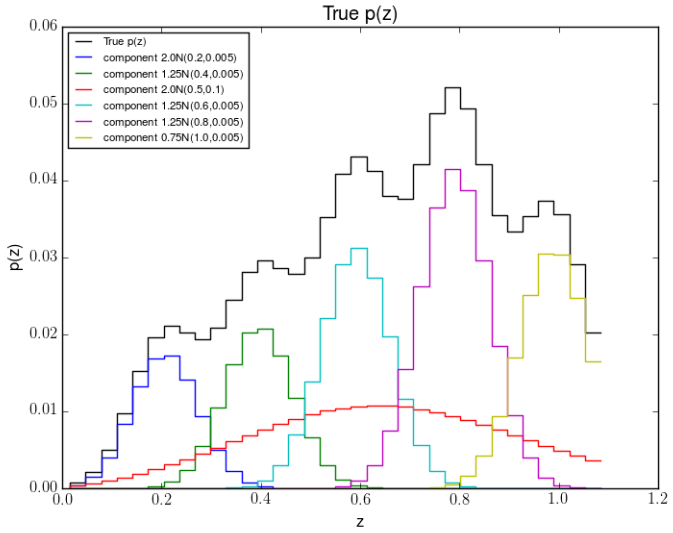


FIG. 1.— We construct $p^0$ as a sum of Gaussians.

Here, we would also like to calculate the posterior $p(N|\underline{D})$ for direct comparison with prior work in the literature. The posterior may be calculated via Eq. **??** in terms of the prior distribution $p(N)$. Since we cannot know $p(\underline{D})$, we sample the desired distribution using Monte Carlo-Markov chain (MCMC) methods.

$$p(N|\underline{D}) = \frac{p(\underline{D}|N)p(N)}{p(\underline{D})} \tag{5}$$

### 2.2. *Fake Data*

We consider the $K = 35$ redshift bins $B_k = [z^B_{k-1}, z^B_k]$ for which **?** calculated posteriors for the redshift of each galaxy based on observations of the apparent magnitude in the five photmetric filters of SDSS. We shall take $\vec{\theta}$ to be a discretized parametrization of $N(z)$, where $\theta_k$ is the average value of $N(z)$ over the redshift range of bin $B_k$. Thus the expected number of galaxies in each bin is $J_k = \theta_k \Delta_k$, where $\Delta_k \equiv z_{k+1} - z_k$, transforming Eq. **??** into Eq. **??** (although we actually work with the log of $\vec{\theta}$ throughout).

$$p(B_k) \equiv \frac{J_k}{J} = \frac{\theta_k \Delta_k}{J} \tag{6}$$

In order to simulate data, we select set a true, physically motivated redshift probability distribution $p^0(z)$, where $p(z)$ corresponds to $N(z)$ for $J = 1$. The convolution of a linear function and a sum of Gaussians is quite general and accurately generates features observed in the true $N(z)$. (Cite this!) The instance of $p^0(z)$ here is shown in Eq. **??**, where the constant $C_c$ indicates the relative amplitude of the Gaussian component centered at $z_c$ with variance $\sigma_c^2$. The linear function is evaluated at the centers of the bins $\bar{z}_k = (z_{k+1} - z_k)/2$ and is included to ensure that $\lim_{z\to 0} N(z) = 0$. The construction of this probability density is illustrated in Fig. **??**.

$$p^0(z_k) = \sum_{c=1}^{6} \bar{z}_k C_c \int_{z_k}^{z_{k+1}} \mathcal{N}(z_c, \sigma_c^2) dz \tag{7}$$

We evaluate $p^0(z)$ over each bin to obtain $\vec{\theta}^0$ according to Eq. **??**, whose construction is illustrated in Fig. **??**. The number of galaxies $J$ is selected from a Poisson distribution centered at an arbitrary $J_0 = 10,000$ galaxies.

$$\theta_k \equiv J \int_{z_k}^{z_{k+1}} p^0(z) dz \tag{8}$$

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 galaxy $j$ by randomly sampling the $K$ bins with weights given by the true redshift function as $\int_{z_k}^{z_{k+1}} p^0(z) dz$. Examples of some such samples for a fixed $J = 1,000$ are shown in Fig. **??**.

We then assign each galaxy a true redshift $z_j$ chosen uniformly from within the bin $B_{b_j}$ to which it was assigned. The true redshift of each galaxy is shifted by a random error according to Eq. **??** to simulate inaccuracy in measurements,

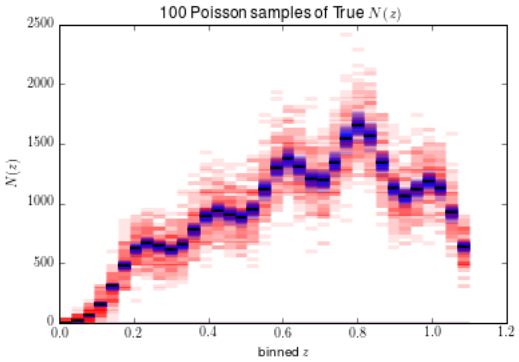


FIG. 2.— Samples of the number of galaxies in each bin are generated as Poisson draws.

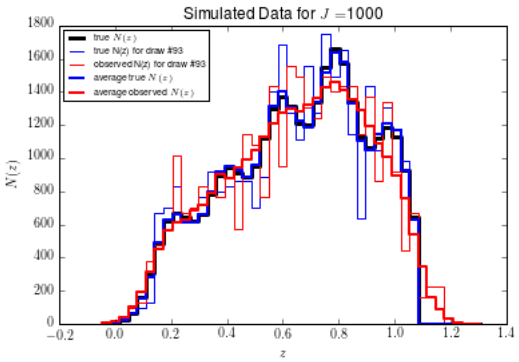


FIG. 3.— The true $\vec{\theta}$ chosen for this test is plotted here, along with histograms of the true redshifts sampled from it and the shifted redshifts resulting from the fake data generation procedure.

yielding a shifted redshift $z'_j$ for each galaxy. Fig. **??** shows histograms of the fake data generated. It is worth noting that any method aiming to calculate the posterior should be unable to do better than the "observed redshifts" that go into generating the data. SOMEHOW THIS IS NOT TRUE FOR THIS DATA, BUT HOW?

$$\begin{aligned} z'_j &= z_j + e_j \\ p(e_j) &= \mathcal{N}(0, \delta_j) \\ \delta_j &= \frac{\sum_{b_k=1}^{K} (z_{b_k} - z_{b_k-1})}{K}(1+z_j) \end{aligned} \tag{9}$$

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 $1+z_j$ to simulate the fact that uncertainty increases with redshift. According to the above, "the data" are thus comprised of $J$ pairs $(z'_j, \delta_j)$. The observed probability that galaxy $j$ has a redshift in bin $k$ is given by Eq. **??**. As a final step, all likelihoods are normalized such that they integrate to unity over the redshift range spanned by the bins. Fig. **??** shows a few examples of simulated likelihoods.

$$p(\vec{d}_j|B_k) = \int_{z_k}^{z_{k+1}} \frac{1}{\sqrt{2\pi\delta_j^2}} \exp\left[-\frac{(z'_j - z)^2}{2\delta_j^2}\right] dz \tag{10}$$

The likelihood function for the entire dataset given in Eq. **??** may be re-expressed as Eq. **??** in terms of $N(z)$ rather than $p(z)$. It is valuable to verify that the likelihood is maximized for the true $N(z)$ that generated a particular set of simulated data.

$$\begin{aligned} p(\underline{D}|N(z)) &= \exp\left[-\int N(z)dz\right] \prod_{j=1}^{J} \int p(\vec{d}_j|z)N(z)dz \\ p(\underline{D}|\vec{\theta}) &= \exp\left[-\sum_{k=1}^{K} \theta_k \Delta_k\right] \prod_{j=1}^{J} \sum_{k=1}^{K} p(\vec{d}_j|B_k)\theta_k\Delta_k \end{aligned} \tag{11}$$

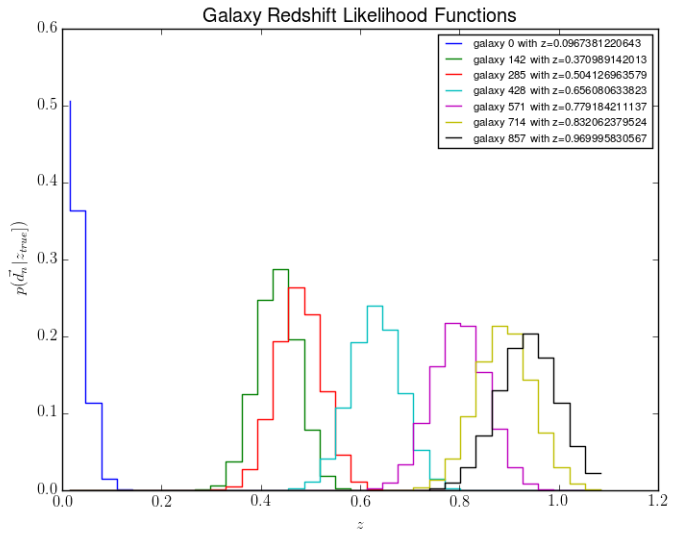


FIG. 4.— Several random redshift likelihood functions are shown here. Note that the width of the Gaussian increases with redshift.

$$\ln p(\underline{D}|\vec{\theta}) = -\sum_{k=1}^{K} \exp\left[\ln\theta_k + \ln\Delta_k\right] + \sum_{j=1}^{J} \ln\left[\sum_{k=1}^{K} \exp\left[\ln p(\vec{d}_j|B_k) + \ln\theta_k + \ln\Delta_k\right]\right]$$

## 3. RESULTS

The Metropolis-Hastings algorithm is applied to sample the posterior. The procedure is initialized with the log of the average probability $\ln\theta_k^0 = -\ln K \;\forall\; k$ (i.e. the flat distribution), which shall be denoted $\ln\vec{\theta} \equiv \ln p(z)$. The log of the numerator of the posterior of Eq. **??** is calculated according to Eq. **??** and is denoted as $\ln\tilde{p}(\vec{\theta}|\underline{D})$.

$$\ln\tilde{p}(\vec{\theta'}|\underline{D}) \equiv \ln p(\vec{\theta'}) + \sum_{j=1}^{J} \ln\left[\sum_{k=1}^{K} p(\vec{d}_j|z_k)\; p(z_k|\vec{\theta'})\right] \propto \ln p(\vec{\theta'}|\underline{D}) \tag{12}$$

$$\tag{13}$$

At each iteration $i$, proposal distribution $\vec{\theta^i}$ is generated from the multivariate normal distribution shown in Eq. **??**. This covariance structure is chosen to enforce continuity of the histogram heights. Some examples of samples from the initialization value are shown in Fig. **??**.

$$\begin{aligned} \ln\vec{\theta^i} &= \mathcal{N}(\ln\vec{\theta^{i-1}}, \underline{\Sigma}) \\ \Sigma_{kk'} &= q\exp^{-\frac{a(k-k')^2}{2}} \end{aligned} \tag{14}$$

The following pseudo-code outlines the algorithm.

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

Here, the threshold was accepting $R = 100$ proposal distributions. 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.

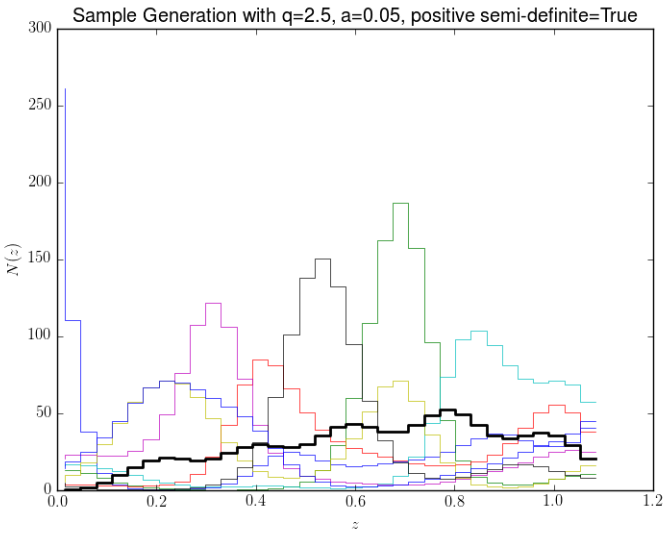


FIG. 5.— Several random samples of $\vec{\theta}$ from the distribution of Eq. **??** are shown here.

## 4. DISCUSSION

It is desirable to compare this result to what would have been obtained by the method of **?**, which directly calculates the posterior for the entire dataset using the posteriors for each galaxy according to Eq. **??**.

$$p(B_k) = \sum_{r=1}^{R} p_r(B_k|\vec{d}_j) \tag{15}$$

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

$$p_r(B_k|\vec{d}_j) = p(\vec{d}_j|B_k)p(\vec{\theta}^r|\underline{D}) \tag{16}$$

Fig. **??** compares the result of summing the posteriors as in Eq. **??** with the result of the MCMC solutions of Eq. **??**. The method of **?** 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
Marshall, P., github.com/drphilmarshall/Pangloss/issues/23
Sheldon, E.S., Cunha, C., Mandelbaum, R., Brinkmann, J., Weaver, B.A., arxiv:1109.5192
FILL IN MORE OF THESE!

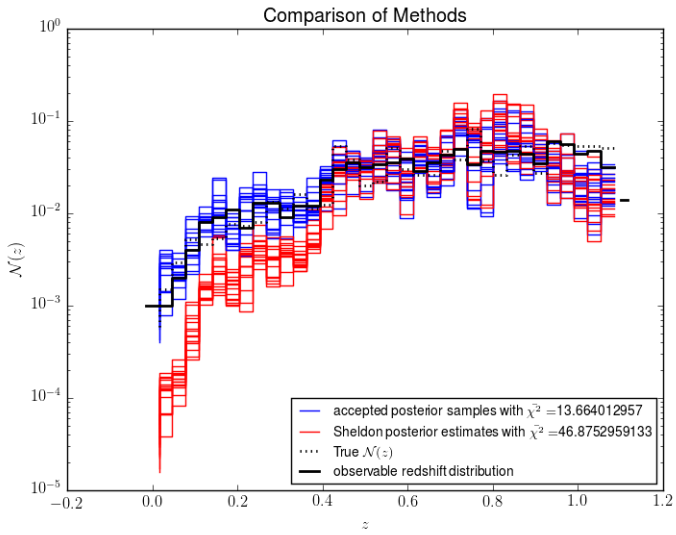


FIG. 6.— The result of applying Eq. **??** is shown in red, the average accepted posterior sample from the method presented here is shown in blue, and $p(z)$ for the observable redshifts of Eq. **??** is shown in black. The sum of squared differences between the result of each method and the true value are also shown; one can see that the **?** approach has larger errors.