The Astrophysical Journal, 695:747–754, 2009 April 10 doi:10.1088/0004-637X/695/1/747

A UNIFIED FRAMEWORK FOR PHOTOMETRIC REDSHIFTS

Tamás Budavári
Department of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA; budavari@jhu.edu

Received 2008 November 16; accepted 2009 January 14; published 2009 April 1

ABSTRACT

We present a rigorous mathematical solution to photometric redshift estimation and the more general inversion problem. The challenge we address is to meaningfully constrain unknown properties of astronomical sources based on given observables, usually multicolor photometry, with the help of a training set that provides an empirical relation between the measurements and the desired quantities. We establish a formalism that blurs the boundary between the traditional empirical and template-fitting algorithms, as both are just special cases that are discussed in detail to put them in context. The new approach enables the development of more sophisticated methods that go beyond the classic techniques to combine their advantages. We look at the directions for further improvement in the methodology, and examine the technical aspects of practical implementations. We show how training sets are to be constructed and used consistently for reliable estimation.

Key words: galaxies: statistics – methods: statistical

Online-only material: color figures

1. MOTIVATION

The concept of photometric redshift estimation is over four decades old. Since Baum (1962) the methodology has changed only incrementally but its role in astronomy has completely spun around. The astronomy community originally received the idea with serious skepticism, which, over time, thanks to a series of breakthroughs in the field (e.g., Koo 1985; Connolly et al. 1995a), slowly faded. Today the next generation telescopes plan to perform photometric observations only and completely rely on these kind of estimation techniques for most of their key science projects including cosmology and large-scale structure.

While getting ready for extracting most of our new scientific knowledge from photometric measurements, we have to examine the current limitations of the various techniques and understand the underlying assumptions. Essentially all currently existing implementation can be categorized into two classes of methods: empirical estimators and template fitting. Reviewing the history of the research area is outside the scope of this study; see Weymann et al. (1999) for a rich cross section of the field instead; now we look at the basic concepts and the differences in the traditional methodologies. Empirical methods map the relation of the observed and desired properties using a training set; e.g., piecewise linear or polynomial fitting, or via other regression methods like artificial neural nets, support vector machines, etc. Template-fitting techniques rely on prior knowledge encoded in the model's spectral energy distributions (SEDs) that can be matched to observations. Why are the current implementations of these two so different? There is no fundamental reason, e.g., one could generate training sets from model templates. Why do only template-fitting algorithms use photometric uncertainties and not the empirical ones? Why do people estimate the redshifts independently from other physical properties, e.g., often use empirical redshift estimates and then template spectra for type determination? We know these quantities are correlated and should be dealt with in a consistent way. The answers to these questions are usually direct consequences of limitations in the models and the measurements. If the model SEDs matched all the observations, we would know everything about all the objects in the universe. The uncertainties would be used more often if they provided reliable extra information.

The "Photo-Z" label currently associated with the above methods, should gain a new meaning. We should expect more from the codes than a single estimate per object. The implementations need to provide the full joint probability density functions of all desired physical parameters, so we can develop new statistical tools that utilize all the information available.

In this paper, we are not concerned with what observables are the best to use or which filter set is optimal for special cases of the generalized photometric inversion problem, which depend on the specific science cases, instead we derive a probabilistic formalism to address the common issues. In Section 2, we introduce the methodology and derive the formulas for determining the photometric constraints on physical properties. Section 3 describes the traditional empirical and template-fitting algorithms as special cases of the proposed framework, and the advanced techniques that go beyond their limits. In Section 4, we illustrate the concepts and detail the practical aspects. Section 5 concludes our study.

Throughout the paper, we use the capital P letter for probabilities and the lower-case p letter for probability density functions, or PDFs for short.

2. METHODOLOGY

We start by formulating the problem as general as possible. The challenge is to constrain physical properties of sources with some observables in a data set denoted by Q, hereafter the query set. Since model spectra would never be perfectly suitable for all desired parameters, one will need a training set, T. In fact there is no reason to demand that these data sets have the same observables. The mapping is provided by some model, M. For example, magnitudes of different photometric systems can be mapped on to one another, say, $UJFN$ observations to $ugriz$ (Fukugita et al. 1995) by empirical formulas. In general, let $\boldsymbol{x}$ be a set of observables in the training set T that also contains extra information about the physical properties $\boldsymbol{\xi}$, and let $\boldsymbol{y}$ denote the observables of the query set Q. Our model is parameterized by a vector $\boldsymbol{\theta}$;

$$\begin{aligned} T: &\quad \{\boldsymbol{x}_t, \boldsymbol{\xi}_t\}_{t\in T} \\ Q: &\quad \{\boldsymbol{y}_q\}_{q\in Q} \\ M: &\quad \boldsymbol{\theta}. \end{aligned}$$

The model M can predict the observables $\boldsymbol{x}$ and $\boldsymbol{y}$ for a given parameter via the density $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, M)$ and has a prior on its parameters $p(\boldsymbol{\theta}|M)$. For example, one can build models based on the Coleman et al. (1980) or Bruzual & Charlot (2003) templates that can be used to calculate the colors of sources at a given redshift in any particular photometric system. However, the modeling goes beyond just estimating the values for a given parameter, because the observational uncertainties also enter the formula. Later on, we will discuss in details how to establish various models; for now, the above functions are assumed to be known. Furthermore, let us assume that the training set samples the entire space of the observables, and discuss the selection effects later.

Our goal is to derive the probability density function (PDF) of the physical properties $\boldsymbol{\xi}$ for a given query point q with $\boldsymbol{y}_q$ observations using our model M. This function, $p(\boldsymbol{\xi}|\boldsymbol{y}_q, M)$, is the solution of the generalized photometric inversion problem and the subject of this section. The next two paragraphs discuss probabilistic concepts analogous to elements of template fitting and empirical estimation, respectively, in the context of our probabilistic formalism. Next, we address the burning issues of selection effects and feasibility.

2.1. Mapping the Observables

The first step is to make the connection between the observables. It can be done formally by calculating the probability density of $\boldsymbol{x}$ for the query point q. We do this via the equality of

$$p(\boldsymbol{x}|\boldsymbol{y}_q, M) = \frac{p(\boldsymbol{x}, \boldsymbol{y}_q|M)}{p(\boldsymbol{y}_q|M)}, \tag{1}$$

where the right-hand side contains integrals of known functions over the model's parameter domain

$$p(\boldsymbol{x}, \boldsymbol{y}_q|M) = \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|M)\, p(\boldsymbol{x}, \boldsymbol{y}_q|\boldsymbol{\theta}, M) \tag{2}$$

and over $\boldsymbol{x}$ for the marginalization

$$p(\boldsymbol{y}_q|M) = \int d\boldsymbol{x}\ p(\boldsymbol{x}, \boldsymbol{y}_q|M). \tag{3}$$

We see how this is superior to the techniques analogous to the traditional way. The usual solution involves fitting for the best-match model parameter using, for example, maximum likelihood estimation (MLE), and accepting that parameter at face value to derive the estimates. Here, we consider all possible model parameters and add up their contributions.

We note that the above general mapping formula is valid in case of improper priors, too, in the sense that the posterior is always properly normalized to unity. If one has no prior knowledge about the model parameters, and wishes to use a noninformative prior, e.g., flat $p(\boldsymbol{\theta}|M) = 1$, formally he/she is allowed to do so; see more on the priors later on.

2.2. Physical Properties

Next we establish the relation between the observable and the desired physical parameters. The traditional way is to assume the properties of interest to be a function of the observables. Some of the existing methods utilize explicit functions such as a polynomial or piecewise linear, while others use more obscure mappings such as a decision tree or an artificial neural net. Conceptually, they are just assuming a fitting function

$$\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}(\boldsymbol{x}), \tag{4}$$

which is tuned to reproduce the elements of the training set as best as possible. The problem with this assumption is that there is no guarantee that the same $\boldsymbol{x}$ observables always correspond to the same $\boldsymbol{\xi}$ properties. In fact, we know that degeneracies are present in most data sets. Clearly, the above assumption is an unnecessary restriction over the general relation of $\boldsymbol{x}$ and $\boldsymbol{\xi}$ denoted by $p(\boldsymbol{\xi}|\boldsymbol{x})$. In other words, the traditional model is

$$p(\boldsymbol{\xi}|\boldsymbol{x}) = \delta(|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}(\boldsymbol{x})|) \tag{5}$$

using Dirac's δ symbol.

A better way is not to restrict the distribution arbitrarily to an unknown surface but to leave the formula general. We can establish the proper relation by observing the fact that

$$p(\boldsymbol{\xi}|\boldsymbol{x}) = \frac{p(\boldsymbol{\xi}, \boldsymbol{x})}{p(\boldsymbol{x})}. \tag{6}$$

The right-hand side is a ratio of two densities that (both) can be estimated from the training set, e.g., using Voronoi tessellation or kernel density estimation (KDE).

Having derived the above relation, one can compute the final PDF of interest as the integral over the possible observables in the training set

$$p(\boldsymbol{\xi}|\boldsymbol{y}_q, M) = \int d\boldsymbol{x}\ p(\boldsymbol{\xi}|\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{y}_q, M). \tag{7}$$

When it is possible to accurately characterize this distribution by a Gaussian function or some mixture model, one can compress the numerical results into a few parameters. When the PDF is unimodal, which is often not the case, the expectation value should suffice for an estimate

$$\bar{\boldsymbol{\xi}}(\boldsymbol{y}_q) = \int d\boldsymbol{\xi}\ \boldsymbol{\xi}\ p(\boldsymbol{\xi}|\boldsymbol{y}_q, M). \tag{8}$$

The above equation is similar to kernel regression (Nadaraya 1964) in case of using KDE, except it is a generalization to incorporate the uncertainties in the data sets.

Photometric redshifts and other such properties are often used in statistical studies for their availability for a large number of sources, even though they provide relatively loose constraints on individual objects. The full PDFs of the sources are best suited to derive the ensemble properties of entire catalogs or even specific subsamples. The distribution of the properties over a set of measurements Q is given by the average

$$p(\boldsymbol{\xi}|Q, M) = \left\langle p(\boldsymbol{\xi}|\boldsymbol{y}_q, M)\right\rangle_{q\in Q}. \tag{9}$$

Hence, there is no need for an extra deconvolution step to recover the underlying distribution of the objects in a sample, because their average PDF is exactly that. A common example is the estimation of the redshift distribution dN/dz for various subsamples, say, at different distances. When selection bias is not an issue for the scientific analysis, e.g., lensing studies, one can even choose the subsets to optimize the contrast of the averaged PDFs.

2.3. Selection Effects

The inherent limitations of a finite training set pose a serious problem for any estimator, which is often neglected. Our formalism introduced earlier is no exception, hence we now turn to examine the effects. The selection function is the probability of a source, with observables $\boldsymbol{x}$ making it into the training set, $P(T|\boldsymbol{x})$. The region that the training set can sample is the window function $P(W|\boldsymbol{x})$, which takes the value of 1 where the selection function is nonzero, and 0 otherwise. For example,

$$P(W|\boldsymbol{x}) = \begin{cases} 1 & \text{if } V(\boldsymbol{x}) < 22 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

for a survey that has a magnitude limit of 22 in a V band.

The selection function is expected to enter our method at two separate places: the marginalization over $\boldsymbol{x}$ and via the density estimates used for the relation $p(\boldsymbol{\xi}|\boldsymbol{x})$. The former appears to be inevitable, but causes problems only at the boundaries of the selection criteria. If the integrand $p(\boldsymbol{x}|\boldsymbol{y}_q, M)$ in Equation (7) vanishes within the integration domain of the window function $P(W|\boldsymbol{x})$, the results are valid. Otherwise the estimated PDF is biased in an unknown way. The probability of q being inside the window function is the right indicator of the problem occurring

$$P(W|\boldsymbol{y}_q, M) = \int d\boldsymbol{x}\; P(W|\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{y}_q, M). \tag{11}$$

When this probability is close to 1, the training set provides good support for the photometric inversion problem, but when the value is low, the query point is known to be outside the regime of the training set.

The relation between the desired properties and the observables is the other issue, as it is only probed on the training set. The relation as seen on the training set depends on the true relation and the selection function via the equation

$$p(\boldsymbol{\xi}|\boldsymbol{x}, T) = \frac{p(\boldsymbol{\xi}|\boldsymbol{x})\, P(T|\boldsymbol{x}, \boldsymbol{\xi})}{P(T|\boldsymbol{x})}. \tag{12}$$

If the selection function strictly depends only on $\boldsymbol{x}$, we have $P(T|\boldsymbol{x}, \boldsymbol{\xi}) = P(T|\boldsymbol{x})$ and find that the empirical relation is identical to the true one on the selection domain. If the sampling frequency is low, the measured relation is noisier and less robust numerically.

This is a critical point, which is worth emphasizing once again: the $p(\boldsymbol{\xi}|\boldsymbol{x}, T) = p(\boldsymbol{\xi}|\boldsymbol{x})$ equality holds only if $\boldsymbol{\xi}$ does not influence the selection in any way, not even indirectly via some hidden parameter. A counter example is the common case of cutting on morphological parameters in the selection function, while only considering the fluxes for $\boldsymbol{x}$. Another interesting consequence is that one cannot use only the colors to estimate, say, photometric redshifts, if a magnitude cut was involved in the selection of the training set. Yet another issue is cosmic variance, which might cause the relation to depend on the position in the sky. The solution in all cases is to revise the selection of the training set, if possible, or to add the hidden observables into $\boldsymbol{x}$, and extend the model to include them.

3. MODELS IN THE TRADITIONAL LIMITS AND BEYOND

Previously, we have hinted at how models can be constructed but, until now, they have just been assumed to be known. A model is a combination of the limitations in our observations, both in the training and query sets, and the parameterization of the observables. From discussing the topic in the most general way, we now turn to the practicalities of real-life astronomical observations.

Today the errors of extracted fluxes of photometric measurements are independent estimates of the uncertainties in the separate passbands. Typically, Gaussian errors are assumed, and the catalogs would quote 1σ values for every source. Analyzing the repeated observations in the Sloan Digital Sky Survey (SDSS; York et al. 2000), Scranton et al. 2005 have shown that this simple picture is wrong, and the off-diagonal elements of the covariance matrix are significant. This is not surprising. One of the major components in the photometric uncertainty is the error in the determination of the aperture. If the multicolor measurements share a common aperture, e.g., SDSS model magnitudes that are best suited for colors, the flux measurements will be inevitably correlated. Thus, an improved error model of the photometric observations is described by a multivariate normal distribution, $N(\boldsymbol{x}|\bar{\boldsymbol{x}}, \mathbf{C}_x)$, with a mean of $\bar{\boldsymbol{x}}$ and covariance matrix $\mathbf{C}_x$. The next generation survey telescopes that plan to visit the sources on multiple occasions will be able to better determine the full covariance matrices from actual observations to improve our understanding of the errors. Hence, for now it is general enough to consider error estimates that are fully described by the covariances.

In this reasonable approximation, the $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, M)$ mapping is also a normal distribution with a full covariance matrix that includes cross-catalog terms, if necessary, that go beyond the calibration work on the individual catalogs. If the apertures are locked together for better color determination, one has to obtain the dependencies via a data set that contains sources with all $\boldsymbol{x}$ and $\boldsymbol{y}$ measurements. However, when the processing pipelines are independent, one can assume that the uncertainties in $\boldsymbol{x}$ and $\boldsymbol{y}$ are also independent, and write a realistic M as the product of the two Gaussians:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, M) = {} & N_x\left(\boldsymbol{x}|\bar{\boldsymbol{x}}(\boldsymbol{\theta}), \mathbf{C}_x(\boldsymbol{\theta})\right) \\ & \times\; N_y\left(\boldsymbol{y}|\bar{\boldsymbol{y}}(\boldsymbol{\theta}), \mathbf{C}_y(\boldsymbol{\theta})\right). \end{aligned} \tag{13}$$

The dependences in the means $\bar{\boldsymbol{x}}(\boldsymbol{\theta})$ and $\bar{\boldsymbol{y}}(\boldsymbol{\theta})$ are straightforward to model and, even in the most complicated case, are similar, in spirit, to the traditional template-fitting procedures. For example, when considering a synthetic model of galaxies, one has to vary the redshift, age, optical depth, and so on, to derive high-resolution model spectra for different parameters, and then convolve them with the broadband filters to get the fluxes.

Clearly modeling the covariance matrices is more complicated and would require many more parameters to model accurately. If $\boldsymbol{\theta}$ is a minimal set of parameters that is enough to describe $\bar{\boldsymbol{x}}(\boldsymbol{\theta})$ and $\bar{\boldsymbol{y}}(\boldsymbol{\theta})$, there are some other hidden parameters or hyperparameters that are also needed for the covariances. The fully Bayesian way is to establish the relation of the covariance matrix and the hyperparameters along with a hyperprior (the prior on the hyperparameters), and to marginalize over the extra dependence. Even though, this relation between the elements of the covariance matrix and the observables could, in principle, be modeled based on the catalogs, it may prove impractical. The empirical Bayes approach, admittedly more optimistic but easily quantifiable, is to find the most likely hyperparameter and substitute it into the dependence. In practice, for every parameter $\boldsymbol{\theta}$, one can find the values of $\bar{\boldsymbol{x}}(\boldsymbol{\theta})$ and $\bar{\boldsymbol{y}}(\boldsymbol{\theta})$ and the closest measurement points, whose covariance matrices are good estimates. If the covariance matrix changes slowly with $\boldsymbol{x}$ compared to its widths, one can safely calculate the values at the catalog points by using the corresponding error matrices,

$$p(\boldsymbol{x}_t, \boldsymbol{y}_q|\boldsymbol{\theta}, M) = N_x\,(\boldsymbol{x}_t|\bar{\boldsymbol{x}}(\boldsymbol{\theta}), \mathbf{C}_t) \times N_y\left(\boldsymbol{y}_q|\bar{\boldsymbol{y}}(\boldsymbol{\theta}), \mathbf{C}_q\right). \quad (14)$$

The only concern with this approximation is the noise on the elements of the covariance. If needed, one could improve on the stability by smoothing or fitting locally over the catalog entries.

The consequences of the model approximation in Equation (14) are most intriguing from the implementation aspect of the methodology. As long as we only evaluate the PDFs at the observed locations, the calculations are more straightforward and computationally less expensive.

3.1. Numerical Evaluation

The field of numerical evaluation of complicated multidimensional integrals that usually emerge in Bayesian analysis such as ours is well studied. The solution typically involves some randomized algorithms that range from simple direct sampling from the prior to adaptive strategies often based on Markov chain Monte Carlo (MCMC) methods, e.g., Gibbs sampling. Although this topic is beyond the scope of the present discussion, we briefly touch on the basic idea to illustrate the concepts and provide some insight on how to derive the final results numerically, namely the value of $P(W|\boldsymbol{y}_q, M)$ and the function $p(\boldsymbol{\xi}|\boldsymbol{y}_q, M)$.

The clever construction of the chain in the MCMC algorithm yields model parameters $\{\boldsymbol{\theta}_i\}$ that can be considered independent random realization drawn from the posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y}_q, M)$ in our case. With the chain in hand, one can readily approximate the integral by the average over the MCMC samples. The mapping of the observables then becomes

$$p(\boldsymbol{x}|\boldsymbol{y}_q, M) = \left\langle N_x\,(\boldsymbol{x}|\bar{\boldsymbol{x}}(\boldsymbol{\theta}_i), \mathbf{C}_t)\right\rangle, \quad (15)$$

where t is the index of the training point $\boldsymbol{x}_t$ closest to $\bar{\boldsymbol{x}}(\boldsymbol{\theta}_i)$. When the query point is well within the regime of the training set, this approximation is valid. What happens otherwise? Often the uncertainties are larger outside the selection criteria, e.g., the photometric errors beyond the flux limit. By using the covariance matrix of the closest training point, one actually artificially decreases the contribution to the integral making $p(\boldsymbol{x}|\boldsymbol{y}_q, M)$ tighter. While the accuracy of the calculation is affected, the change is such that it reduces the value of the integral in $P(W|\boldsymbol{y}_q, M)$, which is the measure of reliability. Hence, if we measure a large value, we can be confident of the result. Having said that we note that in practice the covariances probably do not change fast enough to pose a significant problem in this calculation for the objects along the edge of the selection function, and farther away the probabilities are very small anyway.

Once we know that the estimation is in the safe regime, we can compute the $p(\boldsymbol{\xi}|\boldsymbol{y}_q, M)$ integral ignoring the window function completely by summing up at preset $\boldsymbol{\xi}_r$ points in our region of interest, e.g., a fine redshift grid, as

$$p(\boldsymbol{\xi}_r|\boldsymbol{y}_q, T, M) \propto \sum_{t\in T} p(\boldsymbol{\xi}_r|\boldsymbol{x}_t, T)\frac{p(\boldsymbol{x}_t|\boldsymbol{y}_q, M)}{p(\boldsymbol{x}_t|T)}, \quad (16)$$

where the $p(\boldsymbol{x}_t|T)$ densities and the matrix $p(\boldsymbol{\xi}_r|\boldsymbol{x}_t, T)$ are obtained from the numerical density estimates once for the training set; see Equation (6). Here, we made use of the fact that the $\{\boldsymbol{x}_t\}$ points are (naturally) drawn from the distribution $p(\boldsymbol{x}|T)$.

In order to perform these summations efficiently for many query points, one has to utilize fast searching mechanisms in the space of the observables. The situation is complicated by the strong correlation in the observables and the varying Mahalanobis metric, yet, a significant speedup can be achieved by adequate multidimensional indexing of the color–space as described in Csabai et al. (2007).

3.2. Template Fitting

In classical SED-fitting approaches, one does not technically have a training set. Although formally it can be generated from a grid of model parameters $\{\boldsymbol{\theta}_t\}$ as $\{\boldsymbol{x}_t, \boldsymbol{\xi}_t\} = \{\bar{\boldsymbol{x}}(\boldsymbol{\theta}_t), \bar{\boldsymbol{\xi}}(\boldsymbol{\theta}_t)\}$, where $\bar{\boldsymbol{\xi}}(\boldsymbol{\theta})$ is often simply a subset of $\boldsymbol{\theta}$, e.g., the redshift is just one of the parameters in the models of SEDs. Traditionally, this artificial training set has no errors associated with the reference points, hence we have

$$p(\boldsymbol{x}|\boldsymbol{\theta}, M) = \delta(|\boldsymbol{x} - \bar{\boldsymbol{x}}(\boldsymbol{\theta})|) \quad (17)$$

and, assuming $\bar{\boldsymbol{x}}(\boldsymbol{\theta})$ has an inverse,

$$p(\boldsymbol{\xi}|\boldsymbol{x}_t, M) = \delta(|\boldsymbol{\xi} - \boldsymbol{\xi}_t|). \quad (18)$$

The analytical calculation yields an intuitive result, where the grid points are weighted by their likelihood multiplied by the corresponding prior

$$p(\boldsymbol{\xi}|\boldsymbol{y}_q, M) \propto \sum_{t\in T} \delta(|\boldsymbol{\xi}-\boldsymbol{\xi}_t|)\, p(\boldsymbol{\theta}_t|M)\, N\!\left(\boldsymbol{y}_q|\bar{\boldsymbol{y}}(\boldsymbol{\theta}_t),\mathbf{C}_q\right). \quad (19)$$

In the limit of a flat prior, this is the classic MLE case, which is equivalent to the χ^2 minimization techniques used in most SED-fitting implementations today, where the measurements are compared to the simulated observations at the grid points to select the optimum. One obvious exception is the algorithm of Benítez (2000), which actually applies an explicit empirical redshift prior in this equation, hence it is often referred to as "Bayesian."

The selection of a set of templates is another simple prior but on the spectral type, even if well hidden, implicit, and not often admitted. Researchers routinely seek for templates that provide the best redshift estimates. Strictly speaking, this is cheating. The selection should be based on how well the templates represent the data in the space of the observables, and not based on their performance in the estimation. Naturally, there is a connection, but not in that direction. The templates that follow the data will likely provide better estimates; however, templates that yield good estimates are not guaranteed to match the data. The development of a class of methods by Budavári et al. (1999, 2000, 2001) and Csabai et al. (2000) can be considered early attempts to achieve a better SED prior. Here, the templates are statistically modified to represent the observations more accurately, while not optimized for redshift estimation whatsoever. Clearly, these are just the first steps in this direction. Instead of just assigning 1 and 0 weights to the templates by either including them or not (respectively) as typically done today, one can explicitly formalize more realistic priors over a broader range of SEDs that are driven by scientific knowledge and/or ensemble statistics.

An obvious but rather important improvement in the new framework is the ability to naturally introduce and utilize the uncertainties of the template spectra. We know that the models are not perfect, and this can be easily characterized. As an example, one can use the same prescription for the spectral synthesis, but build on a various stellar libraries to analyze the differences. When using empirical templates, the implementation is even more evident. We fold in the uncertainties by abandoning the simplified relation in Equation (17) and creating a more realistic model with the estimated finite errors.

3.3. *Empirical Method*

The new methodology in the limit of the classic empirical algorithms goes well beyond the usual techniques, which consist of simply establishing the fitting function in Equation (4). We can utilize those fits (or preferably estimate the densities numerically to map the full relation), but we can also properly consider the uncertainties.

The parameterization of a minimalist model is done by a position in the space of the observables, i.e., $\boldsymbol{\theta}$ is the same type of quantity as $\boldsymbol{x}$ and $\boldsymbol{y}$, e.g., *UBVI* fluxes. Namely, we choose $\bar{\boldsymbol{x}}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\bar{\boldsymbol{y}}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. Even though the observables in $\boldsymbol{x}$ and $\boldsymbol{y}$ are the same quantities, the mapping is still required to fold in the photometric errors. With an improper flat prior $p(\boldsymbol{\theta}|M) = 1$, the mapping of the observables is integrated analytically

$$p(\boldsymbol{x}_t|\boldsymbol{y}_q, M) = \int d\boldsymbol{\theta}\; N(\boldsymbol{x}_t|\boldsymbol{\theta}, \mathbf{C}_t)\, N(\boldsymbol{\theta}|\boldsymbol{y}_q, \mathbf{C}_q). \qquad (20)$$

While this model is clearly very simple, it is quite powerful and conceptually more sound than a number of traditional methods. We will use it for illustrations in the upcoming discussions.

Other simple forms of priors can also be handled analytically, e.g., linear and Gaussian that may be reasonable approximations at least locally. Otherwise we resort to the numerical evaluation.

3.4. *Advanced Methods of the Future*

The problem with the classic empirical methods is the requirement of having the same set of observables for both the training and the query sets. The limitations of the SED-fitting techniques come from the fact that the models cannot perfectly describe the relation of observables and the physical properties.

In the realm of our unified framework, we can have more advanced methods that combine these two previously separate classes of techniques. We can introduce new algorithms to take advantage of the training points even if their photometric observables differ from those in the query set. The idea is the following: True to the spirit of empirical methods, we utilize the training set to provide the relation between the physical properties we wish to constrain and some observables; see Equation (6). In addition to this empirical relation, we apply a mapping from the observables of the query set to that of the training set based on SED modeling, like in the template fitting procedures. For example, if the training set contains *UJFN* magnitudes, one can map them to *ugriz* using Equation (1).

The intriguing observation to make here is that one does not even need realistic physical models to start with, because the physics is in the training set and not the model. Let us consider a model M, which is a complete basis on the observed wavelength range, e.g., Legendre polynomials or Fourier series with the parameterization by their coefficients. The manifold of the physical spectra is naturally contained within. In practice, this model needs to be only sufficiently complete and band-limited so that real SEDs can be well described; this is a weak prior that we can set up based on all the spectra we observed and simulated before. A model spectrum corresponding to a certain parameter value in M can be convolved with the appropriate transmission curves to yield the observables $\bar{\boldsymbol{x}}(\boldsymbol{\theta})$ and $\bar{\boldsymbol{y}}(\boldsymbol{\theta})$, even if they are unphysical. Hence, formally we have the basis of our mapping, $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, M)$. As long as the data provide good enough constraints on the model parameters, the mapping is valid and the algorithm follows the routine. When the observations barely constrain the model parameters and large volumes of unphysical SEDs have significant likelihood, the mapping will be wrong. The solution is to apply a prior to consider only the physical SEDs. Using the entire catalog, one can derive an empirical physical prior statistically, which we will discuss in the next section.

These new advanced methods overcome the usual difficulties in photometric redshift estimation, and offer a way out of the half-century-old dilemma. They are a natural extension of everything that has worked before in the field: a straightforward combination of the two previously separate methodologies.

4. DISCUSSION

Next, we demonstrate the new framework in action by applying a simple model to real-life data, which is followed by discussions of the qualities of training sets and the prior.

4.1. *A Case Study*

To illustrate the concepts introduced earlier, we apply the aforementioned minimalist empirical model to a sample of galaxies. We choose SDSS sources for their well-studied photometric uncertainties. Following Scranton et al. (2005), we estimate the full covariance matrix for all objects, and utilize them in the subsequent analysis. We randomly select a quarter of the entire Main Galaxy Sample (MGS; Strauss et al. 2002) of DR6 to be the training set, roughly 100 thousand objects. Our query set is a smaller disjoint random subset for illustration purposes. First, we map the observables (magnitudes to magnitudes) analytically using our simple model in Equation (20). The calculation is done inside the DR6 database by SQL user-defined functions.

Next, we compute the conditional PDFs by a dual-tree KDE implementation (Gray & Moore 2003; Lee & Gray 2006) at preset locations defined by the T training and Q query sets in magnitude space and a uniform high-resolution redshift grid. The practical complication with any density estimation is the fact that it is scale-dependent and changes with the metric. We are further limited in our applications to fix bandwidths for the conditional density estimation in the current implementation of the estimator. We adopt a bandwidth of $h = 0.004$ in a metric that scales the magnitudes to the redshift. In other words, the resolution in redshift space is set by h, the full width half maximum of the normal distributions, and we rescale the magnitudes by a factor of $f = 0.08$ to reasonably match the density of the sources in the separate subspaces. This simple technique is expected perform reasonably well within the regime where the sources are suitably dense but not in the outskirts where a larger variable bandwidth is needed in magnitude space. The theory of more sophisticated conditional density estimation is well-studied (e.g., Fan et al. 1996), and advanced adaptive implementations are in the works to help out (Lee & Gray 2008, private communication).

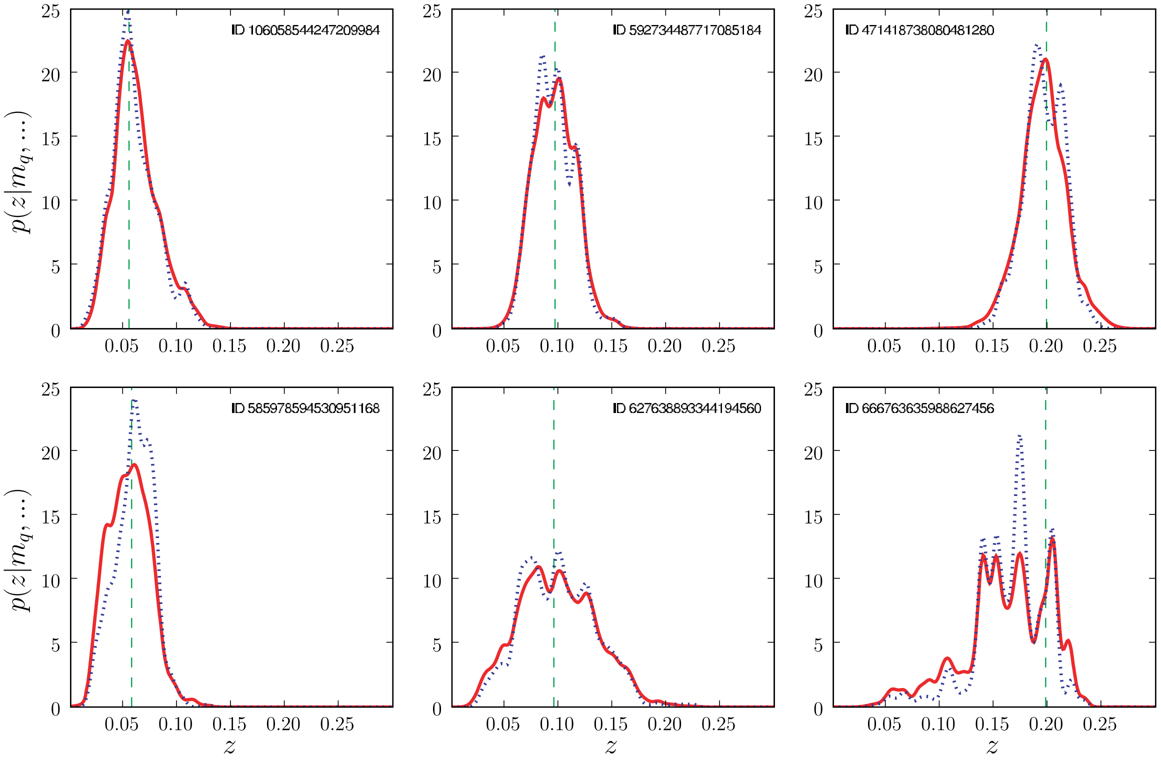

Figure 1. Probability density as a function of the redshift for early- and late-type galaxies (*upper* and *lower* panels, respectively) at different distances marked by the vertical *dashed* lines. For every object, the *dotted* line shows the empirical relation of $p(z|\boldsymbol{x} = \boldsymbol{m}_q)$, and the *solid* line illustrates the final result of $p(z|\boldsymbol{y} = \boldsymbol{m}_q, M)$ after properly folding in the photometric uncertainties via the mapping in the model.

(A color version of this figure is available in the online journal.)

Figure 1 illustrates the nature of the $\boldsymbol{x}$–$\boldsymbol{\xi}$ relation, in this case the multicolor measurements and redshift $p(z|\boldsymbol{m}_q)$, as well as the final redshift distribution, $p(z|\boldsymbol{m}_q, M)$, incorporating the photometric uncertainties in our model. We see that the redshift is really not a simple function of the magnitudes but rather a more general relation. This is even more so for observables that constrain the physical properties less than the *ugriz* measurements. The relation itself (shown as a dotted line) might provide an overly optimistic view of the uncertainties at times and usually much noisier than the final PDF (shown in solid) that sums up these relations with appropriate weights. The top panels show intrinsically red galaxies at three different redshifts, which were selected based on the mixing angle of the first two principal components, also known as the `eClass` in the SDSS terminology. The bottom panels show the more problematic blue galaxies at similar redshifts. Note the consistent performance of the estimator on the red sources as a function redshift in comparison to the blue galaxies that have broader PDFs at higher redshifts and are noisier, especially at the largest distances.

In the bottom rightmost panel, the distribution is not even centered around the spectroscopic redshift, but skewed toward lower values. This object is very close to the edge of the training set, and the result would be considered unreliable due to the lack of calibrators at higher redshift that would still be within the sources photometric uncertainties.

4.2. Sampling Frequency

A very attractive feature of the new PDF estimator derived earlier in Equation (8) is its conceptual independence from the sampling of the calibrators. Many statistical tools rely heavily on having a representative training set and only provide unbiased results in that limit. In our case, the training points simply provide locations where the evaluation is feasible and their density is essentially just a resolution factor. The sampling frequency of the training set only affects the accuracy of the numerical integral in Equation (16) but not in a systematic way as long as the query point is well within the boundaries of the window function. A denser training set will provide higher resolution in the summation, but there is a practical limit beyond which one expects no improvement. The reason is that the new calibrator sources are essentially identical to the ones already in the training set.

The number of spectroscopic measurements to be carried out for calibration purposes is limited by finite resources. It is vital to acquire reliable training sets for the new generation photometric studies. A good training set has a well-defined selection function, using criteria based on only the observables one plans to model for the estimation, and, within that, a smart adaptive sampling strategy to optimize the coverage in observable space. Clearly, the densest regions can be subsampled, but one needs all training points in the outskirts of the manifold for broad support. For this reason, the simplest random subsampling of the underlying population will not suffice. Instead, a stratified sampling strategy is to be pursued.

To demonstrate that the methodology is robust to this kind of systematic changes in the training set, we create a stratified subset and perform the previous analysis the same way. The sampling is done by including sources randomly based on their local density $p(\boldsymbol{x}|T)$ in magnitude space. A galaxy is included in the training set only if the ratio of some constant p_0 and the local density is larger than a randomly generated real number, U_{01}, uniform between 0 and 1, i.e., $p_0/p(\boldsymbol{x}|T) > U_{01}$. We set the value of p_0 to yield a subsample that is half the size of the original data set. Figure 2 shows the results for the previously

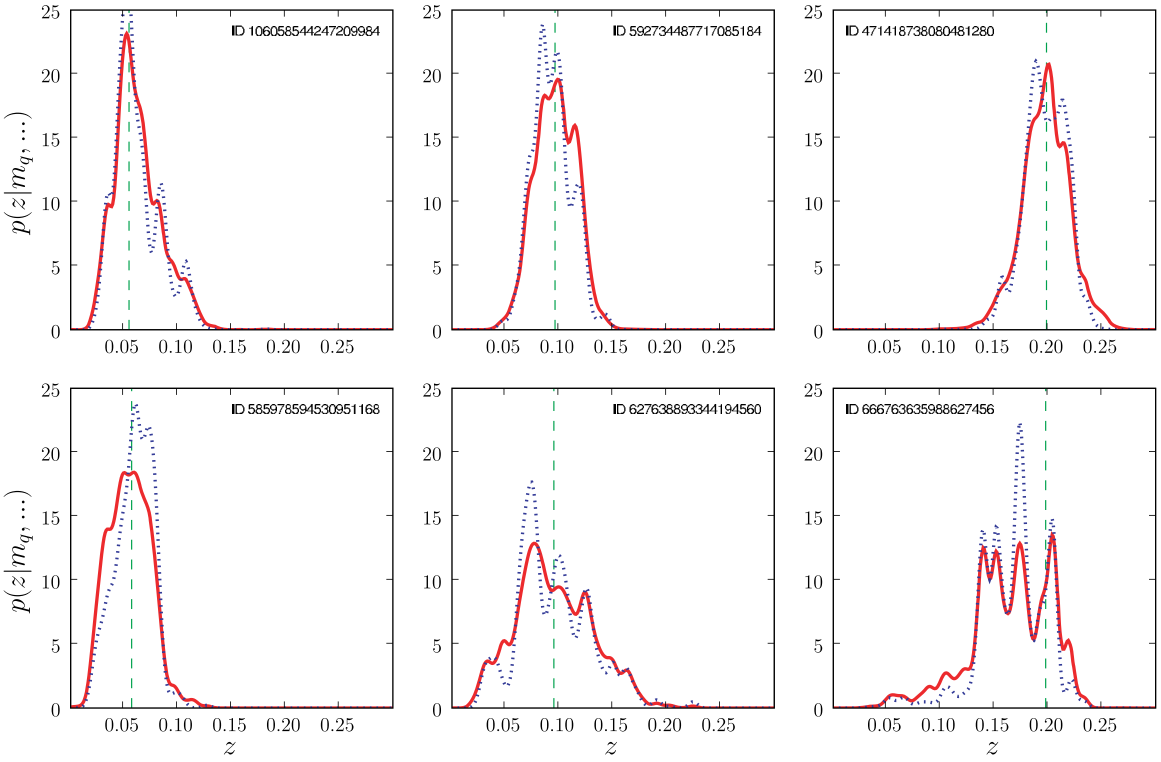

Figure 2. Results obtained from the stratified training set look much like the those from the full sample, which shows that the representativeness does not matter; instead the training sets should be optimized for the broad coverage of the observable volume with highest sampling rates in the outskirts.

(A color version of this figure is available in the online journal.)

selected sources based on the smaller stratified subset. The basic shape of the curves is practically the same in most cases, only somewhat noisier but without systematics. One exception is the blue galaxy at around $z = 0.1$, where the subsampling somewhat amplifies the effect of the large wall in SDSS at $z = 0.08$. The blue galaxy at the highest redshift is essentially unchanged (except for the part at the lowest redshift where the density in magnitude space is larger to start with) because the stratified sampling (by construction) has no effect on its already very sparse neighborhood.

Optimal sampling is difficult to achieve. In fact, it is difficult even to define. In addition to the photometric uncertainties, the desired resolution of the physical quantities also sets limits on the sampling frequency. This is prominent in the case of degenerate regions where an extended part of the physical parameter space is cramped into a small volume of observables. Simulations built on realistic SED models can help cross-check these factors, and evaluate the performance of the estimator ahead of time. In the ideal case, one would create stratified training sets in the space of the physical parameters instead of the observables, which should be more feasible in the near future with improved spectral modeling (e.g., Charlot & Bruzual 2009, in preparation).

4.3. Empirical Priors

The distribution of sources in a training set may be artificial and, as we just argued, should be optimized for coverage with a practical upper bound on the density tuned to the photometric inaccuracies and source diversity. However, the distribution in the query set is often physical and can be used to derive an empirical prior for our model. The basic observation is that the density of sources in the query set, $p(\boldsymbol{y}|Q)$, should match the predicted density of the model, $p(\boldsymbol{y}|M)$. The latter is calculated for any prior as the convolution,

$$p(\boldsymbol{y}|M) = \int d\boldsymbol{\theta}\ p(\boldsymbol{y}|\boldsymbol{\theta}, M)\, p(\boldsymbol{\theta}|M). \qquad (21)$$

If we substitute $p(\boldsymbol{y}|Q)$ measured from the sources on the left-hand side of the equation, the only unknown is the prior, which we can solve for using the elegant deconvolution technique of Richardson (1972) and Lucy (1974).

To see why a physically sensible prior is important, let us consider the density of sources in the training set within the window function. Since the density is proportional to the product of the underlying $p(\boldsymbol{x})$ density and the selection function,

$$p(\boldsymbol{x}|T) \propto p(\boldsymbol{x})\, P(T|\boldsymbol{x}), \qquad (22)$$

a significant volume of the window function is not sampled by the training set, where $p(\boldsymbol{x})$ is zero. Without a reasonable prior, the mapping $p(\boldsymbol{x}|\boldsymbol{y}_q, M)$ could yield wrong weights for unphysical observables in the summation of Equation (7). Hence, any model needs some physical information. Even if one is hesitant to take the empirical prior at face value, the domain of the model parameters should be carefully considered. In case of template fitting, this happens implicitly, even if not optimally, via the selection of the set or manifold of template spectra, but can be also done for even the empirical algorithms.

5. CONCLUSIONS

Starting from first principles of Bayesian probability theory, we built a description and obtained the solution of the generic photometric inversion problem, where the physical properties of sources are constrained based on observational measurements. The new approach yields a formalism that encapsulates the field of photometric redshift estimation, and contains the traditional methods as special cases. In our systematic analysis of the mathematical problem, we put previous techniques in context and pointed out the directions for improvement in each.

The proposed extensions to the current methods represent significant progress in more respects. We avoid the common assumption of the physical properties being a single-valued function of the observables by treating their relation in a more general way. Thus the formalism is not prone to fail in regions, where the data sets are degenerate. We showed how to estimate the corresponding probability density of this relation. In addition, the uncertainties of the observables are propagated all the way to the results via explicit modeling of the accuracies. We discussed various aspects of the modeling from the simplest empirical case to the application of SEDs.

This general framework allows for the construction of novel, more advanced methods that combine the attractive qualities of empirical and template-fitting algorithms. One can build empirical estimators based on training sets that have different observables from the query set, e.g., *UJFN* photometry to *ugriz*, via SED modeling. We can improve the methods by creating more and more realistic models that include, for example, the strengths of the emission lines in galaxies (following Győry et al. 2009) and their inclination angles (based on Yip et al. 2009, in preparation) among the model parameters to properly marginalize over the nuance parameters for a more reliable mapping of the observables.

The current limitations come from the lack of good understanding of the photometric uncertainties. From previous studies, we know that the flux measurements in various passbands are correlated, yet, most catalogs only quote errors on the individual fluxes. For more precise scientific measurements via tighter photometric constraints, we need better photometric error models in the future. Upcoming survey telescopes will observe all sources multiple times, hence will be able to get a better handle on the errors and their covariances. Understadning these systematics is probably one of the highest priority tasks in the preparation for the upcoming era of photometric science.

The proper solution of the generalized photometric inversion problem may be straightforward on paper, but efficient implementations of realistic models with appropriate priors involve many advanced concepts in statistics, and can only be built on the most recent and on-going developments in computer science, e.g., multi-dimensional indexing in databases. Even then the computations are not trivial to carry out, and have significantly higher demand for compute power than previous methods. The immediate future work is to have such a unified framework developed and ready for the next generation imaging surveys.

The author is grateful to Alexander S. Szalay, István Csabai and Andrew J. Connolly for originally introducing the problem, and for the countless inspiring discussions on various aspects of the topic over the course of many years ever since. The presented work has grown out of their collaboration on *Hubble Space Telescope* and Sloan Digital Sky Survey data. Also the author thanks Dongryeol Lee and Alexander G. Gray of the FastLab Team at the Georgia Institute of Technology for making their dual-tree KDE implementation available for this work, and for their continuing support to the astronomy community. This study was supported by the Gordon and Betty Moore Foundation via GBMF 554.

REFERENCES

Baum, W. A. 1962, IAU Symposium, 15 (New York: IAU), 390
Benítez, N. 2000, ApJ, 536, 571
Budavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., & Dickinson, M. E. 1999, in ASP Conf. Proc., 191, Photometric Redshifts and High Redshift Galaxies, ed. R. J. Weymann, L. J. Storrie–Lombardi, M. Sawicki, & R. Brunner (San Francisco, CA: ASP), 19
Budavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., & Dickinson, M. E. 2000, AJ, 120, 1588
Budavári, T., et al. 2001, AJ, 122, 1163
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Coleman, G. D., Wu, C.-C., & Weedman, D. W. 1980, ApJS, 43, 393
Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., & Munn, J. A. 1995a, AJ, 110, 2655
Csabai, I., Connolly, A. J., Szalay, A. S., & Budavári, T. 2000, AJ, 119, 69
Csabai, I., Dobos, L., Trencséni, M., Herczegh, G., Józsa, P., Purger, N., Budavári, T., & Szalay, A. S. 2007, Astron. Nachr., 328, 852
Fan, J., Yao, Q., & Tong, H. 1996, Biometrika, 83, 1, 189–206
Fukugita, M., Shimasaku, K., & Ichikawa, T. 1995, PASP, 107, 945
Gray, A., & Moore, A. W. 2003, SIAM Int. Conf. on Data Mining (Philadelphia: SIAM)
Győry, Z., et al. 2009, AJ, submitted
Koo, D.C. 1985, AJ, 90, 148
Lee, D., & Gray, A. 2006, Proc. 22nd Annu. Conf., Uncertainty in Artificial Intelligence (UAI-06), Arlington, Virginia
Lucy, L. B. 1974, AJ, 79, 745
Nadaraya, E. A. 1964, Theory Probab. Appl., 9, 141
Richardson, W. H. 1972, J. Opt. Soc. Am., 62, 55
Scranton, R., Connolly, A. J., Szalay, A. S., Lupton, R. H., Johnston, D., Budavári, T., Brinkman, J., & Fukugita, M. 2005, arXiv:astro-ph/0508564
Strauss, M. A., et al. 2002, AJ, 124, 1810
Weymann, R. J., Storrie–Lombardi, L. J., Sawicki, M., & Brunner, R., (ed.) 1999, in ASP Conf. Proc. 191, Photometric Redshifts and High–Redshift Galaxies (San Francisco, CA: ASP)
York, D. G., et al. 2000, AJ, 120, 1579