CHARACTERIZING AND PROPAGATING MODELING UNCERTAINTIES IN PHOTOMETRICALLY DERIVED REDSHIFT DISTRIBUTIONS

Augusta Abrahamse1,2, Lloyd Knox1, Samuel Schmidt1, Paul Thorman1, J. Anthony Tyson1, and Hu Zhan3

1 Department of Physics, 1 Shields Avenue, University of California, Davis, CA 95616, USA
2 Universidad Privada Boliviana, Av. Capitan Victor Ustariz Km. 6.5, Cochabamba, Bolivia
3 National Astronomical Observatories, Chinese Academy of Sciences, A20 Datun Rd, Chaoyang District, Beijing 100012, China

Received 2011 January 5; accepted 2011 March 29; published 2011 May 24

ABSTRACT

The uncertainty in the redshift distributions of galaxies has a significant potential impact on the cosmological parameter values inferred from multi-band imaging surveys. The accuracy of the photometric redshifts measured in these surveys depends not only on the quality of the flux data, but also on a number of modeling assumptions that enter into both the training set and spectral energy distribution (SED) fitting methods of photometric redshift estimation. In this work we focus on the latter, considering two types of modeling uncertainties: uncertainties in the SED template set and uncertainties in the magnitude and type priors used in a Bayesian photometric redshift estimation method. We find that SED template selection effects dominate over magnitude prior errors. We introduce a method for parameterizing the resulting ignorance of the redshift distributions, and for propagating these uncertainties to uncertainties in cosmological parameters.

Key words: cosmological parameters – cosmology: observations – galaxies: distances and redshifts – galaxies: photometry – gravitational lensing: weak
Online-only material: color figures

1. INTRODUCTION

The surprising discovery of the accelerated expansion of the universe (Riess et al. 1998; Perlmutter et al. 1999) has invigorated efforts to determine the history of the expansion rate to high precision. By providing a greater understanding of the cosmic acceleration, further observational input holds exciting prospects for improving our understanding of the fundamental laws of physics as well as the fate and possibly the origin of the universe (Albrecht et al. 2006). Many observational efforts are aimed at determining the cosmic–redshift relation via a variety of distance-dependent signals including the apparent magnitude of supernovae, the angular location of the acoustic feature in the galaxy number density correlation function, and the amplitudes of tomographic cosmic shear power spectra. See Bassett & Hlozek (2010), Howell et al. (2009), Heavens (2009), Zhan et al. (2009), and Huterer (2010) for recent reviews.

Determining the redshifts associated with these signals can be challenging. In many cases, due to the large numbers and the faintness of the objects being used for distance determinations, we must extract redshift information not from spectroscopy, but from photometry in multiple broad bands (Baum 1962; Connolly et al. 1995). Redshifts determined from such data are called photometric redshifts, or simply photo-z’s. Recent work has emphasized the stringent requirements on the quality of redshift information needed to avoid significantly degrading or biasing inferences about the dark energy posited to explain cosmic acceleration (Fernández-Soto et al. 2002; Huterer & Takada 2005; Ma et al. 2006; Albrecht et al. 2006; Zhan 2006; Abdalla et al. 2008; Bridle & King 2007; Ma & Bernstein 2008).

Traditional tomographic analysis of photometric data for cosmic shear and galaxy clustering begins with a sorting of galaxies into photo-z bins. This binning is a useful stage of data compression which, if done correctly, leads to very little loss of information. These bins can be small in number and hence fairly coarse. As Ma et al. (2006) have shown (in the context of a low-dimensional dark energy model and a relatively shallow survey) cosmological parameter uncertainties do not decrease significantly in going from five to more photo-z bins. However, significant information loss can occur if the redshift probability distribution for a given galaxy, $P(z)$, is discarded immediately after binning.

The analysis procedure we present here also begins with a sorting of galaxies. We refer to this sorting process as “collecting” rather than “photo-z binning” to emphasize two points: (1) the sorting criterion (or possibly criteria) is not restricted to redshift estimates and (2) even if the sorting is based on a redshift estimate, the distribution of the collection of galaxies with redshift will extend beyond the photo-z redshift range used to define the collection.

However the collection is defined, a weak-lensing (WL) map and a galaxy count map can then be made for each collection. From these maps, all the auto and cross-power spectra can then be calculated (Hu 1999) as well as other summary statistics such as bispectra (Takada & Jain 2004) and shear peak counts (Wang et al. 2009).

The crucial redshift information that we need in order to calculate model predictions for these power spectra (and any other summary statistics) is the redshift distribution, $dN/dz$, for each collection (Ma et al. 2006; Huterer et al. 2006; Zhan 2006). In this paper, as discussed in Cunha et al. (2009) and Wittman (2009), we sum individual $P(z)$’s for the galaxies in each collection in order to infer the $dN/dz$ for that collection.

Our estimates of $dN/dz$ are model dependent. They depend on assumptions about the distribution of galaxies with respect to spectral type, apparent magnitude in a reference band and redshift. We analytically demonstrate that, if the modeling assumptions that go into our $dN/dz$ estimates are correct, then these estimates are unbiased (see the Appendix).

Of course, the modeling assumptions cannot be perfectly correct. Thus, a crucial element of our methodology is a treatment of the uncertainty in $dN/dz$ that arises from uncertainty in the modeling. Our treatment allows for this uncertainty to be propagated through to the rest of the data reduction process. The
method for determining the uncertainty in \(dN/dz\) in a way that can be fully propagated to scientific conclusions is the chief contribution of this paper.

Past work on propagation of redshift errors through to the cosmological parameters has typically relied on highly idealized models for the distribution of the redshift errors. The redshift distributions considered are often far simpler than the distributions expected for real photometric surveys (Ilbert et al. 2006). For instance Ma et al. (2006) assume the distribution to be Gaussian, parameterized by a mean and scatter as a function of redshift. Huterer et al. (2006) assume that photo-z errors take the form of a bias that varies with \(z\). In reality, degeneracies in color space for multi-band photometric determination of redshifts can result in redshift error distributions that are asymmetric and multi-peaked, with peaks separated by \(\Delta z > 1\). As demonstrated by a growing body of research, it is important to take into account these catastrophic photo-z errors and to add complexity to the photo-z uncertainty models (Schneider et al. 2006; Amara & Réfrégier 2008; Ma & Bernstein 2008; Sun et al. 2009; Bernstein & Huterer 2010; Nishizawa et al. 2010; Hearin et al. 2010).

Although sufficiently robust spectroscopic sampling may reduce the need for modeling assumptions and the resulting uncertainties, for proposed large-scale future weak-lensing surveys such as LSST, PanSTARRS, and DES, there is much to be gained from using very large numbers of faint and distant galaxies for the distance determinations. In the case of LSST, several billion galaxies with \(i\)-band magnitudes as faint as 25 will be used. It is very challenging to spectroscopically determine redshifts for a fair sample of such faint objects. Spectroscopic samples preferentially contain galaxies with clear spectral features and are therefore biased toward specific galaxy types. Moreover, they are significantly volume incomplete several magnitudes brighter than the faint limits of the photometry. In other words, training sets used to develop photometric model inputs may be unrepresentative of the larger galaxy population, resulting in uncontrollable errors in the photo-z estimation. Thus, a method for quantifying such modeling uncertainties is an absolute necessity for understanding the resulting uncertainties in the redshift distributions and the subsequent uncertainties in the cosmological parameters.

We present here a method for parameterizing the redshift error distributions that is completely general and thus accommodates the existence of catastrophic errors. The parameterization arises naturally out of the consideration of specific modeling assumptions and data. For specificity, we use the LSST data model (described in Section 3) when giving examples of sensitivity to model assumptions.

We show that changes in modeling assumptions result in variations in the redshift distribution, \(dN/dz\), which can be captured by a principal component analysis (PCA), where just a few modes are sufficient to represent these effects. The \(dN/dz\) can be reproduced from linear combinations of these modes with the weights on each mode becoming the parameters of the new representation. The uncertainty on each weight is extracted from the variations resulting from uncertainties in modeling assumptions, and this information is then usable for any likelihood analysis into which photo-z uncertainties enter. These new variables are continuous and lend themselves to a Fisher matrix or MCMC analysis of the parameter space.

In Section 2, we review Bayesian (model-based) methods for inferring redshifts from multi-band photometry, and we introduce PCA as a method for naturally capturing the modeling uncertainties in photometric redshift estimation. In Section 3, we present a model demonstrating how uncertainty in the selection of a set of spectral energy distribution (SED) templates can be captured by PCA. In Section 4, we explore uncertainties in the type and magnitude distributions that can be used as priors on photometric redshift estimates (Benitez 2000) and demonstrate again the utility of PCA in parameterization of the modeling uncertainty. In Section 5, we relate our results to work that has propagated photo-z uncertainties through to constraints on dark energy. In Section 6, we summarize our findings.

2. INFERRING REDSHIFTS FROM MULTI-BAND PHOTOMETRY

Our method takes as a starting point a calculation of the probability distribution of the redshift of an object given its measured colors and certain modeling assumptions. We then present a simple estimator for the redshift distributions, \(dN/dz\), of collections of such galaxies. For us, the “collections” will consist of galaxies all in the same photometric redshift bin. Of course, our modeling assumptions may be incorrect. Thus, we present a method for parameterizing the resulting uncertainty in the \(dN/dz\) for each photometric redshift bin. With the uncertainty in the \(dN/dz\) parameterized, and the uncertainty in the parameters specified, the problem is now well posed for further propagation of the uncertainty in the photo-z’s to uncertainty in the cosmological parameters. The following subsections lay out each of these steps.

2.1. Bayesian Calculation of Redshift Probability Distributions

Let us consider \(t\) to be a variable (or set of variables) that uniquely specifies the SED of a galaxy. Then specifying for a galaxy its \(t\), redshift \(z\), and apparent magnitude in a single reference wavelength band, \(m_1\), is sufficient for determining the fluxes, \(f_\lambda\), in a set of well-defined bands, where \(\lambda = 1\) we will take to be the reference band. If we model the measured fluxes (the data), as due to the true fluxes plus noise terms \(f_\lambda^d = f_\lambda + n_\lambda\) and we assume that the \(n_\lambda\) are Gaussian random variables with mean zero and \(\langle n_\lambda \rangle = \sigma_\lambda^2\delta_{\lambda\lambda}\), then

\[
P(f_\lambda^d | t, z, m_1, \sigma_\lambda^2) = \frac{1}{(2\pi)^{N/2} \prod_\lambda \sigma_\lambda} \exp \left[ -\frac{\sum_\lambda (f_\lambda^d - f_\lambda(t, z, m_1))^2}{2\sigma_\lambda^2} \right].
\] (1)

We can turn this around, using Bayes’ theorem, to get an expression for the probability distribution of the unknowns \(t, z, m_1\) in terms of the knowns, the \(f_\lambda^d\)’s, and whatever assumptions we care to make, which we shall call \(J\) (for Information). According to Bayes’ theorem, which follows from axiomatic properties of probability distributions,

\[
P(t, z, m_1 | f_\lambda^d, I) = \frac{P(f_\lambda^d | t, z, m_1, \sigma_\lambda^2)P(t, z, m_1 | I)}{P(f_\lambda^d | I)}.
\] (2)

The probability distribution on the left-hand side is called the “posterior” distribution (because it is the probability distribution after we have the data), and the \(P(t, z, m_1 | I)\) is called the “prior” distribution (because it is the probability distribution for the parameters prior to our collection of the data). The prior is where all our modeling assumptions are encapsulated. The term in the denominator we can ignore since it has no dependence on
the model parameters. While \( m_1 \) is formally a measured quantity derived from one of the band fluxes, it is treated as an unknown here because we want to map variations in \( P(z) \) by perturbing it within its Gaussian noise distribution.

If we are interested in only the redshift, we can obtain its (marginal) distribution by integrating over all possible values of \( t \) and \( m_1 \) so that

\[
P(z) \equiv P(z|\mathbf{f}_z^d, I) = \int dt \int dm_1 P(t, z, m_1|\mathbf{f}_z^d, I),
\]

where we have written the distribution simply as \( P(z) \), suppressing the dependence on the data and \( I \) (modeling assumptions) for notational convenience.\(^4\) Although we have ignored the \( P(\mathbf{f}_z^d|I) \) factor, we can now make up for that by choosing a normalization factor such that \( \int_0^\infty P(z)dz = 1 \). A number of authors have used Bayes’ theorem in this way to obtain the marginalized probability distribution of the redshift (see for instance Benitez 2000; Edmondson et al. 2006; Chapin et al. 2004).

2.2. Estimating \( dN/dz \)

Often the redshift distribution is reduced to a single value of redshift, \( z_p \), perhaps the highest peak in \( P(z) \), and then this value of \( z \) is referred to as a Bayesian photometric redshift estimate. Here, we use \( z_p \) to sort galaxies into photometric redshift bins (collections) but use all the information in the \( P(z) \) for each galaxy to estimate \( dN/dz \) for the collection.

As various authors have noted (Bordoloi et al. 2010; Cunha et al. 2009; Mandelbaum et al. 2008; Wittman 2009), the use of the full \( P(z) \) can improve on the systematic errors in photometric redshift estimation. Errors in photometric redshifts are often due to degeneracies in color space. (For instance, when an object’s spectrum is represented by the fluxes in low-resolution optical bands, the Lyman break at \( z \approx 3 \) may be indistinguishable from the 4000 Å break at low redshift.) Although this might result in a drastically miscalculated point estimate redshift (i.e., catastrophic error), the \( P(z) \) for most galaxy types near this redshift is multimodal. Because inference of cosmological parameters (for instance through weak-lensing observations) relies less on the redshift of specific galaxies and more on the distribution of galaxies in a collection, the use of the full \( P(z) \) is a vast improvement over point estimates. Consider Figure 1: the maximum of the \( P(z) \) is at low redshift, whereas the true redshift is near \( z \approx 3 \). In a point estimate approach this would be a catastrophic error, but the full \( P(z) \) assigns almost half of the probability to the galaxy being near its true redshift.

Enumerating the galaxies in a given collection (redshift bin) with index \( g \), we define our estimate for \( dN/dz \) in that collection as

\[
\frac{dN}{dz}(z) = \sum_g P_g(z).
\]

The estimate is motivated by the fact that the probability distribution of the redshift of a galaxy drawn at random from the collection is given by \( \frac{1}{N} \sum G g \) \( P_g(z) \), where \( N \) is the number of galaxies in the collection. The Appendix proves that this intuitive estimator is not biased as long as the prior and the likelihood in Equation (2) are known perfectly.

The advantage to calculating the \( dN/dz \) in this way can be seen in Figure 2, which shows the \( dN/dz \) calculated for a mock galaxy catalog of 100,000 galaxies consisting of the six CWWSB templates which follow the type distributions as a function of magnitude given by Benitez (2000; but with a fixed \( dN/dz \) peaking at \( z = 2.0 \) for simplicity). The full \( P(z) \) is calculated for each galaxy, and the galaxies are sorted into redshift bins, or “collections,” based on the peak value of the \( P(z) \) (in other words, the photo-z point estimate). We compare the actual distribution, \( dN/dz_{\text{true}} \), to the \( dN/dz \) calculated from summing and histogramming the single-point photo-z estimates, \( dN/dz_{\text{peak}} \). Because the calculated distribution, \( dN/dz_{\text{peak}} \), is constructed from the same photo-z point estimates that were used to assign each object to a collection (bin), the estimated redshift value necessarily lies within the redshift range for which the collection (bin) is defined; thus the calculated \( dN/dz_{\text{peak}} \) will always be zero outside that redshift range (or photometric “redshift bin”). This is a major failing of the point estimator method since the true redshift distribution of galaxies, \( dN/dz_{\text{true}} \), may have tails that extend to redshifts above and below the redshift values for which the collection is defined. By using the full \( P(z) \), however, a galaxy placed in a collection can have a non-zero probability extending to values outside of redshift range of the collection. Thus, the distribution calculated using the \( P(z) \) of each galaxy, \( dN/dz_{\text{prob}} \), can more accurately reflect the true redshift distribution.

Despite the use of the full \( P(z) \) there are many possible sources of error in the photo-z estimation that could still cause us to miscalculate the distribution of galaxies, \( dN/dz \). These include use of a small number of galaxies, mischaracterization of the statistical properties of the flux noise, and errors in the modeling assumptions summarized by the prior, \( P(t, z, m_1|I) \). The latter includes prior assumptions about galaxy distributions as a function of type, redshift, magnitude, surface brightness, etc., and also includes the choice of SED templates. Errors from these modeling assumptions are the central concern of this paper.

2.3. Parameterizing Modeling Uncertainty

We capture the effect of modeling uncertainty by explicitly varying the modeling assumptions and seeing how the \( dN/dz \)’s
Here \( i \) indexes the points in \( z \) space for which \( dN/dz \) is defined and \( \alpha \) runs over the variations. Subtracting the mean of each row from each element in the row of \( \eta \), we have
\[
D_{\alpha i} = \eta_{\alpha i} - \bar{\eta}_i,
\]
where
\[
\bar{\eta}_i = \frac{1}{m} \sum_\alpha \eta_{\alpha i}
\]
and \( m \) is the number of variations. The covariance matrix is
\[
C = \frac{1}{m-1} D D^\top.
\]

Computing the eigenvalues \( (\lambda) \) and eigenvectors \( (U) \), the first principal component \( U_1 \) corresponds to the largest eigenvalue, the second principal component \( U_2 \) corresponds to the next largest eigenvalue, and so on. Each data vector (indexed by \( \alpha \)) can be written as a linear combination of the principal components, i.e.,
\[
\eta_{\alpha i} = \bar{\eta}_i + \sum_{\beta=1}^N B_{\alpha \beta} U_{i \beta},
\]
where the weights, \( B_{\alpha \beta} \), are given by the linear transformation
\[
B = U^\top \eta.
\]

There are \( N \) principal components of length \( N \), where again, \( N \) is the number of discretized values of \( z \). If the eigenvalues decrease sufficiently quickly from biggest to smallest, then most of the variation in \( \eta_{\alpha i} \) is described by the first few modes. By keeping only the first \( k \) modes, we are able to reduce the dimensionality of the parameter space. We can reproduce the data via
\[
\eta_{\alpha i} = \bar{\eta}_i + \sum_{\beta=1}^k B_{\alpha \beta} U_{i \beta},
\]
where \( \eta_{\alpha i} \equiv \eta_{\alpha i} \) for a small value of \( k \) if the eigenvalue spectrum is sufficiently steep.

The above description was for a single collection (“photo-z bin”) defined for a given redshift range. It can be repeated for each of the \( \gamma = 1, ..., n_z \) photo-z collections; we will distinguish quantities for collections defined for different redshift ranges with a superscript \( \gamma \).

We can now finally define our parameterization of the redshift distribution as the amplitudes of the modes \( U_{i \beta}^{\gamma} \) defined by the above process. We will call these amplitudes \( b_{\beta}^{\gamma} \) and they influence the redshift distributions \( \eta_i^{\gamma} \equiv dN/dz_i^{\gamma} \) via
\[
\eta_i^{\gamma} = \bar{\eta}_i + \sum_{\beta=1}^k b_{\beta}^{\gamma} U_{i \beta}.
\]

Each modeling variation, \( \alpha \), leads to a particular value of \( b_{\beta}^{\gamma} = B_{\alpha \beta}^{\gamma} \) and we can use these “realizations” to model the prior distribution of the \( b_{\beta}^{\gamma} \), in this lower, \( k \)-dimensional space. The \( b_{\beta}^{\gamma} \) are not fixed at the discrete values, but are allowed to vary continuously to describe the full space of model uncertainty. This distribution will have zero mean due to the mean removal that occurs prior to the principal component decomposition. We can estimate a covariance matrix via
\[
C_{\gamma \beta, \gamma' \beta'} \equiv \langle b_{\beta}^{\gamma} b_{\beta'}^{\gamma'} \rangle \simeq \frac{1}{m-1} \sum_{\alpha=1}^m B_{\alpha \beta}^{\gamma} B_{\alpha \beta'}^{\gamma'}.
\]
2.4. Relation to the Estimation of Cosmological Parameters

For the analysis of cosmic shear data, we assume that it is possible to take a set of model shear power spectra, \( C_{\ell}^{\gamma \nu} \), and calculate its likelihood, given the shear data. We further assume that one can take a specification of cosmological parameters, and a redshift distribution for each photo-z collection, and use these to calculate the model \( C_{\ell}^{\gamma \nu} \). From the power spectra likelihood, any prior constraints on the cosmological parameters (for example from cosmic microwave background observations), and a prior constraint on the redshift distribution parameters, one can then form a joint posterior distribution for the cosmological plus redshift distribution parameters:

\[
\ln P_{\text{posterior}}(\theta_{\text{cos}}) = \ln L(\text{shear data}|C_{\ell}^{\gamma \nu}(\theta_{\text{cos}}, b_{\gamma}^p)) \\
+ \ln P_{\text{prior}}(\theta_{\text{cos}}) + \ln P_{\text{prior}}(b_{\gamma}^p),
\]

where

\[
\ln P_{\text{prior}}(b_{\gamma}^p) = b_{\gamma}^p C_{\beta}^{-1} b_{\gamma}^p b_{\gamma}^p / 2.
\]

With the ability to calculate this joint posterior, one would then be able to explore the constraints on this parameter space given by the data via Markov Chain Monte Carlo, for example.

3. EXAMPLE 1: TEMPLATE UNCERTAINTY

One of the potentially largest sources of modeling errors in photometric redshift estimation may come from the selection of the template set in the SED fitting method, the technique currently used for deep and faint redshift surveys. At present, many photo-z codes use a small number of template spectra from nearby galaxies, e.g., the popular CWW+SB set (Coleman et al. 1980; Kinney et al. 1996) used in part of our analysis. Though the SEDs are sometimes “tweaked” to better fit the data (e.g., Budavári et al. 2000), expanding beyond such simple template sets almost always involves the use of population synthesis models (Brammer et al. 2008; Ilbert et al. 2009), due to the lack of sizeable numbers of high signal-to-noise spectra at high redshift. The difficulty of spectroscopic follow-up of faint, high-redshift galaxies means these models may be incomplete and unrepresentative of true galaxy evolution. Such biases can have a significant effect on the photo-z estimations (see for instance MacDonald & Bernstein 2010). Because the SED shape for some population of galaxies may evolve, or a population not represented at low redshifts/high luminosities is present in the deep photometric data, the choice of template set is an important source of modeling uncertainty.

There are also convergence issues. Since the photo-z is a nonlinear inverse problem, a small amount of photometry noise will drive a noisy assignment of templates for each galaxy. The amount of this induced template noise is a non-linear function of the photometry noise. This is a general feature of threshold systems: the signal-to-noise ratio peaks at a small non-zero value of input noise (Wiesenfeld & Moss 1995). At zero noise the photo-z error is large and due entirely to errors in the modeling assumptions, whereas at high noise levels the photo-z error is large because no template assignment converges. In this paper, we focus on photo-z errors due to errors in models of SED templates and magnitude and type priors at constant photometry noise given by the LSST data model.

To begin to explore the sorts of uncertainties present in an actual survey with noise and where galaxy types may be left out or incorrectly modeled, we start by generating a mock catalog based on 20 SED templates. We then calculate the photometric redshifts while systematically leaving out one each of the SEDs at a time. When one of the templates is removed, the galaxies in the catalog matching this SED will be fit to another incorrect SED, creating errors in the photometric redshifts and variations in the resulting \( dN/dz \). The result is a set of 21 distinct photometric redshift estimates for the galaxies in the catalog, which we can use to develop a parameterization and model of uncertainty for the redshift distribution. (We note that this is just one of the many ways in which template noise can be examined.)

3.1. Mock Data

To generate our set of templates, we begin with a catalog of galaxies from GOODS (Giavalisco et al. 2004) with measured spectroscopic redshifts and 17 band photometry, covering \( U \) band to IRAC mid-infrared wavelengths, available as part of the Photo-z Accuracy and Testing (PHAT) program5 (Hildebrandt et al. 2010). Although there are hundreds of galaxies in the sample, for tractability we develop a more manageable set of 20 SED templates that are representative of the larger data set.

We derive our template set using the methodology of Assef et al. (2010, 2008) in which the 17 band data for each galaxy is fit with a non-negative linear combination of four basis templates (Elliptical, Spiral, Irregular, and active galactic nucleus (AGN)) shifted to the measured redshift. Modifying the code available at http://www.astronomy.ohio-state.edu/~rjass/irft/index.html to allow for the extraction of the coefficient associated with each component of each component and eliminating galaxies with a strong AGN component, we arrive at a sample of 375 galaxies, each described by a linear combination of E, Sbc, and Im templates. Examination of this reduced three-dimensional space reveals that the galaxies’ coefficients are not significantly clustered. In order to choose a more manageable, but still representative, set of templates that spans the data set, we run a simple \( k \)-means algorithm (MacQueen 1967) to choose 20 sets of coefficients that will comprise our mock galaxy template set. The resulting templates are shown in Figure 3. To facilitate the following step, we categorize the templates in terms of their dominant morphological type, E, or \( T_1 \), Sbc or \( T_2 \), and Im, \( T_3 \), with the reddest templates 1–6 corresponding to \( T_1 \), templates 7–13 to \( T_2 \), and the bluest templates 14–20 corresponding to \( T_3 \).

We generate magnitudes for the LSST ugrizy filter set, with depth and error properties expected for the “gold” sample described in Section 3 of the LSST Science Book (LSST Science Collaborations and LSST Project et al. 2009), which consists of expected full 10 year depth and cut to \( i < 25.0 \) mag. The 5\( \sigma \) limiting magnitudes in the ugrizy bands are 26.1, 27.4, 27.5, 26.8, 26.0, 24.8. The “gold” sample galaxies will have \( S/N \gtrsim 20 \) in multiple filters, which is highly recommended for meaningful photo-z measurements. We begin by generating an \( r \)-band apparent magnitude that follows \( p(m) \propto 10^{0.37m} \) to approximate a typical number count distribution. We then choose one of the 20 template types for each galaxy by Monte Carlo, such that the morphological types, \( T_1 \), \( T_2 \), and \( T_3 \) are distributed according to

\[
p(T|m_0) = f_r \exp[-k_r(m_0 - 20)].
\]

We then assign via Monte Carlo a redshift to each galaxy, but such that

\[
p(z|T, m_0) \propto z^m \exp \left\{ - \frac{z}{z_m(m_0)} \right\}^{q^y}
\]

\( ^5 \) Further information on PHAT, including the data set used here is available at http://www.astro.caltech.edu/twiki_phat/bin/view/Main/WebHome.
is satisfied, where \( E_{ft} \) \( S_{bc} \), or \( Im \). The free parameters, \( f_t \), \( k_t \), \( z_0 \), and \( \alpha_t \), are fit to a training set of spectroscopically measured redshifts. We take these to be the values given in Benitez (2000; listed in Table 1). Given the redshift, type, and \( m \)-band magnitude, the expected magnitudes in the remaining \( u, g, i, z, y \) bands are assigned accordingly.

### 3.2. Results

With the fiducial catalog generated, we can begin to explore how template noise affects the photo-\( z \) distribution. We run the photo-\( z \) estimation code BPZ (Benitez 2000) to calculate the \( P(z) \) for each galaxy in the catalog. We perform 21 iterations—once with the fiducial model, e.g., all 20 SEDs included for the estimation, and then 20 variations, each realized by leaving out one of the 20 SED templates.

Each incomplete set of SEDs produces slightly different errors in the photo-\( z \) estimations, which result in variations in the \( dN/dz \) for each collection. The \( dN/dz \) is stored as a vector of 134 “microbins” with a width of .03 from \( z = 0 \) to \( z = 4 \). Hence the sum,

\[
N_\gamma = \sum_{j=1}^{134} dN/dz_{\gamma j},
\] (18)

returns the total number of galaxies per collection \( \gamma \), and summing over the galaxies in each collection returns the total number of galaxies in the catalog, e.g., 100,000 = \( \sum_{\gamma} N_\gamma \).

In Figure 4, the \( dN/dz \) are shown for each realization of the templates. The bottom panels show the residuals, \( dN/dz_i - dN/dz_{\text{fid}}, \) where \( i \) refers to a realization where a single template has been left out for the photo-\( z \) estimation. It can be seen that the residuals are on the order of a few percent when compared to the \( dN/dz \).

The distribution of galaxies can be described by the mean and variance of the \( dN/dz_i \) per collection \( \gamma \) and are different for each template set realization. (Note that the collection index \( \gamma \) has been suppressed for clarity.)

\[
\langle z_i \rangle = \int dz z dN/dz
\] (19)

and

\[
\Delta_i = \sigma_i^2 = \int dz (z - \langle z_i \rangle)^2 dN/dz,
\] (20)

where \( i \) refers to the specific template choice. We calculate the mean and the standard deviation of the previous quantities, i.e.,

\[
\mu_X = \frac{1}{21} \sum_{i=1}^{21} (X_i)
\] (21)

\[
\sigma_X = \sqrt{\frac{1}{21} \sum_{i=1}^{21} (X_i - \bar{X})^2},
\] (22)

where \( X \) refers to either the mean or the variance of each variation. These values are shown in Figure 5. (To keep the axes on the same scale, we actually plot the average offset for the bin, \( \mu_{<z>} - z_{\text{center}.} \)

The spread in the means is typically on the order of \( 10^{-3} \), although for collection 4 it is larger. This reflects the difficulties in calculating photometric redshifts for the redshift range for which collection 4 is defined, \( 2.4 < z < 3.2 \). Specifically, the Lyman break at \( z \sim 3 \) is easily confused with the 4000 Å at low redshift. (This is the cause of the low redshift “bump” in the \( dN/dz \) in collection 4 seen in Figure 4.) This effect is even more

---

**Table 1**

| Spectral Type | \( f_t \) | \( \alpha_t \) | \( z_0 \) | \( k_{nw} \) | \( k_t \) |
|--------------|-------|-------|-------|--------|-------|
| E/SO         | 0.35  | 2.46  | 0.431 | 0.0913 | 0.450 |
| Sbc, Scd     | 0.50  | 1.81  | 0.390 | 0.0636 | 0.147 |
| Irr          | 0.15  | 0.91  | 0.0626| 0.123  |

**Figure 3.** 20 LRT SED templates derived from GOODSNS and linear combinations of the Assef basis templates. The galaxies are further broken into three type classifications which follow the redshift distributions as parameterized in Table 1.

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

**Figure 4.** LRT template sets: top panels show the \( dN/dz \)’s from summing the \( P(z) \) of galaxies in each collection defined for a given redshift range. The bottom panels show the residuals for each realization—that is, the difference between the fiducial (all 20) template set and the reduced (19 only of the 20) template sets, \( \Delta dN/dz_i = dN/dz_{\text{fid}} - dN/dz_i \), for each template set realization.

(A color version of this figure is available in the online journal.)
apparent in the standard deviation of the variance for collection 4. The variance is sensitive to probability moving into and out of outlying redshift “islands.” To a lesser extent, this effect is also visible in the spread of the variance for collection 1. The effect is smaller, however, due to the large numbers of bright galaxies that also populate collection 1 and are never confused with distant galaxies. (Note the asymmetry of this degeneracy, which makes this effect more pronounced in the higher redshift bins: bright objects are unmistakably at low redshift, whereas faint objects may either be low redshift intrinsically faint galaxies or high redshift intrinsically bright galaxies.) We calculate the principal components of the \(dN/dz\) due to the variations of the template sets. The fraction of the variance captured by a given mode \(i\) is \(\lambda_i = \sum_j \lambda_j\), where \(\lambda_j\) is an eigenvalue of the covariance matrix corresponding to the \(i\)th mode. It is difficult to assign a physical interpretation to the modes themselves, which show a complex and data-dependent structure. The modes will be a function of the methodology used to vary input model, e.g., we could have chosen to leave out two templates at a time and the modes returned would have shown a different structure; however, they will always represent the “directions” in the data, rank ordered by the amount of variance captured. As an example, the first three modes of collection 4 are plotted in Figure 6. Some features of the underlying data can be seen in the modes. The high amplitude of the modes at low redshift, for instance, reflects the fact that the template set variations move probability of certain galaxies into and out of this low-redshift outlying region.

As noted above, the principal components are useful for dimensionality reduction. For all collections, the first three principal components cumulatively account for 80% or more of the variance. The eigenvalue spectrum is plotted in Figure 6, where it can be seen that higher number modes quickly approach zero.

Recall from Section 2.3 that the first step of PCA is to subtract the mean from each \(dN/dz_\alpha\), where \(\alpha\) refers to a specific realization of template set. We define \(d_{\alpha}\) to be the set of mean-subtracted data, i.e., \(dN/dz_\alpha = (dN/dz)\). Each \(d_{\alpha}\) can be reconstructed from a specific linear combination of principal components. Performing the reconstruction of the \(dN/dz\)’s of the low resolution template (LRT) variations with the first three modes, it is possible to recover 80% or more of the variance per collection. In general, the precise amount of variance recovered with the first three modes will depend on the specific methodology used to vary the photo-z modeling assumptions and also on the data set. In general, the variance and the fidelity of the reconstruction can be “tuned” by retaining a larger or smaller number of the modes, e.g., if 1% accuracy is required, we can retain seven or eight modes that capture 99% of the variance. In the case of three modes retained and five tomographic bins, the photo-z uncertainty is parameterized by just 15 free parameters. This is a relatively tractable number considering analyses propagating redshift uncertainties through to the cosmological parameters have used 60 or more free parameters (Ma et al. 2006; Zhan 2006; Ma & Bernstein 2008).

A potential concern is that there is a mode (or modes) of \(dN/dz\), discarded by the truncation, which, though not contributing much to the variance of \(dN/dz\), causes a change to the shear power spectrum (or other observable) with the same shape as caused by a cosmological parameter variation. Thus, although the amplitude of the variation is small, it could lead to bias in parameter estimation. We think such a scenario is unlikely, but one could test for it when this work is extended all the way to cosmological parameter estimation.

Additionally we note that knowledge of the specific redshift range of \(dN/dz\) to which a cosmological parameter is most
Figure 7. LRT set: depicts $d_{\Delta N}/dz$ as the case of all 20 SED templates. The blue (solid) line is the actual residual. The cyan (dashed) line is the residual reconstructed from the first three principal components. Although some of the details (wiggles) are missed, the general shape is reproduced. The addition of more three principal components. While some of the detailed structure is missed, the overall shape is reproduced. The cyan (dashed) line is the reconstruction from the first modes over 80% of the variance can be accounted for with modes 1–3—little is described by the higher number modes. (A color version of this figure is available in the online journal.)

What about the variance of the variation? In the top panel of Figure 8, we plot the standard deviation of the variations $d_{\Delta N}/dz$ relative to those at $z = 4$ relative to those at $z = 1$. It may be worth exploring such an optimization in the future.

Figure 8 shows $d_{\Delta N}/dz$ for the case of all 20 SED templates. The blue (solid) line is the actual residual, and the cyan (dashed) line is the reconstruction from the first three principal components. While some of the detailed structure is missed, the overall shape is reproduced. The addition of more modes would continue to improve the reconstructed $d_{\Delta N}/dz$.

What about the variance of the variation? In the top panel of Figure 8, we plot the standard deviation of the set of $d_{\Delta N}/dz$ at each point in redshift space. This essentially captures the uncertainty in the photometric redshift calculations as a function of redshift. However, it is not necessary to retain all the modes to capture this level of error; rather, with just three modes over 80% of this uncertainty can be represented. This is illustrated in the bottom panel of Figure 8 where we reconstruct the $d_{\Delta N}/dz$ only using modes 4 and higher and recalculate the standard deviation as a function of $z$ using these reconstructed $d_{\Delta N}/dz$. It can be seen that the standard deviation from the reconstruction is greatly reduced for all collections—in other words, a large portion of the variance is captured by the first three principal components and little is accounted for by the higher order modes.

4. EXAMPLE 2: MAGNITUDE AND TYPE PRIOR DISTRIBUTION UNCERTAINTY

Another modeling assumption that can affect the photometric redshift estimations is the assumed distribution for redshifts as a function of type and magnitude. In the Bayesian framework, these distributions can be applied as a prior in the SED template fitting method with the intent to improve photometric redshift estimates. In Benitez (2000), these priors are derived semi-empirically, that is to say, a functional form is chosen and the parameters are fit with a training set of spectroscopic redshifts. The probability of a galaxy being at a certain redshift is given by

$$p(z|C, m) \propto \sum_T p(T|m_0)p(C|z, T)$$  \hspace{1cm} (23)

and using the product rule

$$p(z, T|m_0) = p(T|m_0)p(z|T, m_0).$$  \hspace{1cm} (24)

The distributions, $p(T|m_0)$ and $p(z|T, m_0)$, are constrained using auxiliary data. Errors may be introduced into these priors in a number of ways. For instance, a biased spectroscopic redshift sample (one that underrepresents a certain galaxy population present in the photometric sample) could cause errors in the inferred type distributions. An incorrect analytical form could also be chosen that is unable to accurately represent the prior distributions; or simple statistical errors, such as those resulting from the finite training set, could cause the prior parameters to be known only to low accuracy. In this section, we show again that PCA can be used to describe uncertainties in redshift estimates resulting from such errors in the magnitude and type priors.
To investigate the effects of imperfect knowledge of the Bayesian prior, we generate a mock “training” set of 100,000 galaxies. The simulated data are created according to the procedure described in Section 3: the generating r-band magnitudes, types, and redshifts via Monte Carlo, while following the distributions for $p(T|m_0)$ and $p(z|T, m_0)$, given in Equations (16) and (17), with the free parameters of the distribution given in Table 1.

Instead of using the LRT template set as a basis for generating the 17 band photometry, we use the CWW+KIN set that are included with the BPZ software, assigning elliptical/reddest as type 1 and SB2/bluest as type 6. Rather than limiting ourselves to an overly simplified model with only six discrete types, which would result in very few type misidentifications, we create a continuous color distribution by linearly interpolating between the galaxy types (a continuous number between 0.8 and 6.2) to assign color to each galaxy.

### 4.2. Results

We use the mock catalog as a starting point for probing the uncertainty in the prior parameters, $f_t$, $k_t$, $z_0$, and $\alpha_t$, and investigating the effects of this uncertainty on the photometric redshift estimations. We run BPZ to fit each galaxy to one of the standard six HDFN templates. (Because of the interpolation step (described above), some galaxies will necessarily be mistyped and these errors, on top of statistical errors associated with a finite training set, could further affect the estimation of the prior parameters.)

With this output, we use the Monte Carlo Markov Chain (MCMC) method (Metropolis Hastings algorithm) to find the best-fit values and the confidence bounds on $f_t$, $k_t$, $z_0$, and $\alpha_t$. (Note that although there are many estimators that may have worked in our case, MCMC was chosen because of the high dimensionality of the parameter space: 14 free parameters.) The results are presented in Table 2. It can be seen that these values are close to, but not the same as, the fiducial parameters in Table 1.

To probe the effects of variations in the priors, we select 100 values of the parameters drawn from the Markov chain. To make sure we consider both best and worst case scenarios, we also add the best-fit values and parameter values at the upper and lower 2σ confidence levels (note that when all the parameters are at the 2σ confidence limit, this reflects a much larger than 2σ fluctuation.)

We run the photo-z redshift code, BPZ, on our simulated galaxy catalog for each instance of prior parameters. As in the template uncertainty case, for each galaxy in our catalog, the product of the code is a probability distribution, $P(z)$, for $0 < z < 4$.

As for the LRT template variations, we sort the galaxies into collections by defining five redshift ranges evenly spaced from $z = 0$ to $z = 4$ and by using the criteria that the peak of the $P(z)$ fall within one of these ranges to assign each galaxy to a given collection or “redshift bin.” As for the previous case, the $dN/dz_i$ for each collection, $i$, is calculated by summing the $P(z)$ for each galaxy in the collection.

Variations in the priors create changes in the estimated $dN/dz$. The $dN/dz_i$ and the residuals, $dN/dz_i − dN/dz_{\text{best fit}}$, is plotted in Figure 9. The differences between the overall $dN/dz$’s plotted here and those of the LRT model (shown in Figure 4) are due to the mock data and the template set used for the photo-z estimation (not the priors). While we avoid the overly simplistic six-template (CWW+SB) model by generating data to include linear combinations of the SEDs, we only compare to the six discrete HDFN templates to estimate the redshifts. Thus, not every galaxy is accurately represented in the template library. It is the non-representativeness of these templates that cause the deviations seen in Figure 9. (Note that in the LRT case, we left one template out of 20 instead of using six to represent a continuous mix of types.) The variations in the $dN/dz$ due to variations in the prior are small, as can be seen by the difference between the $dN/dz$ using the best-fit prior parameters and the $dN/dz$’s from each Monte Carlo sampling $dN/dz_i − dN/dz_{\text{best fit}}$. They are a factor of five or more smaller when compared to the variations due to the template uncertainty. Thus, at least for the case of a well-sampled training set, small errors in the prior are not as important as having well-calibrated SED template set.

Once again, we perform PCA on the set of $dN/dz$ to derive a lower dimensional parameterization of the photo-z uncertainties. We find that for variations of the priors, only two modes are necessary to account for over 80% of the variance. The top panels of Figure 10 show the first two principal components for each collection. The bottom panels show the eigenvalue spectrum for each collection.

One caveat to this approach is that we have not included the differences in the $dN/dz$’s resulting from the difference between the fiducial prior parameters and the best-fit parameters. However, these differences in the $dN/dz$ remain small and are of the same order as the variations we do consider. We note that in a case where the type distributions of the training set

### Table 2

| Spectral Type | $f_t$ | $k_t$ | $z_0$ | $\alpha_t$ | $k_{\text{tot}}$ | $k_1$ |
|---------------|-------|-------|-------|------------|------------------|------|
| E/SO          | 0.305 | 2.350 | 0.416 | 0.0935     | 0.368            |
| Sbc, Scd      | 0.503 | 1.640 | 0.356 | 0.070      | 0.152            |
| Irr           | 0.192 | 0.902 | 0.0687| 0.119      |                  |
are more drastically biased, the prior parameter distribution may have a greater impact. What we have presented here is a method for exploring such uncertainties and parameterizing their effects on the redshift distributions in such a way that they could be propagated through to error bounds on cosmological parameters.

5. CONNECTION TO OTHER WORK

Many authors have considered the propagation of photo-z errors through to the cosmological parameters, see for instance Huterer & Takada (2005), Ma et al. (2006), Albrecht et al. (2006), Zhan (2006), Abdalla et al. (2008), Bridle & King (2007), and Ma & Bernstein (2008). A commonly quoted result is from Ma et al. (2006), where it was shown that for a photo-z distribution modeled as a Gaussian with redshift-dependent mean and standard deviation, that for a two-parameter dark energy model (the dark energy density $\Omega_{DE}$, its equation of state today $w_0 = p_{DE}/\rho_{DE}$), the 1σ errors on $z_{\text{bias}}(z)$ and $\sigma(z)$ must be less than 0.01 per “bin” in order to avoid a degradation in the dark energy parameters of more than 50%. Note that this result does not apply to the mean and bias for the tomographic redshift bins (collection in the terminology of this paper), rather the analysis refers to the mean and bias on “microbins” of width $\delta z = 0.1$. Moreover, it does not include the powerful self-calibration enabled by a joint analysis of WL and baryon acoustic oscillations, due to the shared large-scale structure.

To put in perspective the variations we produce via changes in template sets and magnitude priors, we compare the size of the resulting variations in the $dN/dz_{\text{ph}}$ per collection to those of the model analyzed in Ma et al. (2006). The $dN/dz_{\text{ph}}$ per collection is given by

$$
\frac{dN}{dz} (z) = \int_{z_{\text{ph}}^{-}}^{z_{\text{ph}}^{+}} \frac{dN}{dz_{\text{tot}}} (z) \ p(z_{\text{ph}}|z),
$$

(25)

where

$$
\frac{dN}{dz_{\text{tot}}} \ z^{\alpha} \exp\left[-(z/z_0)^{\beta}\right].
$$

(26)

The photo-z distribution is modeled as

$$
p(z_{\text{ph}}|z) = \frac{1}{\sqrt{2\pi} \sigma_z} \exp\left[-\frac{(z - z_{\text{ph}} - z_{\text{bias}})^2}{2\sigma_z^2}\right]
$$

(27)

and is defined for “micro” bins of width $\delta z = 0.1$, each with two photo-z parameters, the bias, $z_{\text{bias}}(z)$, and the scatter, $\sigma(z)$. The $dN/dz_{\text{ph}}$ for each “macro” bin, or collection, is obtained by interpolating to the mean and bias for the tomographic redshift bins (collection in the terminology of this paper) rather than the photo-z distribution is modeled as $p(z_{\text{ph}}|z)$, to return a continuous function defined for the entire redshift range.

To calculate the fluctuations in each collection resulting from 1σ errors on $z_{\text{bias}}(z)$ and $\sigma(z)$ for each “micro” bin, we start with the photo-z model described above. We assume a fiducial case with 40 “micro” bins between $z = 0$ and $z = 4$ with $\delta z = 0.1$, $z_{\text{bias}}(z) = 0$ and $\sigma(z) = 0.05(1 + z)$. We then randomly draw 21 samples (consistent with the fact that we generated 21 samples from variations of template set (we could have used 103 to correspond to variation of the prior parameters, but given that the distribution is Gaussian, the extra variations add little information) from an 80-dimensional Gaussian distribution with $\sigma_{\text{bias}} = \sigma_{\text{scatter}} = 0.01$. We linearly interpolate to create a continuous function in $z$, $P(z_{\text{ph}}|z)$, and we perform the integral in Equation (25) each time to obtain a set of 21 $dN/dz_{\text{ph}}$ for each collection.

We define, as in Section 3, $d_{\alpha} = dN/dz_{\alpha} - \langle dN/dz \rangle$, i.e., the mean-subtracted data. The $dN/dz_{\alpha}$ and the $d_{\alpha}$ are shown in Figure 11. It can be seen that $d_{\alpha}$ are of the same order (although somewhat larger) than those of the template variations (Figure 4). However, there are no catastrophic “islands,” i.e., non-zero values, for the $dN/dz$ far outside of the redshift range for which the collection is defined. This is also reflected in Figure 12 which shows the standard deviation in the mean redshift (Equation (19)) and the variance (Equation (20)) for each collection. What is perhaps most interesting to note is the difference between the Gaussian case and the model-based variations in collections 4 and 5, and to a lesser extent collection...
1, which are subject to the redshift degeneracies associated with faint galaxies, as discussed in Section 3. In particular, these differences are apparent in the standard deviation in the parameterized uncertainties of similar order to those considered in somewhat simplified examples and not intended to be taken as a statement about what this would mean for constraints on the dark energy. We consider two sources of modeling uncertainty—the SED template set and magnitude priors—and show that both cause errors in the photometrically estimated $dN/dz$. Though both cases utilize simplified models to probe their impact, the results indicate that template selection effects are dominant to those of the magnitude priors (mean-subtracted data, $dα = dN/dz_α - (dN/dz)$, are nearly an order of magnitude larger for the template variations). This result comes with the caveat that the underlying parameterization of the prior distribution was perfectly known for the template variations and we only consider the effects of uncertainties in prior parameters (differences could be larger in the case where the functional form of the distribution is incorrect or training set is biased in some way). Additionally, we vary the template set and the priors independently, although the effects of these variations may be correlated. A more sophisticated model would enable us to examine such modeling uncertainties jointly.

We have shown that in both cases, the variations in the $dN/dz$ can be characterized by a PCA, which selects out the directions of maximum variance. Using the principal components as the new basis, the $dN/dz$ can be parameterized by the weights, $b_i$, where the $dN/dz_i$ for each collection is constructed according to Equation (9). The uncertainties in the model can then be described in terms of a covariance matrix for the $b_i$ and can be used in a likelihood analysis for weak lensing or other cosmological observations involving photometric redshifts.

PCA allows for dimensionality reduction (here we find that two or three modes are sufficient to account for 80% of the variation in the data). We note that the precision of the reconstruction can be easily “tuned” by retention of a larger or smaller number of modes. The exact number of modes needed for a given level of fidelity to the original data will depend on the exact data set and modeling uncertainties considered. However, it is a simple matter to choose a desired level of variation captured by the parameterization: $λ_α/\sum λ_α$, where $λ_α$ is the eigenvalue associated with a given mode. For the case of five redshift bins (or collections) and three principal components, the uncertainty in the photometric redshifts can be captured by 15 numbers (a computationally tractable result). Moreover, this parameterization is completely general and does not resort to ad hoc forms such as Gaussians. Thus, we present a way of parameterizing realistic modeling uncertainties and propagating these uncertainties through to the cosmological parameters.

More generally, this method of estimating photo-$z$ uncertainty and propagating to cosmological parameters could be applied to the more realistic observational case where two or more probes of geometry and structure formation (such as cosmic shear tomography and baryon acoustic oscillations) are combined. This cross-calibration removes degeneracies, significantly reducing sensitivity to systematic error. Finally, there are statistical calibrations that can be done in regions of the sky where there are spectroscopic samples, and this information can be imported to the method we have developed. In such cases the mean redshift of the galaxy population in every collection may be calibrated by cross-correlating that sample with a bright spectroscopic sample in angle and redshift (Matthews & Newman 2010). Leveraging the new deep-wide spectroscopic surveys, one could go beyond a calibration of the mean redshift of that sample to knowledge of the distribution $P_α(z)$ of galaxies in that collection as a function of type. This information could then be incorporated into the prior.

6. DISCUSSION AND CONCLUSIONS

With the potential to significantly improve our knowledge of the expansion history of the universe, future observational programs are necessary for making theoretical progress on the puzzle of dark energy. However, systematic errors have the potential to wreak havoc on these results. In particular, photometric redshift methods, essential to many future experimental programs, rely on modeling assumptions that can cause both biases and catastrophic errors in the redshift estimates. In order for future observations to deliver on their promise, these errors must be reduced or well understood.

In this work, we have presented a method of exploring and parameterizing modeling uncertainties associated with photometric redshift estimates. We consider two sources of modeling uncertainty—the SED template set and magnitude priors—and show that both cause errors in the photometrically estimated $dN/dz$. Though both cases utilize simplified models to probe their impact, the results indicate that template selection effects are dominant to those of the magnitude priors (mean-subtracted data, $dα = dN/dz_α - (dN/dz)$, are nearly an order of magnitude larger for the template variations). This result comes with the caveat that the underlying parameterization of the prior distribution was perfectly known for the template variations and we only consider the effects of uncertainties in prior parameters (differences could be larger in the case where the functional form of the distribution is incorrect or training set is biased in some way). Additionally, we vary the template set and the priors independently, although the effects of these variations may be correlated. A more sophisticated model would enable us to examine such modeling uncertainties jointly.

We have shown that in both cases, the variations in the $dN/dz$ can be characterized by a PCA, which selects out the directions of maximum variance. Using the principal components as the new basis, the $dN/dz$ can be parameterized by the weights, $b_i$, where the $dN/dz_i$ for each collection is constructed according to Equation (9). The uncertainties in the model can then be described in terms of a covariance matrix for the $b_i$ and can be used in a likelihood analysis for weak lensing or other cosmological observations involving photometric redshifts.

PCA allows for dimensionality reduction (here we find that two or three modes are sufficient to account for 80% of the variation in the data). We note that the precision of the reconstruction can be easily “tuned” by retention of a larger or smaller number of modes. The exact number of modes needed for a given level of fidelity to the original data will depend on the exact data set and modeling uncertainties considered. However, it is a simple matter to choose a desired level of variation captured by the parameterization: $λ_α/\sum λ_α$, where $λ_α$ is the eigenvalue associated with a given mode. For the case of five redshift bins (or collections) and three principal components, the uncertainty in the photometric redshifts can be captured by 15 numbers (a computationally tractable result). Moreover, this parameterization is completely general and does not resort to ad hoc forms such as Gaussians. Thus, we present a way of parameterizing realistic modeling uncertainties and propagating these uncertainties through to the cosmological parameters.

More generally, this method of estimating photo-$z$ uncertainty and propagating to cosmological parameters could be applied to the more realistic observational case where two or more probes of geometry and structure formation (such as cosmic shear tomography and baryon acoustic oscillations) are combined. This cross-calibration removes degeneracies, significantly reducing sensitivity to systematic error. Finally, there are statistical calibrations that can be done in regions of the sky where there are spectroscopic samples, and this information can be imported to the method we have developed. In such cases the mean redshift of the galaxy population in every collection may be calibrated by cross-correlating that sample with a bright spectroscopic sample in angle and redshift (Matthews & Newman 2010). Leveraging the new deep-wide spectroscopic surveys, one could go beyond a calibration of the mean redshift of that sample to knowledge of the distribution $P_α(z)$ of galaxies in that collection as a function of type. This information could then be incorporated into the prior.
We thank Andreas Albrecht, Jim Bosch, and Ami Choi for useful discussions, and Perry Gee for computing support. This work was supported by the National Science Foundation grant 0709498. J.A.T. and S.S. were supported by the Research Corporation. H.Z. is supported by the Bairen program from the Chinese Academy of Sciences, the National Basic Research Program of China grant No. 2010CB833000, and the National Science Foundation of China grant No. 11033005.

APPENDIX

ESTIMATOR BIAS

Here, we show that the simple estimator Equation (4) is unbiased if the prior and likelihood are both correct. Assuming an underlying redshift distribution $dN/dz$, we get a posterior distribution for the $i$th galaxy

$$P_i(z) \equiv P(f_i | z) = P(f_i | z) dN A(f_i | z),$$

where $A(f_i | z)$ is a normalization factor so that

$$\int dP(z) = 1.$$  \hspace{1cm} (A1)

For simplicity we have suppressed $t$ and $m_i$, which can be integrated out via Equation (3). In what follows, we also drop the subscript $\lambda$ and superscript $d$ of $f$ for convenience. We group galaxies that meet certain criteria, e.g., within a volume in color space, together and call the collection a “redshift bin.” Redshift is only one application; the collection criteria can be more general. In terms of symbolics, we use $f \in F$ to denote the criteria. On the one hand, $dN/dz$ in this $F$ collection is

$$\frac{dN_F}{dz} = \int_{f \in F} df \frac{dN}{dz} P(f | z).$$

(A3)

On the other hand, the estimator is

$$\frac{d \tilde{N}_F}{dz} = \sum_{f \in F} P(z | f).$$

(A4)

Calculating the expectation value for this estimator is the same as calculating its value in the limit of very large numbers of galaxies, assuming that the galaxies follow the redshift distribution $dN/dz$. In this large-$N$ limit we can replace

$$\sum_{f \in F} \int_{f \in F} df \int dz \frac{dN}{dz} P(f | z),$$

so that

$$\left\langle \frac{d \tilde{N}_F}{dz} \right\rangle = \int_{f \in F} df \int dz \frac{dN}{dz} A(f) P(f | z) P(f | z) \frac{dN}{dz}$$

$$= \int_{f \in F} df \frac{dN}{dz} P(f | z) = \frac{dN_F}{dz}.$$  \hspace{1cm} (A5)

Since the above derivation does not use any specific form of the likelihood function, and since the collection $F$ is generic, the estimator is not biased. In reality, the prior $dN/dz$ and the likelihood $P(f | z)$ are not accurately known, and systematic errors in them can still cause biases in the results.

REFERENCES

Abdalla, F. B., Amara, A., Capak, P., Cypriano, E. S., Lahav, O., & Rhodes, J. 2008, MNRAS, 387, 969

Albrecht, A., et al. 2006, APS Meeting Abstracts, G1002 (http://wFIRST.gsfc.nasa.gov/science/DETF_Report.pdf)

Amara, A., & Réfrégier, A. 2008, MNRAS, 391, 228

Assef, R. J., et al. 2008, ApJ, 676, 286

Assef, R. J., et al. 2010, ApJ, 713, 970

Bassett, B. A., & Hlozek, R. 2010, in Dark Energy: Observational and Theoretical Approaches, ed. P. Ruiz-Lapuente (Cambridge: Cambridge Univ. Press), 246

Baum, W. A. 1962, in IAU Symp. 15, Problems of Extra-Galactic Research, ed. G. C. McVittie (Cambridge: Cambridge Univ. Press), 390

Benitez, N. 2000, ApJ, 536, 571

Bernstein, G., & Huterer, D. 2010, MNRAS, 401, 1399

Bordoloi, R., Lilly, S. J., & Amara, A. 2010, MNRAS, 406, 881

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503

Bridge, S., & King, L. 2007, New J. Phys., 9, 444

Badavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., & Dickinson, M. 2000, AJ, 120, 1588

Chapin, E. L., Hughes, D. H., & Aretxaga, I. 2004, in Multiwavelength Cosmology, ed. M. Plionis (Astrophysics and Space Science Library, Vol. 301; Dordrecht: Kluwer), 121

Coleman, G. D., Wu, C., & Weedman, D. W. 1980, ApJ, 43, 393

Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., & Munn, J. A. 1995, AJ, 110, 2655

Cunha, C. E., Lima, M., Oyaizu, H., Frieman, J., & Lin, H. 2009, MNRAS, 396, 2379

Edmondson, E. M., Miller, L., & Wolf, C. 2006, MNRAS, 371, 1693

Fernández-Soto, A., Lanzetta, K. M., Chen, H., Levine, B., & Yahata, N. 2002, MNRAS, 330, 889

Giavalisco, M., et al. 2004, ApJ, 600, L93

Hearin, A. P., Zentner, A. R., Ma, Z., & Huterer, D. 2010, ApJ, 720, 1351

Heavens, A. 2009, in Data Analysis in Cosmology, ed. V. J. Martínez, E. Saar, E. M. Gonzales, & M. J. Pons-Borderia (Lecture Notes in Physics, Vol. 665, Berlin: Springer), 585

Hildebrandt, H., et al. 2010, A&A, 523, 31

Howell, D. A., et al. 2009, arXiv:0903.1086

Hu, W. 1999, ApJ, 522, L21

Huterer, D. 2010, Gen. Rel. Grav., 42, 2177

Huterer, D., & Takada, M. 2005, ApJ, 617, 23

Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, MNRAS, 366, 101

Ilbert, O., et al. 2009, ApJ, 690, 1236

Kinney, A. L., Calzetti, D., Bohlin, R. C., McQuade, K., Storchi-Bergmann, T., & Schmitt, H. R. 1996, ApJ, 467, 38

LSST Science Collaborations and LSST Project, et al. 2009, arXiv:0912.0201

Ma, Z., & Bernstein, G. 2008, ApJ, 682, 39

Ma, Z., Hu, W., & Huterer, D. 2006, ApJ, 656, 21

MacDonald, C. J., & Bernstein, G. 2010, PASP, 122, 485

MacQueen, J. B. 1967, in Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, ed. L. M. Cam & J. Neyman (Berkeley, CA: Univ. California Press), 281

Mandelbaum, R., et al. 2008, MNRAS, 386, 781

Matthews, D. J., & Newman, J. A. 2010, ApJ, 721, 456

Nishizawa, A. J., Takada, M., Hamana, T., & Furusawa, H. 2010, ApJ, 718, 1252

Perlmuter, S., et al. 1999, ApJ, 517, 565

Riess, A. G., et al. 1998, AJ, 116, 1009

Schneider, M., Knox, L., Zhan, H., & Connolly, A. 2006, ApJ, 651, 14

Sun, L., Fan, Z., Tao, C., Kneib, J., Jouvel, S., & Tilquin, A. 2009, ApJ, 699, 958

Takada, M., & Jain, B. 2004, MNRAS, 348, 897

Wang, S., Haiman, Z., & May, M. 2009, ApJ, 691, 547

Wiesenfeld, K., & Moss, F. 1995, Nature, 373, 33

Wittman, D. 2009, ApJ, 700, L174

Zhan, H. 2006, J. Cosmol. Astropart. Phys., JCAP08(2006)008

Zhan, H., et al. 2009, in Astro2010: The Astronomy and Astrophysics Decadal Survey, Science White Papers, No. 132