There's a bug with `start_paper` causing the title/authors/abstract to not be rendered. Also, it is a problem that it requires internet access to compile.

Introduction

Ongoing and upcoming photometric galaxy surveys such as the Large Synoptic Survey Telescope (LSST) will observe tens of billions of galaxies without spectroscopic follow-up to obtain redshifts necessary for studies of cosmology and galaxy evolution. Such surveys rely on the methods of photometric redshift (photo-z) estimation. Photo-zs are subject to a number of systematic errors, some caused by the data analysis procedures and others intrinsic to the data itself. Due to these issues, the photo-z community favors estimation of redshift probability distributions, or photo-z PDFs, that include information about the potential for such systematic errors for each galaxy in the survey.

Given the tremendous size of the surveys in question, storage of these probability distributions raises a number of nontrivial questions. Previous treatments of photometric redshifts resulted in each galaxy's catalog entry having one additional floating point number to store, or perhaps a few based on a handful of photo-z codes; how many numbers are necessary to characterize a photo-z PDF to the degree of precision dictated by the survey's science goals? Photo-z PDF codes do not all produce outputs in the same parametrization; what storage format best preserves the characteristics of a photo-z PDF?

Little attention has been paid to these matters, with a few exceptions. (Carrasco Kind & Brunner 2014)

In this work, we outline a method by which a survey team may optimize the choice of parametrization and number of stored parameters for an anticipated catalog of photo-z PDFs. We also present the publicly available `qp` Python package for performing this optimization for an arbitrary survey. The approach is demonstrated for LSST mock data.

Methods

We have developed the `qp` Python library to facilitate manipulation of photo-z PDFs. A `qp.PDF` object is defined by its parametrizations. By way of interpolation, `qp` can convert a representation of a photo-z PDF under one parametrization to a representation of that photo-z PDF under a different parametrization. The supported

parametrizations are described in Sec. 2.1. qp also includes a few built-in metrics of the accuracy of a representation of a photo-z PDF if its value under a given parametrization is designated as "true." The included metrics are described in Sec. 2.2.

Approximation Methods

First, we establish a vocabulary for the definitions of approximation methods. Each *parametrization* of a photo-z PDF is defined in terms of the *format* $\mathcal{F}$, *metaparameters* comprising $\vec{C}$, and *parameters* comprising $\vec{c}$. Each parametrization in turn corresponds to a *representation*

$$\hat{p}_{\mathcal{F},\vec{C}}(z) \equiv \mathcal{F}(z;\vec{C},\vec{c}) \tag{1}$$

of the photo-z PDF, denoted as $\hat{p}(z)$ for brevity.

qp is capable of converting a photo-z PDF between the following five formats: step functions, samples, grid evaluations, mixture models, and quantiles. These formats are described below in terms of the number N_f of stored parameters c_i per photo-z PDF, which are presumed to be floating point numbers. The metaparameters are set of N_M numbers C_m necessary to convert the stored photo-z PDF parameters $\vec{c}$ into a probability distribution function over redshift. Because the metaparameters for a catalog of photo-z PDFs released by a survey will only need to be stored once, it does not matter how large N_M is. For each format, we address the following questions:

- When/where has this format appeared in the literature as a published catalog format, native photo-z PDFcode output format, and science application input format?
- What exactly is stored under this format, per galaxy (the parameters) and per catalog (the metaparameters)?
- In what ways is this format advantageous, and what are its weaknesses?

Regular Binning

By far the most popular format for photo-z PDFs is that of a piecewise constant step function, also called a histogram binning. It is the only format that has been used for public release of photo-z PDF catalogs (Tanaka et al. 2017; Sheldon et al. 2012);

it is unclear whether this is a consequence or a cause of the fact that it is the most common format for using photo-z PDFs in cosmological inference, as tomographic binning is a universal step between the photo-z PDF catalog and calculation of any two-point correlation function.

The metaparameters of the binned parametrization are the ordered list of $N_M = N_f + 1$ redshifts $(z_1, z_2, \ldots, z_{N_M-1}, z_{N_M})$ with $z_1 < z_2 < \cdots < z_{N_M-1} < z_{N_M}$ and $z_{m+1} = z_m + \Delta_m$ serving as endpoints shared by all galaxies in the catalog, each adjacent pair of which is associated with a parameter $c_i = \int_{z_i}^{z_{i+1}} p(z)\, dz$. If the binning is "regular," then $z_{i+1} = z_i + \Delta$ for some constant scalar $\Delta \equiv$. As this is the only type of binning that has been used in the literature, it is the only one we consider.

The standard assumption that $p(z) = 0$ when $z < z_1$ or $z > z_{N_M}$ implies that the c_i are normalized according to $\Delta \sum_i c_i = 1$. Note that this is not equivalent to the erroneous normalization condition $\sum_i c_i = 1$, that commonly appears in the literature.

All known native photo-z PDF formats are easy to convert to the histogram representation. However, the corrolary to this statement is that it is an inherently lossy format. A photo-z PDF with features smaller than the bin width may not accurately represent the underlying probability distribution, as significant structure may be stored as a single bin value. This problem may be particularly troubling as the quality of the approximation will not be unbiased among catalog entries; rather, it is likely to correlate with the properties of the photo-z PDF in question.

The binned parametrization may also be considered wasteful in terms of data storage. A photo-z PDF with a compact probability distribution, with much of the probability contained within a region far narrower than the redshift range $z_{N_M} - z_1$, may have many of its catalog entries c_i being identically zero. The storage footprint of this data is wasted in that it is redundant, with each c_i requiring the same space even though $c_i = c_{i'}$ for many pairs (i, i').

Finally, the binned parametrization requires the researcher to choose the minimum and maximum possible redshifts of the galaxy sample. These are physical quantities that are unknown, so it would be preferable to not have to choose them at the stage of producing the photo-z PDF catalog.

Samples

Samples $(z_1, z_2, \ldots, z_{N_f-1}, z_{N_f})$ are another common storage format for photo-z PDFs. *Cite upcoming DES paper.* Samples are the native output format of many machine learning algorithms dependent on random choices, such as random forests.

Such approaches typically produce large numbers of samples, far more than can realistically be stored by any survey, so a subsample is commonly stored. A small number of samples from a broad photo-z PDFmay not be representative of the overall shape, but it can be appropriate for photo-z PDFs with narrow features. As in the case of the histogram parametrization, there is a significant risk of bias in the quality of the approximation with respect to properties of the photo-z PDF.

Though it is possible to construct a catalog where each galaxy has a different number N_f of stored samples, optimizing the choice of N_f given the shape of the photo-z PDF in its native format is nontrivial. As this has not been done in the literature, we leave its investigation to future work.

Evaluation on a Regular Grid

Another storage option is evaluations of a continuous functional form of a photo-z PDF at N_f redshifts $z_1, z_2, \ldots, z_{N_f-1}, z_{N_f}$. Thus far, only a regular grid of redshifts (i.e. $z_{i+1} = z_i + \Delta$ for a constant Δ) shared among the entire catalog have been used in the literature, so we do not consider irregular grids or grids that are not homogeneous over the catalog. In some cases, this is the native output format of a photo-z PDF code. *Apparently BPZ does this because that's what Melissa and Sam gave me, but I don't have anything to cite saying this.*

The gridded parametrization is in some ways similar to the histogram format of 2.1.1, suffering from the same risks of losing small-scale features (and the corresponding risk of systematic bias in the quality of the approximation), wasting a substantial fraction of the N_f allocated parameters, and choosing the grid endpoints. Unlike the piecewise constant parametrization, the grid evaluations are not necessarily normalized, which can be a challenge for null tests and science use cases. Additionally, it is not in general easy to convert to this format from the native format of a photo-z PDF code, as it requires a continuous function for the photo-z PDF.

Mixture Model

There is some history of using a mixture model parametrization for photo-z PDFs. *I'm still seeking something to cite here – I thought the native output of BPZ was the means and standard deviations of the top three Gaussian components, but this does not appear to be so.* In this work, we consider only the common Gaussian mixture model, though the `qp` framework can accommodate mixtures of all probability distribution functions that have been implemented as `scipy.stats.rv_continous` objects.

A Gaussian mixture model may be a natural choice for the native output of a template-fitting code or one based on a distance metric in the space of photometry, and it is easy to convert it to the previously discussed parametrizations. However, it may be difficult to fit a mixture model to photo-z PDFs in other native formats, and the mixture model is only an accurate approximation of the photo-z PDFs are actually comprised of components of the included parametrizations.

Besides the mixture of Gaussians, other functions have been investigated before in the sparse basis representation of Carrasco Kind & Brunner (2014). Though it has promising compression properties, we do not consider it in this work for several reasons. Decomposition with `SparsePZ` does not guarantee that the stored parametrization be a probability distribution in the mathematical sense of always being positive definite and integrating to unity, which we consider a necessary condition both for use of photo-z PDFs in research and for comparison to other methods using the `qp` metrics. The sparse basis representation also assumes that the native format of a photo-z PDF is evaluations on a grid; if this is not true, then the photo-z PDF may undergo additional conversions that introduce loss of information with each approximation.

I also thought it required extensive computational resources, but it appears to be a lot faster than `qp` making quantiles with the same number of parameters – recall that calculating quantiles generally requires an optimization. My reasons are pretty weak; if I had more time I'd have `qp` employ the `SparsePz` method as another parametrization, but I don't know when I'll be able to do it!

Regular Quantiles

One parametrization that has not previously been implemented is that of quantiles, which are defined in terms of the cumulative distribution function (CDF)

$$CDF(z) = \int_{-\infty}^{z} p(z)\, dz. \tag{2}$$

Under the quantile parametrization, an ensemble of photo-z PDFs shares a set of N_f values $0 < q_1 < q_2 < \cdots < q_{N-1} < q_{N_f} < 1$. Each galaxy's catalog entry is the N_f values z_i satisfying $CDF(z_i) = q_i$. In our tests, the quantiles are regular, such that $q_i \equiv i\bar{q}$, where $\bar{q} \equiv (N_f + 1)^{-1}$, but `qp` does not require that this be so.

The quantile parametrization is the inspiration for this work and namesake of `qp`. Though it has not appeared in the photo-z PDF literature prior to this point, it is a natural choice for the compression of probability distributions because it keeps more information in areas of higher probability density, so there is inherently less waste in the information that is stored for each catalog entry and minimal risk of bias in the quality of the approximation across photo-z PDF shapes. Storing quantiles is equivalent to storing piecewise constant data (or function evaluations) on an irregular binning (or irregular grid) optimized to have narrower bins (denser evaluations) in areas of high probability and wider bins (more diffuse evaluations) in areas of low probability, effectively performing the optimization in bin size (grid resolution) while still permitting an entire catalog to share a single set of metaparameters. Unlike samples, it is guaranteed to be an equally good approximation regardless of the shape of the photo-z PDF.

The quantile parametrization is not without its drawbacks. There is as yet no infrastructure for using such a format in a scientific application, but there is no reason to think that this cannot change if the quantile parametrization proves effective. Furthermore, no known photo-z PDF method has quantiles as a native output format, and some native photo-z PDF output formats, like samples, are easy to convert to quantiles, while others, like piecewise constant functions, are not, requiring a numerical optimization for each parameter i. Nonetheless, the quantile parametrization is a new option that merits careful consideration.

Comparison Metrics

We use two metrics to quantify how well an approximation of a photo-z PDF extracted from a stored format represents the original photo-z PDF, characterized by many more parameters than are available for storage, before it was compressed. Given the stored parameters, we perform an interpolation to evaluations of the approximated photo-z PDF $\hat{p}(z)$ on a fine grid $(z_1, z_2, \ldots, z_{N_{ff}-1}, z_{N_{ff}})$ to calculate metrics against the original format $p_0(z)$. The distributions of metric values for each parametrization are compared in Sec. 5.1. The metrics are also calculated for a science application of photo-z PDFs in Sec. 5.2.

RMSE

The root mean square error (RMSE) is a familiar measure of the difference between two functions $p(z)$ and $\hat{p}(z)$,

$$RMSE = \frac{1}{N_{ff}} \sum_{i=1}^{N_{ff}} (p(z_i) - \hat{p}(z_i))^2. \quad (3)$$

The RMSE is symmetric in that it is simply about the difference between the functions, not some distance from one to the other. The RMSE is also not specific to probability distributions. The RMSE is always positive, and a smaller value indicates better agreement between the approximation and the truth.

KLD

The Kullback-Leibler divergence (KLD)

$$D(p_0(z)||\hat{p}(z)) = \int_{-\infty}^{\infty} p_0(z) \log\left[\frac{p_0(z)}{\hat{p}(z)}\right] dz \quad (4)$$

quantifies the loss of information of an approximation of a probability distribution from the true probability distribution. Note that the KLD is not symmetric and requires not only that both functions be true probability distributions but also that there must be some notion of a true reference distribution. The KLD is always positive, and a smaller value indicates better agreement between the approximation and the truth.

Photo-z Test Data

With the expectation that `qp` may suggest a different optimal result for different datasets, we apply it to two mock datasets with different data quality properties. Both datasets were fit using Bayesian Photometric Redshift (BPZ) estimation (Bentez 2000), which employs spectral energy distribution (SED) fitting to a template set. However, the choice of photo-z PDF estimation method is not relevant to this study; so long as the mock photo-z PDFs are of realistic complexity, it does not matter how accurately they describe the probability distribution of galaxy redshifts given their photometric data. We only seek to optimize the fidelity of the stored photo-z PDF relative to the photo-z PDF output by a representative photo-z PDF fitting code. (Other work has been done to compare the accuracy of photo-z PDFs

Figure 1. [single-panel plot with a few examples of what the LSST-only photo-z PDFs look like]

produced by different methods; see Schmidt, et al. 2017 (in prep.), Tanaka et al. (2017).)

As BPZ is a widely used and well established method, we assume that the photo-z PDFs produced by it are of realistic complexity, meaning they take shapes we expect to see in accurate photo-z PDFs from real datasets with similar photometric properties. The mock datasets considered here have already been transformed into a gridded parametrization, as we did not run the photo-z PDF code ourselves to get the raw output. To create a realistically complex testbed catalog, we fit these high-resolution, gridded photo-z PDFs with a Gaussian mixture model so that our catalog has a notion of "true" underlying photo-z PDFs. *At this point, the only reason to do this is the suboptimal implementation of* `qp.PDF.truth`. *It would not be impossible to eliminate this last step.*

Carrasco Kind & Brunner (2014) uses $N_g = 10^6$ *galaxies. How many should we use for the paper?*

LSST mocks

Should we invite Sam to write this section?

LSST will provide six-band optical photometry to a depth of 27.5 magnitudes. *Is this consistent with Buzzard?* The Buzzard simulations The photo-z PDFs were provided in the form of $N_{ff} = 211$ floating point numbers represing the probability on a regular grid of redshifts $0.005 < z < 2.11$. Due to the small number of photometric filters, LSST-only photo-z PDFs are expected to be multimodal; some examples are shown in Fig. 1. We produced true underlying PDFs by fitting a five-component Gaussian mixture model to each photo-z PDF.

LSST+Euclid mocks

Should we invite Melissa to write this section?

While LSST will provide six-band optical photometry to a depth of 27.5 magnitudes, Euclid will provide three-band near infrared photometry to a depth of 24 magnitudes. The photo-z PDFs were provided in the form of $N_{ff} = 351$

Figure 2. [single-panel plot with a few examples of what the LSST+Euclid photo-z PDFs look like]

floating point numbers represing the probability on a regular grid of redshifts $0.01 < z < 3.51$. With even three additional photometric filters, photo-z PDFs derived from LSST+Euclid photometry are expected to be narrow and in most cases unimodal. A few examples of those with nontrivial shapes are shown in Fig. 2, but these represent a small fraction of the galaxies in the sample. We produced true underlying PDFs by fitting a two-component Gaussian mixture model to each photo-z PDF.

Science Metrics

Though the use of photo-z PDFs could potentially extend to all areas of astronomy in which redshifts are used, photo-z PDFs have thus far been used almost exclusively to estimate the redshift distribution function $n(z)$ necessary for calculating the correlation functions used by many cosmological probes. The most common way to estimate the redshift distribution function is to sum the photo-z PDFs according to

$$\hat{n}(z) \equiv \frac{1}{N_g} \sum_{j=1}^{N_g} \hat{p}_j(z), \tag{5}$$

where the estimator is normalized so that it, too, is a probability distribution. Though we do not recommend this approach to estimating the redshift distribution (see Malz and Hogg, et al. (in prep.) for a mathematically consistent alternative), we also calculate the metrics of Sec. 2.2 on this "stacked estimator" of the redshift distribution function under the assumption that inaccuracy and imprecision of a photo-z PDF parametrization in th

Results

In this study, we perform tests comparing the parametrizations of Sec. 2.1 as a function of the number of parameters per galaxy. The tests are conducted using the functionality of the `qp.Ensemble` class that is a wrapper for collections of `qp.PDF` objects.

Individual photo-z PDFs

Figure 3. [histograms of each metric for different numbers of parameters, one panel per combination of metric (horizontal) and N_f (vertical)]

Figure 4. [number of stored floats N_f vs. metric value, one curve per approximation method, one panel per combination of metric (horizontal) and N_f (vertical)]

Parametrizations are compared on the basis of the distributions of the metrics of Sec. 2.2 calculated over all galaxies in the ensemble.

Stacked $\hat{n}(z)$ estimator

Parametrizations are also compared by the accuracy of their stacked redshift distribution estimator $\hat{n}(z)$ relative to that of the photo-z PDFs in their original format.

Conclusions & Future Directions

We have presented a method for determining the most appropriate compression formats and number of stored parameters for large catalogs of photo-z PDFs produced by surveys with limited storage capacity. In the case of well-behaved photo-z PDFs, we observe [RESULTS]. In the case of photo-z PDFs from lower quality data, we observe [RESULTS]. Given the constraint that LSST will be able to store 200 floating point numbers to quantify the redshift of each galaxy and intends to include several photo-z PDF codes, we recommend the [RECOMMENDATION] parametrization.

This work addresses the tradeoff between the accuracy of stored photo-z PDFs and the footprint of the data needed to encode the photo-z PDFs. It does not address the computational resources necessary to perform the storage operation nor to unpack the stored information into the form necessary for science computations. There may be other issues with the loss of information when extracting compressed photo-z PDFs for use in science, with impacts that may differ for each science case.

Further applications of `qp` functionality for manipulations of photo-z PDFs is demonstrated in the LSST-DESC PZ DC1 paper (in prep.).

Acknowledgments

We thank Melissa Graham and Sam Schmidt for providing the mock datasets. This work was incubated at the 2016 LSST-DESC Hack Week.

This is the text imported from `acknowledgments.tex`, and will be replaced by some standard LSST DESC boilerplate at some point.

Author contributions are listed below.
Alex Malz: Initiated project, led development work.
P.J. Marshall: advised on statistics, and project design and management.

References

Bentez, N. 2000, The Astrophysical Journal, 536, 571

Carrasco Kind, M., & Brunner, R. J. 2014, Monthly Notices of the Royal Astronomical Society, 441, 3550

Sheldon, E. S., Cunha, C. E., Mandelbaum, R., Brinkmann, J., & Weaver, B. A. 2012, The Astrophysical Journal Supplement Series, 201, 32

Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2017, arXiv:1704.05988 [astro-ph], arXiv: 1704.05988