$p(z)$ TO $N(z)$

These are a few notes on redshift PDF summation.

1. Definitions

Some definitions so I don't confuse them later

- a: a vector of parameters describing the true, noiseless observable properties of a galaxy; if you know a, you know everything there is to know about the galaxy, including its redshift
- z: true redshift of a galaxy
- d: the data vector of a galaxy in some photometric survey; includes photometric measurements and their uncertainties

All of these come with a subscript i for galaxy i in the sky, and there is a set of all possible a defined as A.

In a Bayesian hierarchical scheme, there is a vector of hyperparameters $h \in H$ that determines $p(a)$ and because we don't know the true h the two always appear together as $p(a, h)$. For now, assume we know the true H.

2. Individual galaxy $p(z_i)$

Let's look at a single galaxy i and estimate its redshift based on photometric measurements d_i in the survey, which is what, for instance, BPZ intends to do, i.e. calculate $p(z_i|d_i)$. The following is a simple equality:

$$p(z_i|d_i, H) = \int_A p(z_i|a)p(a|d_i, H)\, da = \int_A p(z_i|a)p(a, H)p(d_i|a)/p(d_i)\, da \tag{1}$$

Of these

- $p(z_i|a)$ is simply $\delta(z_i - z(a))$ since we said z was fully determined by a
- $p(a, H)$ only depends on the Universe (i.e., the value that the hyperparameters happen to take)
- $p(d_i|a)$ is very simple in the limit of noiseless d_i (a multi-dimensional δ function, because here d_i is equal to some subset of true observable properties a_i) and depends on the survey for noisy d_i
- $p(d_i)$, the distribution of *observed* parameters in our survey, depends a lot on the survey (because there is e.g. a spatially dependent magnitude limit), but it doesn't influence $p(z_i|d_i, H)$ because it doesn't depend on redshift and is fixed by normalization anyway ($1 = \int_{\mathbb{R}} p(z_i|d_i)\, dz_i$).

The following statements about $p(z_i|d_i, H)$ hold:

Date: October 4, 2016.

- $p(z_i|d_i, H)$ gets wider in z_i if our survey gets worse. It does so because $p(d_i|a_i)$ changes
- If we had many equally looking galaxies i with $d_i = d\ \forall i$ then $p(z_i|d_i, H)$ is the same for all i. The true redshifts of these galaxies are all different, and are actually distributed as $p(z_i|d_i, H)$.
- The latter requires that the galaxies are not just equally looking but *fully representative* of the set of galaxies that are equally looking. Galaxies that look like d_i *but are selected by some properties that are not part of* d_i (typically the case for spectroscopic surveys, where the success rate of redshift determination depends on whether or not a galaxy has a bright emission line), you'll get a different $p(z)$ (see Daniel's draft).

3. Ensemble $p_D(z)$

Consider, for a change, the redshift distribution of an ensemble of galaxies. Assume that we have selected these galaxies by observational properties, e.g. we take all galaxies whose $d_i \in D$, where D is some interval of outcomes of our photometric measurements. Call the true redshift distribution of this set of galaxies $p_D(z|H)$. The true redshift distribution is what you would get if you measure the true redshifts of all galaxies with $d \in D$ and bin up the results.

The following is an equality

$$p_D(z|H) = p(z|d \in D, H) = \int_D p(z_i|d_i, H) p(d_i|H)/p(d \in D|H)\ dd_i \tag{2}$$

The $p(d|H)/p(d \in D|H)$ is just to keep things normalized. It's the PDF of d for galaxies in D given hyperparameters H. In the limit of many galaxies $i = 1, ..., N$ that sample this PDF,

$$\int_D p(z_i|d_i, H) p(d_i|H)/p(d \in D|H)\ dd_i \to \frac{1}{N} \sum_{i=1}^{N} p(z_i|d_i, H)\ , \tag{3}$$

so it's fine to sum the individual PDFs.

4. If you are uncertain about H

So far, we assumed that we know $p(a)$ exactly, or equivalently that we know some vector of hyperparameters H exactly that determines $p(a|H)$. This is the endgame of galaxy evolution. What if we don't know H, we only know some $p(a|h)$ and $p(h)$? Then eqn. (1) becomes

$$p(z_i|d_i) = \int_A \int_H p(z_i|a) p(a|d_i, h) p(h)\ da\ dh = \int_A \int_H p(z_i|a) p(a|h) p(h) p(d_i|a, h)/p(d_i)\ da\ dh\ . \tag{4}$$

Because $p(a, h) = p(a|h)p(h)$ and $p(d_i|a, h)$ depend on h this PDF will actually be broader than the distribution of z that we would get if we binned up the true redshifts of all galaxies that look like d_i. Likewise, the stacked PDF will be broader than the actual PDF on an ensemble of galaxies selected by $d \in D$. PDF stacking is not an operation that gets you to the ensemble PDF in a scheme with uncertainty in the hyperparameters. This means for example

- stacking PDFs estimated with different codes will give you the wrong PDF (both for a single galaxy and for an ensemble) because that implicitly includes uncertainty on the hyperparameters (chosen by the different codes)
- A Bayesian hierarchical scheme is more complicated than the equation above because it uses the data to constrain the hyperparameters. But as long as they're not constrained completely, the stacked $p(z)$ of individual galaxies estimated from a Bayesian hierarchical scheme is not an unbiased estimate of their true PDF - it is wider than that.
- Actually, that is already true for the PDF of an individual galaxy. If you take spectra of many equal-looking galaxies then their distribution will be narrower than the above $p(z_i|d_i)$. That is because these spectra actually constrain h (in the limit for infinitely many spectra, they'll fully constrain h up to changes that are degenerate under the redshift distribution of galaxies that look like d_i).
- In the previous sections, we had never assumed that the H was correct, just that we weren't uncertain about it. So the $p(z_i|d_i, H)$ and $p_D(z_i|H)$ also work with little h instead of H.