COMBINING THE POSTERIOR PDFS FOR THE PHOTOMETRIC REDSHIFTS OF INDIVIDUAL GALAXIES

DANIEL GRUEN, PHIL MARSHALL AND CHRIS DAVIS

1. Introduction

A common practice in observational cosmology is to sum the posterior PDFs for the photometric redshifts of individual galaxies in order to estimate the underlying redshift distribution of galaxies. At first glance this might seem like an odd thing to do: normally we take products of probability distributions when combining information from different sources. In these notes we explore the assumptions that are being made when summing posteriors, and investigate ways to cope with these assumptions.

2. Definitions

First, some definitions:

- Let a be a vector of parameters determining 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, which can therefore be written as $z(a)$.
- Let z be the true redshift of a galaxy.
- Let d be the data vector of a galaxy in some photometric survey; it includes photometric measurements and their uncertainties.

All of these quatities come with a subscript i for galaxy i in the sky.

We imagine that there is a "true model," with some unknown parametrization, that can be thought of as having generated the data we have. The goal in inference is to learn this model and its parameters. However, in practice we do not have access to T, and so have to be satisfied with assuming a model M for any particular galaxy's redshift and type, or better, a hierarchical model H that describes the whole population of galaxies. The parameters of M include a, and also any latent variables m. In a Bayesian hierarchical scheme, the redshift distribution of galaxies is modeled explicitly, via assumptions H and a vector of hyperparameters h that determines $P(z, a|h, H)$.

3. Individual galaxy posteriors $P(z_i|d_i, M)$

Let's look at a single galaxy i and infer its redshift based on photometric measurements d_i in the survey, which is what, for instance, BPZ intends to do. This inference involves some

Date: June 17, 2017.

assumptions, which we denote by M, and so results in a characterisation of $P(z_i|d_i, M)$. The following are two simple equalities, the second of which uses Bayes' theorem:

$$P(z_i|d_i, M) = \int P(z_i|a)P(a|d_i, M)\ da = \int P(z_i|a)P(a|M)P(d_i|a, M)/P(d_i|M)\ da \tag{1}$$

Of these,

- $P(z_i|a)$ is simply $\delta(z_i - z(a))$ since we said z was fully determined by a.
- $P(a|M)$ depends on the assumptions that we make when measuring photometric redshifts (such as the magnitude distribution, or the abundance of various galaxy types, or the template spectra used).
- $P(d_i|a, M)$ is very simple in the limit of noiseless d_i: it is a multi-dimensional δ function, because in our notation d_i is equal to some subset of true observable properties a_i. In this limit, $P(d_i|a)$ independent of M. It depends on the survey, and hence M, for noisy d_i.
- $P(d_i|M)$, the distribution of *observed* parameters in our survey given our assumptions, depends a lot on the survey (because there is e.g. a spatially dependent magnitude limit), but it doesn't influence $P(z_i|d_i, M)$ because it doesn't depend on redshift and is fixed by normalization anyway ($1 = \int_{\mathbb{R}} P(z_i|d_i)\ dz_i$), in which context it is sometimes known as the "Evidence," "Fully Marginalized Likelihood," or "FML."

4. Working in the "true model"

Suppose we had a way of computing the posterior PDF for a galaxy redshift that (somehow) assumed the "true model," T.

In this case, the following statements about $P(z_i|d_i, T)$ hold:

- The individual galaxy redshift posterior PDF $P(z_i|d_i, T)$ still gets wider in z_i if our survey gets worse. It does so because $P(d_i|a_i, T)$ changes.
- If we had many apparently identical galaxies i with $d_i = d\ \forall i$ then $P(z_i|d_i, T)$ would be the same for all i. The true redshifts of these galaxies could all be different, but these redshifts would actually be distributed as $P(z_i|d_i, T)$.
- For the latter statement to be true, we require that the galaxies are not just identical in their observed data d, but also *fully representative* of the set of galaxies that look identical. Galaxies that look like d_i *but are selected by some properties that are not part of* d_i would follow $P(z_i|d_i, S)$, where S denotes some different selection than drawing faithfully from $P(d_i|T)$. This is 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.

Now let's consider 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 $P_D(z|T)$. If you measured the true redshifts of all galaxies with $d \in D$, for a complete set of bins D, and plotted up the results, you would get a histogram-like piece-wise constant approximation to the global true redshift distribution, $P(z|T)$.

The following is an equality

$$P_D(z|T) = P(z|d \in D, T) = \int_D P(z_i|d_i, T) P(d_i|T)/P(d \in D|T)\ dd_i \tag{2}$$

The factor $P(d|T)/P(d \in D|T)$ just serves to keep things normalized. It's the PDF for data d for galaxies in D given the "true model" T. In the limit of many galaxies $i = 1, ..., N_g$ that sample this PDF,

$$\int_D P(z_i|d_i, T) P(d_i|T)/P(d \in D|T)\ dd_i \approx \frac{1}{N_g} \sum_{i=1}^{N_g} P(z_i|d_i, T)\ , \tag{3}$$

and repeating this over all bins D gives us a histogram approximation to $P(z|T)$.

This last equation shows that *if we had access to the individual galaxy posterior PDFs given the true model,* then summing them over the whole ensemble (each of which represents a sample dataset d drawn from $P(d|T)$) would give a good estimator of the underlying redshift distribution. So, under what conditions does this "stacked estimator" provide a good approximation to the underlying $P(z|T)$? We can identify two:

1. The galaxies in our dataset must have been drawn from $P(d|T)$, that is, they must be a representative subset of all the galaxies in the population we are trying to understand.
2. The model Universe that we assume when we compute the posterior PDF for each galaxy's redshift must be very close to the real Universe.

In practice, each of these conditions may be hard to achieve. They are not uncorrelated, though: as the discussion above shows, one way to achieve this second condition is to work with a very large "training set" that permits us to make, for the i^{th} galaxy whose redshift z_i is to be inferred, a subset containing many galaxies that have observed data that is very close to the target galaxy's data d_i. The distribution of redshifts in this subset gives us an approximate sampling from $P(z_i|d_i, T)$. This sort of scheme sounds a lot like machine learning, evoking visions of methods like K-nearest neighbors.

5. What if your model is not the true one?

In the previous section, the noiseless observables (or latent variables) a did not appear: we just made various engineering choices to enable us to make an approximation of $P(z_i|d_i, T)$ without ever considering what galaxies are like. If we do wish to understand galaxies, we need a generative model for them, H, which allows us to draw possible galaxy properties a from $P(a|H)$. The hyper-parameters of this model, h, govern the shape of this PDF. Correctly guessing H and inferring all the h's is the endgame of galaxy evolution. In this section, we discuss two scenarios where an assumed model H is employed, and the h are unknown.

5.1. **Stacking despite model uncertainty.** What if we assume a model H that is not the true model T? We can write down $P(a|h,H)$ and $P(h|H)$, and the individual galaxy posterior PDF becomes

$$P(z_i|d_i,H) = \int\int P(z_i|a_i)P(a_i|d_i,h,H)P(h|H)\ da_i\ dh \tag{4}$$

$$= \int\int P(z_i|a_i)P(a_i|h,H)P(h|H)P(d_i|a_i,h,H)/P(d_i|H)\ da_i\ dh\ . \tag{5}$$

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 estimator made by summing posterior PDFs like these 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 an accurate estimator of the underlying redshift distribution in a scheme with uncertainty in the hyperparameters.

This means, for example, that stacking PDFs estimated with different codes will give you the wrong PDF (both for a single galaxy and for an ensemble) because that implicitly encodes uncertainty on the hyperparameters (because different hyperparameters were used by the different codes).

In a Bayesian hierarchical inference, all the parameters a_i (for all i) and hyperparameters h are constrained by the data and integrated out. But as long as they're not constrained completely, then stacking the posterior predictive distributions $P(z_i|d_i,H)$ of individual galaxies that some Hierarchical Bayesian gave you will not give an unbiased estimate of their true PDF – it will be wider than that. (Actually, that is already true for the PDF of an individual galaxy. If you take spectra of many apparently identical galaxies then the distribution of their redshifts will be narrower than the above $P(z_i|d_i,H)$. This 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).

5.2. **Hierarchical inference.** Let's investigate the hierarchical inference itself, to look for clues on how to combine posterior PDFs $P(z_i|d_i,M)$ given to us by people working with an untrue model M. A Bayesian gives you an ensemble of posterior PDFs—what do you do?

Let's go back to the simpler problem of inferring just the underlying redshift distribution, $P(z|H)$, and suppose that we can evaluate the likelihood $P(d|z,H)$ given our model. We recognize $P(z|h,H)$ as our model redshift distribution, whose functional form is determined by h, and that we are uncertain about the h. It is this $P(z|h,H)$ that we need to propagate through into future analyses that require a redshift distribution: the marginalization over h needs to be postponed, so that the degeneracy between the h and any other hyperparameters (such as the cosmological parameters) can be fully explored and captured.

Our goal, then, is to characterize the posterior PDF for h:

$$P(h|d,H) = P(h|H)\int\prod_i P(d_i|z_i,H)P(z_i|h,H)/P(d_i|H)dz_i \tag{6}$$

where we have assumed the independence of all our datasets. The following derivation follows (Foreman-Mackey et al., 2014), culminating in the same "importance sampling trick." Further discussion of its application to the photometric redshift problem can be found in Malz et al (in preparation).[1]

The giant integral in the above equation is separable, and so we can write the integral of the product as a product of integrals:

$$P(h|d,H) = P(h|H)/Z_H \prod_i \left[\int P(d_i|z_i,H) P(z_i|h,H) dz_i \right], \tag{7}$$

where the normalization constant $Z_H = \prod_i P(d_i|H)$. What our Bayesian collaborator provides us with is an ensemble of posterior PDFs $P(z_i|d_i,M)$ obtained under the assumption of some other model, M. If we insert each one into the numerator and denominator of the appropriate integrand we get

$$\int P(d_i|z_i,H) P(z_i|h,H) dz_i = \int \frac{P(d_i|z_i,H) P(z_i|h,H)}{P(z_i|d_i,M)} P(z_i|d_i,M) dz_i \tag{8}$$

If we can draw N_s samples from the supplied posterior (or if it comes to us, approximately, in the form of N_s samples), then this integral can be approximated as a sum:

$$\int P(d_i|z_i,H) P(z_i|h,H) dz_i \approx \frac{1}{N_s} \sum_{k=1}^{N_s} \frac{P(d_i|z_i^k,H) P(z_i^k|h,H)}{P(z_i^k|d_i,M)} \tag{9}$$

This shows that each integral in our product over all galaxies can be approximated by a sum of weights w_i^k, each of which is computed as a combination of three PDFs evaluated at the sample redshift z_k: the conditional prior in the hierarchical model $P(z_i^k|h,H)$, the posterior PDF supplied to us $P(z_i^k|d_i,M)$, and the likelihood $P(d_i|z_i)$:

$$w_i^k = \frac{P(d_i|z_i^k) P(z_i^k|h,H)}{P(z_i^k|d_i,M)}. \tag{10}$$

(Note that $P(d_i|z_i) = P(d_i|z_i,H) = P(d_i|z_i,M)$, since the data only depend on the redshift, not the model which generated them.) Our target "prior" PDF $P(z_i^k|h,H)$ appears in the numerator, showing our route to the inference of h: we might consider sampling, or optimizing, for h via the

$$\log P(h|d,H) = \log P(h|H) + \sum_{i=1}^{N_g} \log \sum_{k=1}^{N_s} w_i^k + \text{const.} \tag{11}$$

The scipy function `logsumexp` provides a numerically stable way of summing the weights.

As consumers of photo-z code outputs, let's consider two cases we might expect to encounter in practice:

[1] https://github.com/aimalz/prob-z

(1) **Tabulated posteriors.** In this case, we are given $P(z_i|d_i, M)$ in the form of a 1D tabulated function. We can interpolate this to compute the denominator of w_k, and we can draw samples from $P(z_i|d_i, M)$ straightforwardly. Evaluating $P(d_i|z_i^k)$ is also straightforward, *provided the data d_i really do only depend on the parameter z_i*.

(2) **Sampled posteriors.** In this case, we don't have to draw samples ourselves, but on the other hand we do need a smooth approximation to the posterior $P(z_i|d_i, M)$ so we can evaluate it in the denominator of each weight – and then we have the same task of computing the likelihood $P(d_i|z_i^k)$.

In general, the photometric data will depend on all observable intrinsic properties of the galaxy in addition to its redshift, which means computing the (marginalized) likelihood term is not easy. A better solution would be to be given the value of the log prior $[\log P(a_i^k|M)]$ along with the full set of galaxy properties a_i required to model the galaxy photometry for each sample. Recall that a_i fully describes the galaxy in the framework of the model.

In this case, we can write

$$w_i^k = \frac{P(d_i|a_i^k)P(a_i^k|h, H)}{P(a_i^k|d_i, M)}, \tag{12}$$

$$\propto \frac{P(a_i^k|h, H)}{P(a_i^k|M)} \tag{13}$$

Here, the likelihood term has cancelled out, and we are left with weights that are simply ratios of the "interim" prior (applied by the photometric redshift code) and the hierarchical prior. We divide out the former and then apply the latter, in what can now be seen to be an importance sampling approach. Note that to be able to compute this, we need posterior samples of a_i (of which z_i is an element), and the log interim prior values as well.

6. Conclusions

- If you have a machine learning photometric redshift code that returns individual posterior PDFs $P(z_i|d_i, I)$ by interpolating in data space over a large and representative training set, then an approximation to the underlying redshift distribution $P(z|T)$ is given by the stacked estimator $\frac{1}{N_g}\sum_{i=1}^{N_g} P(z_i|d_i, I)$. How accurate this redshift distribution is will depend on the size and representativeness of the training set.
- If you have a template-fitting photometric redshift code that returns individual posterior PDFs $P(z_i|d_i, M)$ by assuming model SEDs or types t_i, then the underlying redshift distribution can be modeled hierarchically as $P(z|h, H)$, and the hyperparameters h be inferred using a double sum over galaxies and their weighted posterior samples, $\log P(h|d, H) = \log P(h|H) + \sum_{i=1}^{N_g} \log \sum_{k=1}^{N_s} w_i^k + \text{const}$, where each weight is the ratio of the model distribution for the redshift (and other observables), and the "interim" prior assigned to the galaxy parameters (including the redshift) during the photometric code execution.

- The hierarchical inference above requires posterior photo-z samples to be accompanied by their log prior values.
- Simply stacking the photo-z PDFs output from an inference that assumes the wrong model for the underlying redshift distribution will produce an estimator for that underlying redshift distribution that is likely to be biased, and broader than the true distribution.

References

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, Astrophys. J., 795, 64