Quasar Photo-z Research Notes

Andrew Miller

January 23, 2015

Abstract

1 Related work

Below is a list of astronomy references on photo-z with a brief summary of methodology/contribution.

- Benitez (2000) presents a thorough summary of Bayesian methods for photometric red-shift estimation from spectral templates (as opposed to the so called empirical method of learning a function from colors to z value, e.g. with a neural network). They set up the framework where we are given data $D = \{C, m_0\}$, the colors C (ratio of each band's magnitude to some reference band's magnitude) and the magnitude information of the reference band m_0. The goal is then to infer the distribution $p(z|D)$ using some prior information in the form of a spectral template library. Some interesting notes
  - The prior probability $p(z|m_0)$ should be a function of the magnitude of observations (fainter = farther a priori). They also note that this should be defined in terms of $p(z|\hat{m}_0)$, where $\hat{m}_0$ is the true value of that magnitude, and we observe some noisy version that needs to be integrated over.
  - They then discuss galaxy templates, T, which are essentially types of galaxies (characterized by spectral density). The fully Bayesian photo-z distribution averages over these templates

$$\begin{aligned} p(z|C, m_0) &= \sum_T p(z, T|C, m_0) && (1) \\ &\propto \sum_T p(z, T|m_0) p(C|z, T) && (2) \\ p(z, T|m_0) &= p(T|m_0) p(z|T, m_0) && (3) \end{aligned}$$

    The idea is to then extend T to parameterize the types of spectra expected - basically the weights involved

$$\begin{aligned} p(z|C, m_0) &= \int dS p(z, S|C, m_0) && (4) \\ &\propto \int dS p(z, S|m_0) p(C|z, S) && (5) \end{aligned}$$

where S are like pca weights or other parameters of spectra. They discuss how to do this with the template idea, but not so much on the parameterization of spectra. This is the key piece of information I'm missing, and would be the novel contribution of the paper.

How is $p(C|z,S)$ defined - this is the important part of the likelihood

- They employ the template method as their likelihood. The likelihood, if I understand it correctly, looks like this

$$p(C|z,T) \propto \frac{1}{\sqrt{F_{TT}(z)}} \exp\left(-\frac{1}{2}\sum_{\alpha} \frac{(f_\alpha - a f_{T_\alpha})^2}{\sigma^2_{f_\alpha}}\right) \tag{6}$$

  where $f_0, \ldots, f_{n_C}$ are the observed fluxes, $f_T\alpha(z)$ is the flux of the template T (I think), and a is a magnitude nuisance parameter. They then discuss different algebraic representations of this likelihood. Importantly, I think their relationship between spectral template T and likelihood $p(\text{fluxes}|z,T)$ is sort of empirical, based on how well individual template fluxes match the observed fluxes. I think what I need is the function to go from spectral template and red shift to fluxes.

$$(f_{T,u}, f_{T,g}, \ldots, f_{T,z}) = h(z,T) \tag{7}$$

  or better yet, a parameterization

$$f(s,z)_u, \ldots, f(s,z)_z = h(z,s) \tag{8}$$

  where s are like PCA coefficients or other model parameters of the spectra.

- Equation 22 gives the form of the prior they use for the probability of galaxy type given magnitude of reference flux

$$p(T|m_0) = f_t \exp\left(-k_t(m_0 - 20)\right) \tag{9}$$

- Equation 23 gives the prior probability distribution over red shift given galaxy type and reference flux magnitude

$$p(z|T, m_0) \tag{10}$$

- Budavari et al. (2001) and companion paper Richards et al. (2001) goes into specific detail for SED model based photometric redshifts for quasars. They compare a bunch of methods. In particular, Budavari et al. (2001) describes an algorithm for reconstruction a quasar spectrum template from photometric observations and spectroscopic redshifts. It seems sorta like a dynamic K-means/EM algorithm (they add spectral types as needed), and does a decent job reconstructing bumps where the emission lines are.

  The algorithm is: start with a collection of initial spectral templates (I believe these are rest frame.) $\psi_1(\lambda), \ldots \psi_K(\lambda)$

  1. Categorize all photometric observations in the training set into one of these K categories. Which is the most likely template to describe each observation?
  2. Repair the estimated SEDs of each object (does this mean just de-redshift them?)

3. Replace each reference templates $\psi_k(\lambda)$ with the mean of each of the repaired templates of that class (discovered in step 1). This is like computing the new mean in k-means.
4. Check to see if you need to add a new template or remove an existing template based on some statistical criterion.

A difference to point out - my work proposes to do all of these steps within a rigorous probabilistic framework. We can even incorporate a nonparametric model to add/remove templates (or bases). Also, my model is more of a factor analysis approach as opposed to clustering. I'm not sure what difference this makes.

Some takeaways from this paper:

- High level intuition:

  > For a given redshift, the photometric observation gives constraints on the possible underlying SED, since we expect to get back the measured photometric values by redshifting the SED and convolving it with the filter response function. This constraint obviously depends on the photometric system and, also, the redshift of the object as the rest-frame spectrum is sampled at different wavelengths.

- They land on four classes of quasars (Fig 7), each of which has a slightly different distribution of red shift (figure 6)

Again, this paper doesn't really clarify to me an important aspect - how exactly is the χ^2 objective defined? How do you go from templates, T to fluxes? One improvement that I can make would be learning a prior over spectral model weights w (similarly, template indicators T), and use this prior to hone in on combinations of w that are likely, while leaving the correlation between w and z alone a priori. Similarly, incorporating prior information about magnitudes might disambiguate really low red shifts with higher red shifts.

- Bovy et al. (2012) - extreme deconvolution method
- Suzuki (2006) (QUASAR SPECTRUM CLASSIFICATION WITH PRINCIPAL COMPONENT ANALYSIS (PCA): EMISSION LINES IN THE Lyα FOREST)
- Brescia et al. (2013) use a multi-layer perceptron (four layers) regression setting on a combination of SDSS (from the DR7QSO dataset, I believe), UKIDSS, and WISE photometric datasets, comparing photo-z performance on the following intersections:

  1. SDSS: 1.1 × 105;
  2. SDSS $\cap$ GALEX: 4.5 × 104;
  3. SDSS $\cap$ UKIDSS: 3.1 × 104;
  4. SDSS $\cap$ GALEX $\cap$ UKIDSS: 1.5 × 104;
  5. SDSS $\cap$ GALEX $\cap$ UKIDSS *cap* WISE: 1.4 × 104.

  The largest dataset combined 43 features (mostly band fluxes and magnitudes). The authors mostly discuss the multi-layer model and their training technique, which is L-BFGS and various rounds of cross validation. The authors note their model's inability to generalize to regions of the space for which they don't have data (particularly large magnitudes or out of range z values).

The authors outline a bunch of statistics to compare between methods, including the bias, sample stdev, median of absolute value of two quantities

- $\Delta z = (z_{spec} - z_{phot})$ (residuals)
- $\Delta z_{norm} = (z_{spec} - z_{phot})/(1 + z_{spec})$ (normalized residuals)

And a bunch of percentages of outliers and such based on those statistics. One is "catastrophic outliers", defined as individual samples where $|z_{norm}| > 2\sigma(z_{norm})|$ - outside of two sample standard deviations.

- Kind and Brunner (2014) perform photo-z on a galaxy sample in a two step process

  - use a self-organizing map as an unsupervised preprocessing step to put data on a low (2) dimensional manifold, typically from galaxy meta information like fluxes and profile information. In a sense, galaxies are clustered into little grid cells on this low-dimensional manifold, which are continuous.
  - Spectroscopic z measurements are then somehow assigned to each grid cell for use in the prediction task. Prediction seems to be take galaxy features, map it to the cell, and use some sort of average or statistic of that cell's photo-z measurement. This seems like a $k - nn$ kinda thing? I'm not sure.

  They really emphasize the fact that their SOM is unsupervised, and I'm not sure why it's a selling point. They deliberately cut off their representation learning step from the actual signal they wish to predict, though it does show the representation is capturing relevant structure.

  Though similar in spirit, the self-organizing map doesn't elicit much of a generative or physical interpretation, and really serves as more of a manifold learning. I would not expect generalization outside of the range of z values and flux magnitudes to predict accurately with this method.

- Budavári (2009) (Unified photo-z paper)

- Berk et al. (2001) Describes a composite spectra for quasars using 2200 sample spectra form the SDSS sample in 2001. This paper goes into details about the characteristic emission lines and how they combined (de-redshifted, re-binned, took the median, etc), and what it says about the physical makeup of the object. Major quote

  > The steps required to generate a composite quasar spectrum involve selecting the input spectra, determining accurate redshifts, rebinning the spectra to the rest frame, scaling or normalizing the spectra, and stacking the spectra into the final composite. Each of these steps can have many variations, and their effect on the resulting spectrum can be significant

  Takeaway: details about the particular meaning behind emission lines or inductive biases about smoothness or good models for quasar spectra can probably be mined from this paper.

- Walcher et al. (2011) provides an extensive review of spectral energy density (SED) models for stars and galaxies, and various models and methods that can be used for measuring specific properties. They describe various types of eigendecomposition of galaxy spectra (SVD, trimmed SVD, robust SVD), and various other models of spectra. They also go into detail about the emission/absorption lines.

  One example they go into detail is photometric red shifts.

  > Traditionally, photometric redshift estimation is broadly split into two areas: empirical methods and the template- fitting approach. Empirical methods use a sub-sample of the photometric survey with spectroscopically-measured red- shifts as a training set for the redshift estimators. This sub- sample describes the redshift distribution in magnitude and colour space empirically and is used then to calibrate this relation. Template methods use libraries of either observed spectra of galaxies exterior to the survey or model SEDs (as described in Sect. 2). As these are full spectra, the templates can be shifted to any redshift and then convolved with the transmission curves of the filters used in the photometric survey to create the template set for the redshift estimators.

- Ball et al. (2008a) Is cited in the review Walcher et al. (2011) as a quasar photo-z method using artifical neural networks.
- Myers et al. (2009)
- Ball et al. (2008b)

References

Nicholas M Ball, Robert J Brunner, Adam D Myers, Natalie E Strand, Stacey L Alberts, and David Tcheng. Robust machine learning applied to astronomical data sets. iii. probabilistic photometric redshifts for galaxies and quasars in the sdss and galex. *The Astrophysical Journal*, 683(1):12, 2008a.

Nicholas M Ball, Jon Loveday, and Robert J Brunner. Galaxy colour, morphology and environment in the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 383(3):907–922, 2008b.

Narciso Benitez. Bayesian photometric redshift estimation. *The Astrophysical Journal*, 536(2):571, 2000.

Daniel E Vanden Berk, Gordon T Richards, Amanda Bauer, Michael A Strauss, Donald P Schneider, Timothy M Heckman, Donald G York, Patrick B Hall, Xiaohui Fan, GR Knapp, et al. Composite quasar spectra from the sloan digital sky survey. *The Astronomical Journal*, 122(2): 549, 2001.

Jo Bovy, Adam D Myers, Joseph F Hennawi, David W Hogg, Richard G McMahon, David Schiminovich, Erin S Sheldon, Jon Brinkmann, Donald P Schneider, and Benjamin A Weaver. Photometric redshifts and quasar probabilities from a single, data-driven generative model. *The astrophysical journal*, 749(1):41, 2012.

M Brescia, S Cavuoti, R D'Abrusco, G Longo, and A Mercurio. Photometric redshifts for quasars in multi-band surveys. *The Astrophysical Journal*, 772(2):140, 2013.

Tamás Budavári. A unified framework for photometric redshifts. *The Astrophysical Journal*, 695 (1):747, 2009.

Tamas Budavari, Istvan Csabai, Alexander S Szalay, Andrew J Connolly, Gyula P Szokoly, Daniel E Vanden Berk, Gordon T Richards, Michael A Weinstein, Donald P Schneider, Narciso Benitez, et al. Photometric redshifts from reconstructed quasar templates. *The Astronomical Journal*, 122(3):1163, 2001.

Matias Carrasco Kind and Robert J Brunner. Somz: photometric redshift pdfs with self-organizing maps and random atlas. *Monthly Notices of the Royal Astronomical Society*, 438(4):3409–3421, 2014.

Adam D Myers, Martin White, and Nicholas M Ball. Incorporating photometric redshift probability density information into real-space clustering measurements. *Monthly Notices of the Royal Astronomical Society*, 399(4):2279–2287, 2009.

Gordon T Richards, Michael A Weinstein, Donald P Schneider, Xiaohui Fan, Michael A Strauss, Daniel E Vanden Berk, James Annis, Scott Burles, Emily M Laubacher, Donald G York, et al. Photometric redshifts of quasars. *The Astronomical Journal*, 122(3):1151, 2001.

Nao Suzuki. Quasar spectrum classification with principal component analysis (pca): Emission lines in the lyα forest. *The Astrophysical Journal Supplement Series*, 163(1):110, 2006.

Jakob Walcher, Brent Groves, Tamás Budavári, and Daniel Dale. Fitting the integrated spectral energy distributions of galaxies. *Astrophysics and Space Science*, 331(1):1–51, 2011.