Dictionary Learning for Photometric Redshift Estimation

Joana Frontera-Pons *, Florent Sureau*, Bruno Moraes†, Jérôme Bobin*, Filipe B. Abdalla† ‡, Jean-Luc Starck*
* IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France. Email: joana.frontera-pons@cea.fr
† Department of Physics & Astronomy, University College London, Gower Street, London WC1E 6BT, UK
‡ Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa

***Abstract*—Photometric redshift estimation and the assessment of the distance to an astronomic object plays a key role in modern cosmology. We present in this article a new method for photometric redshift estimation that relies on sparse linear representations. The proposed algorithm is based on a sparse decomposition for rest-frame spectra in a learned dictionary. Additionally, it provides both an estimate for the redshift together with the full resolution spectra from the observed photometry for a given galaxy. This technique has been evaluated on realistic simulated photometric measurements.**

I. Introduction

Measuring the angular positions of galaxies to the required cosmological precision is easily achievable with an optical galaxy survey; measuring their radial positions, on the other hand, is one of the most challenging problems in modern observational cosmology. The way we infer those radial distances is based on their spectral energy distribution (SED): due to the expansion of the Cosmos, galaxies are receding from us and their light is consequently redshifted, similar to a Doppler Effect. These redshifts are directly related to the galaxies' distances, and by measuring it from the spectral characteristics of the received light, we can reconstruct their positions.

Here two different approaches need to be distinguished, with their own characteristics, advantages and challenges. Measuring spectroscopic redshifts consists in observing the full SED of a galaxy and identifying features that allow a secure redshift determination. Galaxy spectra are a consequence of a series of relatively well-understood physical phenomena, mostly concerning the nuclear and chemical reactions inside stars and the types and ages of stellar populations within the galaxy in question (see [1] for a review). Atomic emission and absorption lines give rise to very distinct peaks and troughs in a galaxy SED, and the secure identification of the wavelength of such a feature can easily be translated into a shift compared to the known wavelength of such a transition observed in Earth's laboratories.

Photometric redshift measurements, on the other hand, try to reconstruct the redshift value out of only a handful of numbers representing the integrated flux in broadband filters . This is an ill-posed severely underdetermined inverse problem where both redshift and spectra needs be estimated from a few photometric measurements. Degeneracies abound, making results less precise and possibly biased, but they circumvent the need of a spectrograph and can also reach fainter magnitudes, as light is integrated in broad wavelength ranges. While spectroscopic redshifts are more accurate than photometric redshifts, their acquisition is time consuming and limited to only the brightest objects.

Most of the techniques for photometric redshift estimation are based either on empirical machine learning approaches or obtained through template-fitting methods [2]. Some of the most popular codes take advantage of neural networks [3], [4], regression trees [5] among others. Other information than flux such as galaxy morphology, colors, etc can also be included in their redshift estimation to improve their accuracy. However, the major drawback of these methods is that they have to be trained with of a huge amount of representative labelled data for which the true redshift value needs to be perfectly known. Another family of methods is based on template fitting. They are based on matching physically meaningful redshifted rest-frame templates (i.e. without redshift effects) to the observed spectrum, to obtain both redshift and best fit template. These template spectra are constructed from theoretical libraries. The most widespread photometric redshift estimation template fitting code is is LePHARE [6]. These techniques strongly rely on a good template modelling and a deep understanding of realistic galaxy SEDs.

The main contributions of this article are:

- A new algorithm for photometric redshift estimation based on rest-frames templates learned from data using sparse dictionary learning; the complete spectrum of the galaxies is also recovered;
- The evaluation of the proposed scheme on realistic galaxy photometric simulations.

II. Methodology

Let us first consider the problem of recovering the full spectra of a galaxy, $\mathbf{x} \in \mathbb{R}^{w_s}$, from photometric measurements, $\mathbf{y} \in \mathbb{R}^{w_p}$, and the vectors' dimensions satisfy $w_p << w_s$. Typical values for the spectra dimension are $w_s = 3000$ or $w_s = 4000$ while only $w_p = 5$ or $w_p = 10$ bands are commonly available. This can be formulated as an inverse problem according to:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{x}$ is the original spectroscopy. This signal has been passed through some filters $\mathbf{H} \in \mathbb{R}^{w_p \times w_s}$, yielding a lower resolution version $\mathbf{y}$, that corresponds to the photometry and

$\mathbf{n}$ represents the noise. Hence, we seek to retrieve the original signal $\mathbf{x}$ by solving this super-resolution task. This severly underdetermined ill-posed inverse problem requires constraints on the spectra $\mathbf{x}$ to be solved. We propose to model the spectra as a sparse linear combination of a few learned templates then redshifted to a tested redshift value ; the best approximation of the photometric data giving the estimated redshift. In the following, we first present how we build our learned rest-frame representation for galaxy spectra using sparse dictionary learning, the sparse coding algorithm associated to the recovery of the spectra, and finally how we estimate the redshift.

A. Dictionary learning for rest-frame galaxy spectra

The proposed method relies on learning linear representations on rest-frame training data and the spectra are approximated by a sparse decomposition, $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$. In this context, the dictionary $\hat{\mathbf{D}} \in \mathbb{R}^{w_s \times n_a}$ with n_a atoms is constructed from a training set $\mathbf{X} \in \mathbb{R}^{w_s \times n_t}$. This training set is composed of n_t examples disposed in columns and the dictionary is obtained by solving the joint minimization problem:

$$\hat{\mathbf{D}}, \hat{\mathbf{A}} = \underset{\mathbf{D}\in\mathcal{D},\mathbf{A}}{\arg\min} ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 \text{ s.t.} \forall i,\ ||\boldsymbol{\alpha}_i||_0 \leq \tau \quad (2)$$

where $\hat{\mathbf{A}} \in \mathbb{R}^{n_a \times n_t}$ is the matrix of codes and each column corresponds to the representation for each training example, $\{\boldsymbol{\alpha}_i\}$. $||\cdot||_F$ denotes the Frobenius norm, $||\cdot||_0$ counts the number of non-zero entries of a vector and τ is the targeted sparsity degree, $\mathcal{D}$ designates the set of dictionaries with atoms in the unit ℓ_2 ball. Among the different approaches to solve (2), we use a technique based on the method of optimal direction detailed in [7]. This procedure performs alternately sparse coding by orthogonal matching pursuit and dictionary updating. The sparsity degree specified in the sparse coding stage and the number of atoms in the dictionary are free parameters.

B. Sparse coding for rest-frame galaxy spectra

The original spectroscopic signal $\mathbf{x}$ is then retrieved from the photometric signal $\mathbf{y}$ by imposing sparsity on the learned representation $\boldsymbol{\alpha}$. In addition to the sparsity constraint, positivity on the reconstructed spectra can also be enforced for a more constrained recovery. Although negative values of the spectra may lead to a better photometry reconstruction, these solutions are impossible. Therefore, we need to minimize:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\arg\min} \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{D}\boldsymbol{\alpha}||_2^2 + \lambda\,||\boldsymbol{\alpha}||_1 + I_{\mathcal{C}}(\mathbf{D}\boldsymbol{\alpha}) \quad (3)$$

where $I_{\mathcal{C}}$ denotes the indicator function on the spectra set $\mathcal{C}$ that enforces non-negativity for the galaxy emitted light. The regularization parameter λ controls the trade off between the reconstruction error and the sparsity promoting term. The value of λ has been automatically set to be proportional to the estimated noise level $\hat{\sigma}$ as detailed in [8].
To take into account the different constraints and the differential term in the cost function, the optimisation in (3) is performed with the Generalized Forward-Backward Splitting algorithm introduced in [9] and recalled in algorithm 1. The prox operator associated to the ℓ_1 norm corresponds to soft-thresholding operator; the one associated to the indicator function has no closed-form expression but was computed with an inner FISTA algorithm on the dual problem, as detailed in [10].

Algorithm 1 *Generalized Forward-Backward Splitting*

Initialization : $k = 0$, $\mathbf{t}_1 = 0$, $\mathbf{t}_2 = 0$, $\hat{\boldsymbol{\alpha}} = 0$ and $\lambda = 3\hat{\sigma}^2$,
while Have not converged **do**
$\nabla = -\frac{1}{L}\,\mathbf{D}\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{D}^T\hat{\boldsymbol{\alpha}}_{k-1})$
$\mathbf{t}_1 = \mathbf{t}_1 + \text{prox}_{\frac{\lambda}{L}||\cdot||_1}(2 * \hat{\boldsymbol{\alpha}}_{k-1} - \mathbf{t}_1 - \nabla)$
$\mathbf{t}_2 = \mathbf{t}_2 + \text{prox}_{I_{\mathcal{C}}(\cdot)}(2 * \hat{\boldsymbol{\alpha}}_{k-1} - \mathbf{t}_2 - \nabla)$
$\hat{\boldsymbol{\alpha}}_k = \frac{\mathbf{t}_1 + \mathbf{t}_2}{2}$
end while
return $\hat{\boldsymbol{\alpha}}$

C. Photometric redshift algorithm

Similarly, we can decompose an observed spectrum $\mathbf{x}_z$, at a certain redshift z, according to $\mathbf{x}_z = \mathbf{D}^{(z)}\boldsymbol{\alpha}^{(z)}$. The value of z is computed as the one providing the closest approximation for the observed photometric signal y_z.
More specifically, for every tested value of z, the dictionary $\mathbf{D}$ originally built for rest-frame representations is redshifted to $\mathbf{D}^{(z)}$ and we solve an inverse problem as the one described in (3). Accordingly, we can write for every value of z:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}^{(z)} = &\underset{\boldsymbol{\alpha}}{\arg\min} \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{D}^{(z)}\boldsymbol{\alpha}^{(z)}||_2^2 \\ &+ \lambda\,||\boldsymbol{\alpha}^{(z)}||_1 + I_{\mathcal{C}}(\mathbf{D}^{(z)}\boldsymbol{\alpha}^{(z)}) \end{aligned} \quad (4)$$

and solve (4) with algorithm 1 described above. Ultimately, the value of the redshift z is obtained as the solution of the following equation:

$$\hat{z} = \underset{z}{\arg\min}\ \frac{||\mathbf{y} - \mathbf{H}\mathbf{D}^{(z)}\boldsymbol{\alpha}^{(z)}||_2^2}{||\mathbf{y}||_2^2} \quad (5)$$

Solving problem (5) requires a fine sampling on the range of tested redshifts, which would require solving many problems (4) and would be computationnaly extremely costly. To avoid this, we propose a coarse-to-fine strategy for redshift testing: we evaluate the approximation error for a hierarchical grid of z values. In other words, the whole interval that encompasses all possible values of $z \in [z_{min}, z_{max}]$ has been uniformly sampled with ten steps, and the minimum among this points, $\hat{z}_1$, is retained. Then, the explored interval is reduced around this minimum. The new interval is evenly re-sampled at ten points yielding a new minima. This process is repeated five times allowing us to build a hierarchical grid for z. This method reduces the computational time while keeping a good resolution in terms of z and will be illustrated in the following experimental section.

D. Comparison with LePHARE

In order to assess the performance of the proposed redshift algorithm, the proposed algorithm is compared to LePHARE code [6]. LePHARE is a template-based redshift estimation method. It starts from a library of spectroscopic templates built from a wide range of theoretical observations. It then applies observational corrections to the spectra and integrates them through the defined filter set. For each galaxy, LePHARE integrates all spectra in the library for several redshift test values and finds the combination of a spectrum and a redshift value that provide the best possible fit to the observed photometric data. In this way, each galaxy is assigned a best-fit template and a redshift value.

III. Experimental results

We present in this section the results obtained with galaxy simulated spectroscopy for the training stage and simulated photometry for testing the algorithms.

A. Simulations

In this section we present the data used in our studies. The first step is to define a master catalog for the analyses. We work with the COSMOSSNAP simulation pipeline [11] to generate a data set of simulated galaxy SEDs and corresponding photometric properties. The idea is to take real data as a basis, thereby ensuring that realistic relationships between galaxy type, color, size, redshift and SED are preserved. COSMOSSNAP chooses the COSMOS photometric redshift catalog [12], generated from a combination of 30 bands from diverse astronomical surveys covering the full spectral range from the UV (GALEX), through the optical (Subaru) and all the way to infrared bands (CFHT, UKIRT, Spitzer). This data set is matched to Hubble ACS imaging data, to provide realistic size-magnitude distributions, employing weak-lensing-quality shape measurements [13]. Based on these properties, COSMOSSNAP chooses a spectral template from a predefined library such that the integrated fluxes through the 30 broadband filters above provide the best-fit to the observations. Each galaxy therefore has a "true" redshift and its associated SED, and the distribution of types and redshifts follows the measured distribution in the COSMOS field. This catalog is the basis for all COSMOSSNAP simulations.

To generate realistic photometric properties, the first step is to integrate the best-fit spectral template through a set of broadband wavelength filters that will be used for a given galaxy survey. In actuality, the full transmission curve includes not only filter effects, but also atmospheric transmission (in the case of ground observations), telescope optical effects and more. The full transmission curve is commonly referred to as filter throughput (even though it is not only due to the filter itself). COSMOSSNAP takes a defined set of filter throughput and calculates magnitudes and their corresponding errors for each galaxy in the catalogue. For the purposes of our analysis, we choose to reproduce closely the expected properties of the Large Synoptic Survey Telescope [14] (LSST). Fig. 1 shows the modelled throughputs [15] for our current band selection represented by $\mathbf{H}$ in the problem formulation. Therefore, the redshift value will need to be inferred only from these 6 available broadbands (commonly referred to as 'ugrizY'). At the end of the generation procedure, we have a realistic master galaxy catalogue with magnitudes, colors, shapes and redshifts for 538 000 galaxies on an effective 1.24 deg^2 region of the sky down to an i-band magnitude of 26.5. To further match the expected properties of the LSST Science sample, we limit our catalog to galaxies brighter than 25.3 and with signal-to-noise $(S/N) > 10$ in the i-band. Imposing these restrictions, we obtain a galaxy catalog with a realistic set of photometric properties, and best-fit spectral templates with realistic continuum and emission line properties. We now need to forward-model the observational process in the spectroscopic case in a manner consistent with expected observational conditions.

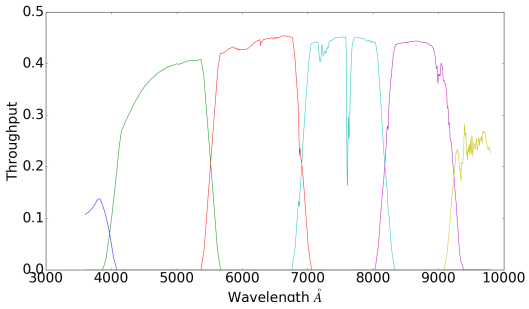

Fig. 1: LSST filter throughputs for the considered photometric scenario.

For obtaining realistic spectral templates, we need to resample and integrate the best-fit SEDs. As given by the simulations, these SEDs are pure functional forms. At the end of the observational process, what we obtain is an integrated flux in logarithmic wavelength bins at a resolution of R. From the simulation run described above, we select two random subsets.

B. Dictionary Learning

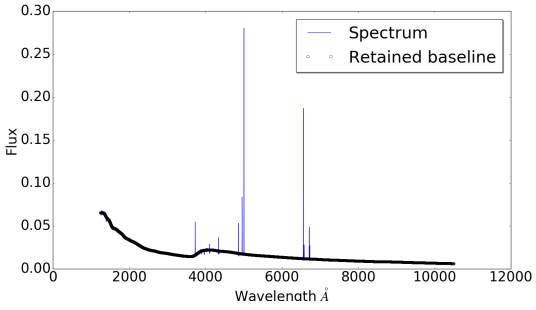

Fig. 2: Example of the subtraction of high-frequency features for rest-frame spectra. The original spectra is represented by a blue solid line and the retained information after emission lines subtraction is displayed with black circles.

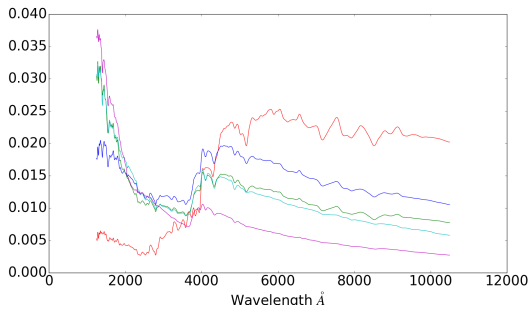

Fig. 3: Example of five atoms learned using dictionary learning and imposing a sparsity degree of 3 on rest-frame spectra.

Firstly, we chose a subset of noiseless low-redshift galaxies that have been blueshifted to $z = 0$ in order to form the training set. Hence, the $\mathbf{X}$ is composed of $n_t = 10000$ clean rest-frame example spectra covering the range $[1250\text{Å}, 10499\text{Å}]$ and $w_s = 4258$. Moreover, high frequency information from these rest-frame spectra has been removed through wavelet filtering retaining four scales and keeping the baseline as illustrated in Fig. 2. Finally, the dictionary $\mathbf{D}$ is learned by specifying the desired sparsity degree $\tau = 3$ and the number of atoms of the dictionary $n_a = 40$. The code developed in C++ was iterated for 100 repetitions which allowed for convergence in the dictionary estimation measured as the averaged approximation error variations through iterations. Fig. 3 displays five atoms from the adapted dictionary used from now on.

C. Redshift estimation

Secondly, the testing is performed on a different randomly selected subset. We have evaluated the algorithm on $n = 1000$ galaxies lying in a redshift range of $z \in [0, 1]$ and including only $w_p = 6$ photometric measures for each galaxy.

Let us now discuss the results obtained for redshift estimation in the simulated catalogue.
The considered strategy of building a hierarchical grid mesh for testing the different z values is illustrated in Fig. 4. The grid search starts by exploring the whole $z \in [0, 1]$ interval and the approximation error as a function of the tested redshift is depicted in Fig. 4 (a). Hence, the minimum is chosen and the considered interval is reduced in Fig. 4 (b). We repeat the process five times to achieve the desired resolution in z. The smoothness of the approximation curves as a function of redshift allows to attain the same minima with this hierarchical approach as the one obtained with a one level grid with a much finer resolution as shown in Fig. 5, although the computational time is significantly lower, which justifies the choice of our approach.

Fig. 6 displays the estimated redshift for all the galaxies in the test set with respect to their true redshift value. The performance of the method is quantified through the bias over the entire test set $\langle \delta_z \rangle = \langle z_{est} - z_{true} \rangle = -0.004$, and the 68th percentile scatter $\sigma_{68} = 0.0475$. Then, one can define the number of catastrophic failures as those galaxies falling outside $3\sigma_{68}$, yielding $\nu = 53$.

Finally, Fig. 7 shows the results of the simulated catalogue with LePHARE photometric estimation. The corresponding bias is $\langle \delta_z \rangle = 0.0421$, the 68th percentile scatter $\sigma_{68} = 0.0708$ and the number of catastrophic failures $\nu = 22$.

It is important to point out two main differences with our algorithm. On one hand, the templates used in the LePHARE code are theoretical while ours are derived directly from the data. Moreover, while LePHARE is based on template fitting, the proposed method allows for a linear combination of more than one template leading to greater flexibility and representational capacity.

IV. Conclusion

We have introduced a new method to compute redshift from photometric data. The proposed algorithm allows to recover the full-spectra of the galaxies from broad-band photometry solving a super-resolution problem. This estimation scheme has been analyzed on simulated galaxies' spectra and compared to classical LePHARE code.

Further developments will explore other representation approaches where the emission lines are included. The performances will be compared to other photometric redshift estimation based on machine learning as ANNz2 [4]. Finally, we aim to investigate the performance of this algorithm on real photometric data.

Acknowledgment

This work is funded by the DEDALE project (contract no. 665044) and LENA (ERC StG no. 678282) within the H2020 Framework Program of the European Commission.

References

[1] H. Mo, F. Van den Bosch, and S. White, *Galaxy formation and evolution*. Cambridge University Press, 2010.
[2] H. Hildebrandt, S. Arnouts, P. Capak, L. Moustakas, C. Wolf, F. B. Abdalla, R. Assef, M. Banerji, N. Benítez, G. Brammer *et al.*, "Phat: Photo-z accuracy testing," *Astronomy & Astrophysics*, vol. 523, p. A31, 2010.
[3] R. Tagliaferri, G. Longo, S. Andreon, S. Capozziello, C. Donalek, and G. Giordano, "Neural networks for photometric redshifts evaluation," in *Italian Workshop on Neural Nets*. Springer, 2003, pp. 226–234.
[4] I. Sadeh, F. B. Abdalla, and O. Lahav, "Annz2: photometric redshift and probability distribution function estimation using machine learning," *Publications of the Astronomical Society of the Pacific*, vol. 128, no. 968, p. 104502, 2016.
[5] A. Boselli, *A panchromatic view of galaxies*. John Wiley & Sons, 2012.
[6] S. Arnouts and O. Ilbert, "Lephare: Photometric analysis for redshift estimate," *Astrophysics Source Code Library*, 2011.
[7] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5. IEEE, 1999, pp. 2443–2446.
[8] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
[9] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.

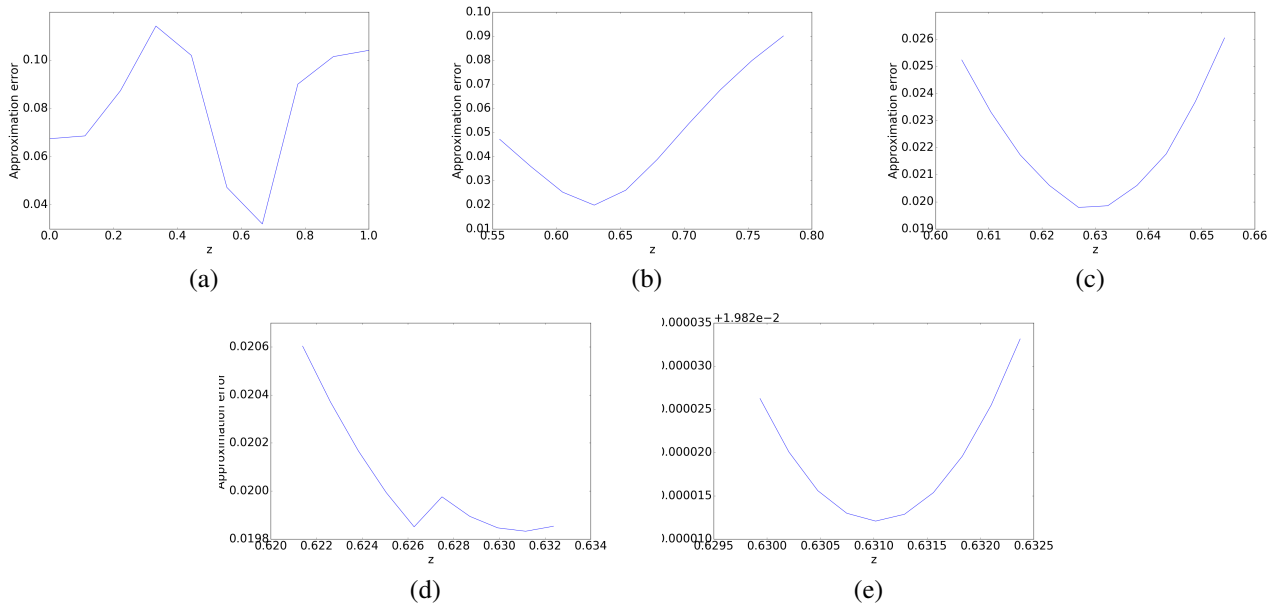

Fig. 4: Different levels in the hierarchical grid mesh for testing the values of z. The whole z range is explored in (a) and the minimum is computed at each layer reducing the considered interval up to the finest resolution in (e).

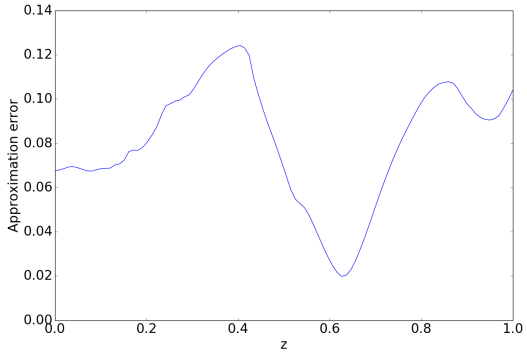

Fig. 5: One-level grid uniformly sampled at 100 steps between $z = 0$ and $z = 1$.

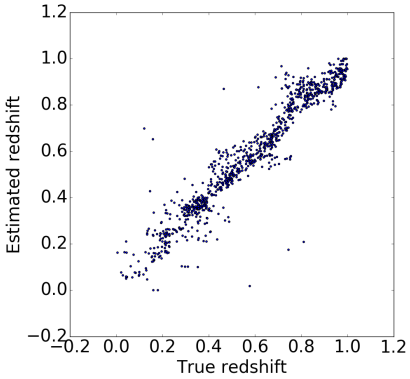

Fig. 6: True vs estimated redshifts for the proposed dictionary learning photometric redshift estimation algorithm.

[10] J. Rapin, J. Bobin, A. Larue, and J.-L. Starck, "NMF with sparse regularizations in transformed domains," *SIAM journal on Imaging Sciences*, vol. 7, no. 4, pp. 2020–2047, 2014.

[11] S. Jouvel, J.-P. Kneib, O. Ilbert, G. Bernstein, S. Arnouts, T. Dahlen, A. Ealet, B. Milliard, H. Aussel, P. Capak *et al.*, "Designing future dark energy space missions-i. building realistic galaxy spectro-photometric catalogs and their first applications," *Astronomy & Astrophysics*, vol. 504, no. 2, pp. 359–371, 2009.

[12] O. Ilbert, P. Capak, M. Salvato, H. Aussel, H. McCracken, D. Sanders, N. Scoville, J. Kartaltepe, S. Arnouts, E. Le Floc'h *et al.*, "Cosmos photometric redshifts with 30-bands for 2-deg2," *The Astrophysical Journal*, vol. 690, no. 2, p. 1236, 2008.

[13] A. Leauthaud, R. Massey, J.-P. Kneib, J. Rhodes, D. E. Johnston *et al.*, "Weak gravitational lensing with cosmos: galaxy selection and shape measurements," *The Astrophysical Journal Supplement Series*, vol. 172, no. 1, p. 219, 2007.

[14] https://www.lsst.org/.

[15] https://github.com/lsst/throughputs.

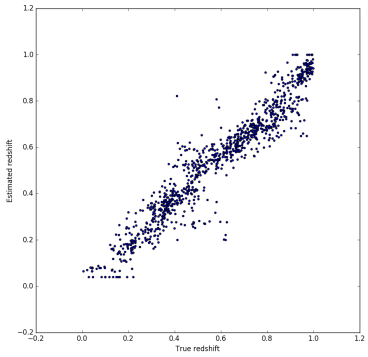

Fig. 7: True vs estimated redshifts from the benchmark LePHARE code.