# A Spectral Model for Multimodal Redshift Estimation

Sven D. Kügler, Nikolaos Gianniotis, Kai L. Polsterer
Astroinformatics Group
Heidelberg Institute for Theoretical Studies gGmbH
Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
Email: {Dennis.Kuegler, Nikos.Gianniotis, Kai.Polsterer}@h-its.org

*Abstract*—**We present a physically inspired model for the problem of redshift estimation. Typically, redshift estimation has been treated as a regression problem that takes as input magnitudes and maps them to a single target redshift. In this work we acknowledge the fact that observed magnitudes may actually admit multiple plausible redshifts, i.e. the distribution of redshifts explaining the observed magnitudes (or colours) is multimodal. Hence, employing one of the standard regression models, as is typically done, is insufficient for this kind of problem, as most models implement either one-to-one or many-to-one mappings. The observed multimodality of solutions is a direct consequence of (a) the variety of physical mechanisms that give rise to the observations, (b) the limited number of measurements available and (c) the presence of noise in photometric measurements. Our proposed solution consists in formulating a model from first principles capable of generating spectra. The generated spectra are integrated over filter curves to produce magnitudes which are then matched to the observed magnitudes. The resulting model naturally expresses a multimodal posterior over possible redshifts, includes measurement uncertainty (e.g. missing values) and is shown to perform favourably on a real dataset.**

## I. Introduction

The exploration of the history of the universe has been mainly driven by the detection and investigation of highly-redshifted extragalactic sources, such as the quasi-stellar objects (QSO, [2]). The study of the distribution of these objects over space and time allows us to draw precise conclusions about how the universe was initially formed and developed since then [3]. Additionally, photometric redshifts have been used in the studies of galaxy clusters [1] and in constraining the galaxy luminosity function [15].

Due to the extreme luminosity of quasars they are perfect traces for the early universe. Therefore, a significant time of research has been spent on estimating their redshifts. While spectroscopic surveys are extremely precise in doing so, they are extremely time-intensive and cannot be used to study a large fraction of the objects known to date. Instead, photometric surveys are used to infer knowledge about the nature and redshift of the quasars. Originally, this was done in a template-based way [7]. In particular, the proposed work bears close resemblance to [9] which addressed the generation of templates from spectra by first "repairing" spectra that have missing values. Also very relevant is the approach in [4] which approaches redshift estimation from a Bayesian angle. Recently the number of data-driven approaches has increased drastically (e.g. [20], [13], [12] and many more). In these works the main focus has been the comparison of methodologies. A popular tool of the community is the random forest [8] due to its reproducibility, precision and favourable computational complexity.

In this work, we acknowledge the fact that the redshift estimation is a regression problem that admits multiple solutions, i.e. there can be more than just one redshift $z$ that explains the observed magnitudes. If we look at the set of possible solutions $z$ as a distribution, then we are acknowledging that this distribution can be multimodal (multiple distinct solutions) as opposed to unimodal (single distinct solution). This manifestation of multiple distinct solutions is the consequence of:

- *Different physical mechanisms*. The aforementioned quasars are actually a preamble for a variety of observed phenomena that are thought to originate from similar sources, simply observed under different circumstances, e.g. viewing angles [18]. The emission of light due to a loss of gravitational energy is common for all the sources. However, there are a lot of different effects that can contribute to the appearance of the spectrum, e.g. the central electron density or the black hole mass and spin e.g. [16]. In addition, it has been observed that the appearance of the quasars changes with redshift [11]. The superposition of only these effects can lead to the observed multimodality and make a physical modelling of quasar emission very cumbersome.
- *Limited number of photometric measurements*. An abundant number of photometric measurements could potentially help us identify a distinct redshift solution. However, with only a limited number measurements available (e.g. 11 measurements do not suffice to pinpoint a unique redshift [19]), ambiguities appear in the guise of multiple distinct solutions.
- *The presence of noise in photometric measurements may introduce multiple solutions*. Typically, the noisier the measurements, the more difficult it becomes to pinpoint the correct redshift. However, the presence of noise does not only introduce uncertainty, but also distinct solutions. As an example, consider the case of two sets of photometric measurements $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ which look very similar in all but one band in which they differ significantly. Assume

further that, due to this significant difference in one band, the two measurements correspond to distinct redshifts $z_1$ and $z_2$. If we were to increasingly add noise to the distinct band, we should observe that beyond a certain level of noise, solution $z_1$ starts becoming a likely candidate solution for measurement set $\boldsymbol{g}_2$; we should also notice that $z_2$ starts becoming a likely solution for $\boldsymbol{g}_1$. In the extreme case, that the only distinct band was completely missing (or equivalently riddled by high uncertainty), the remaining observed bands would point out that there are two distinct solutions $z_1$ and $z_2$ for both $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$.

The above factors have a tremendous impact on the uncertainty and the number of distinct redshift solutions. However, in most works concerning redshift estimation, some of the standard regression models are employed which are typically geared towards optimising an objective that consists in a sum-of-squares error or some variation thereof:

$$E = \iint (f(x) - t)^2 \; p(t|x) \; p(x) \; dt \; dx \; , \tag{1}$$

where $f(\cdot)$ is a regression model, $x$ are the data inputs and $t$ the data targets. By rewriting this loss as the bias-variance decomposition [5], we gain the following insight:

$$\begin{aligned} E = & \iint (f(x) - t + \mathbb{E}[t|x] - \mathbb{E}[t|x])^2 \; p(t|x) \; p(x) \; dt \; dx \\ & + (t - \mathbb{E}[t|x])^2 \; p(t|x) \; p(x) \; dt \; dx \\ = & \iint (f(x) - \mathbb{E}[t|x])^2 + (t - \mathbb{E}[t|x])^2 \; p(x) \; dx \; , \end{aligned} \tag{2}$$

where $\mathbb{E}[t|x]$ is the conditional expectation $\int tp(t|x)dt$ for a given $x$. The first term in Eq. (2) tells us that an optimal regression model $f(\cdot)$ is one that is as close as possible to the conditional average $\mathbb{E}[t|x]$. The second term is the variance in $t$ due the presence of noise. We can thus interpret the fitted model as a conditional Gaussian density $\mathcal{N}(f(x), \sigma^2)$ with the variance given by the second term. Hence, standard regression models $f(\cdot)$ are not appropriate for the redshift estimation problem where multiple redshift solutions are possible.

In the machine learning literature, inverse problems with multiple solutions have been addressed by two prevailing frameworks, namely the mixture density network architecture (MDN) [6] and the hierarchical mixture of experts (HME) [10]. In MDN the target variable is modelled with a mixture of Gaussians. The parameters of the Gaussian mixture are parametrised by the outputs of the neural network. Hence, the parameters of the Gaussian mixture are a function of the data inputs and this results in a flexible model that adapts its distribution to the local characteristics of the input space. HME also offers local adaptation by partitioning the data space and allocating different experts to each partition. This allows for building complex models out of simpler models that specialise on smaller regions of the data space.

As aforementioned, in the case of redshift estimation, the multimodality of solutions is a consequence of the presence of noise and the limited number of measurements and thus not (necessarily) a characteristic of the data space, i.e. *multimodality is not a function of the data space*. In this work, we formulate a model based on simple physical considerations that states in a generative fashion how observed magnitudes arise from spectra. Observational noise is incorporated in a transparent way and the multimodality in the distribution of redshift solutions arises naturally.

## II. Model formulation

### A. Probabilistic PCA for uncertain spectra

For the proposed approach, all spectra are preprocessed and are shifted to their rest-frames (see Section III-B). Therefore, not all wavelengths are observed for all spectra. This prevents us from extracting photometry by integrating the spectra over the filter curves for any redshift. In order to make this integration possible, we propose to "fill in" the unobserved wavelengths which we treat here as missing data. To that purpose, we employ probabilistic principal component analysis (PPCA), with a slight modification that allows us to deal with missing/unobserved data. The idea behind using PPCA is to treat the observed, "incomplete" high-dimensional spectra as noise-corrupted versions of low-dimensional coordinates. In other words, by having at our disposal only "incomplete" spectra, we try to reconstruct a lower-dimensional space that explains the behaviour of the observed data. Once we have identified the lower-dimensional space, we can generate "complete" spectra by mapping a low-dimensional coordinate into the high-dimensional space of spectra.

In the following we give a brief overview of how PPCA is formulated omitting details that can be found in the original formulation in [17]. Following PPCA, we take the view that the observed high-dimensional data $\boldsymbol{\xi}_n \in \mathbb{R}^D$ are the images of $Q$-dimensional $(Q < D)$ coordinates $\boldsymbol{\theta}_n \in \mathbb{R}^Q$ under a linear mapping plus additive Gaussian noise of covariance $\boldsymbol{S}_n$:

$$\boldsymbol{\xi}_n = \boldsymbol{W}\boldsymbol{\theta}_n + \boldsymbol{\mu} + \boldsymbol{S}_n \; , \tag{3}$$

where $\boldsymbol{W} \in \mathbb{R}^{D \times Q}$ and $\boldsymbol{\mu} \in \mathbb{R}^D$ define the linear mapping from the low-dimensional space to the high dimensional data space, and $\boldsymbol{S}_n \in \mathbb{R}^{D \times D}$ is a diagonal covariance matrix whose diagonal elements are equal to the variance in the measurement of $\boldsymbol{\xi}_n$, i.e. $\boldsymbol{S}_n = diag(\sigma^2_{n,1}, \ldots, \sigma^2_{n,D})$. As aforementioned, we do not observe a spectrum $\boldsymbol{\xi}_n$ in its entire wavelength range, i.e. certain values $\xi_{ni}$ are unobserved. Our solution for dealing with unobserved wavelengths is to set the unobserved $\xi_{ni}$ equal to a fixed value $\bar{\xi}_n$ and set the corresponding variance to a high value $\sigma^2_{n,i} = \sigma^2_{high}$ , thus stating our ignorance for these unobserved values at wavelengths $i$.

Still following PPCA, the model is completed by imposing a Gaussian prior on the latent variables $\mathcal{N}(\boldsymbol{\theta}_n|\boldsymbol{0}, \boldsymbol{I})$. This gives rise to the following log-likelihood:

$$\log \mathcal{L}(\boldsymbol{W}, \boldsymbol{\mu}) = \log \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{\xi}_n|\boldsymbol{W}\boldsymbol{\theta}_n + \boldsymbol{\mu}, \boldsymbol{S}_n) \; \mathcal{N}(\boldsymbol{\theta}_n|\boldsymbol{0}, \boldsymbol{I}_D) \; . \tag{4}$$

Treating the low-dimensional coordinates $\boldsymbol{\theta}_n$ as latent variables, PPCA formulates an expectation-maximisation algorithm. In the expectation step, one calculates the expected log-likelihood $\mathbb{E}_{q(\theta)}[\log \mathcal{L}]$ where the expectation is taken over the posterior of the latent variables $q(\boldsymbol{\theta}_n)$ which due to the linear-Gaussian structure of the model is a Gaussian density, $q(\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{\theta}_n|\boldsymbol{m}_n, \boldsymbol{C}_n)$. Hence in the expectation step we need the posterior mean and covariance of each latent variable. These quantities are calculated as:

$$\boldsymbol{C}_n = \left(\boldsymbol{I} + \boldsymbol{W}^T \boldsymbol{S_n}^{-1} \boldsymbol{W}\right)^{-1} \tag{5}$$

$$\boldsymbol{m}_n = \boldsymbol{C}_n \boldsymbol{W}^T \boldsymbol{S}_n^{-1}(\boldsymbol{\xi}_n - \boldsymbol{\mu}) \tag{6}$$

and can be contrasted to the corresponding equations (25) and (26) of the standard PPCA found in [17]. Armed with these posteriors, we are in position to calculate $\mathbb{E}_{q(\theta)}[\log \mathcal{L}(\boldsymbol{W}, \boldsymbol{\mu})]$. In the maximisation step, we optimise[1] $\boldsymbol{W}$ by calculating the gradient $\frac{\partial}{\partial \mathbf{W}}\mathbb{E}_{q(\theta)}[\log \mathcal{L}(\boldsymbol{W}, \boldsymbol{\mu})]$ and employing it in a gradient-based optimiser.

Once the expectation-maximisation algorithm has converged, we are able to map previously unseen (out-of-sample) "incomplete" spectra $\boldsymbol{\xi}_\star$, with covariance matrix $\boldsymbol{S}_\star$, to the low-dimensional space by:

$$\boldsymbol{\theta}_\star = \left(\boldsymbol{I} + \boldsymbol{W}^T \boldsymbol{S}_\star^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^T \boldsymbol{S}_\star^{-1}(\boldsymbol{\xi}_\star - \boldsymbol{\mu}) \ . \tag{7}$$

A reconstruction for $\boldsymbol{\xi}_\star$ can be obtained by mapping back to the data space:

$$\hat{\boldsymbol{\xi}}_\star = \boldsymbol{W}\boldsymbol{\theta}_\star + \boldsymbol{\mu} \ . \tag{8}$$

Figure 1 illustrates how an observed, out-of-sample, spectrum $\boldsymbol{\xi}_\star$ is reconstructed as a "complete" spectrum $\hat{\boldsymbol{\xi}}_\star$.

### B. Physical model

In the previous section, we briefly described how PPCA embeds noise-corrupted spectra in a low-dimensional space and how "complete" reconstructions of spectra can be generated. In fact, we regard the PPCA as a generative model, parametrised by the low-dimensional coordinate $\boldsymbol{\theta} \in \mathbb{R}^Q$, that generates synthetic spectra $\hat{\boldsymbol{\xi}}(\boldsymbol{\theta})$ that closely resemble observed spectra $\boldsymbol{\xi}$. In the following, we detail how this generative model can be exploited in a redshift regression model.

Observed photometric magnitudes are produced by a spectrum $\hat{\boldsymbol{\xi}}(\boldsymbol{\theta})$, generated by our PPCA model, of an unknown $\boldsymbol{z}$ which we want to estimate. Furthermore, photometric magnitudes are modelled as the integration of the spectrum over filter curves. Thus, the flux $\mathcal{I}$ in a band $b$ is computed as:

$$\mathcal{I}_b(\theta, z) = \frac{\int_0^\infty \lambda \hat{\boldsymbol{\xi}}(\boldsymbol{\theta})(\lambda/(z+1)) f_b(\lambda) d\lambda}{\int_0^\infty \lambda f_b(\lambda) d\lambda} .$$

Since the spectra are discrete the transformation $\hat{\boldsymbol{\xi}}(\boldsymbol{\theta})(\lambda/(z+1))$ cannot be continuously done. This is why, we use the replacement

$$\tilde{\lambda} = \lambda/(z+1)$$

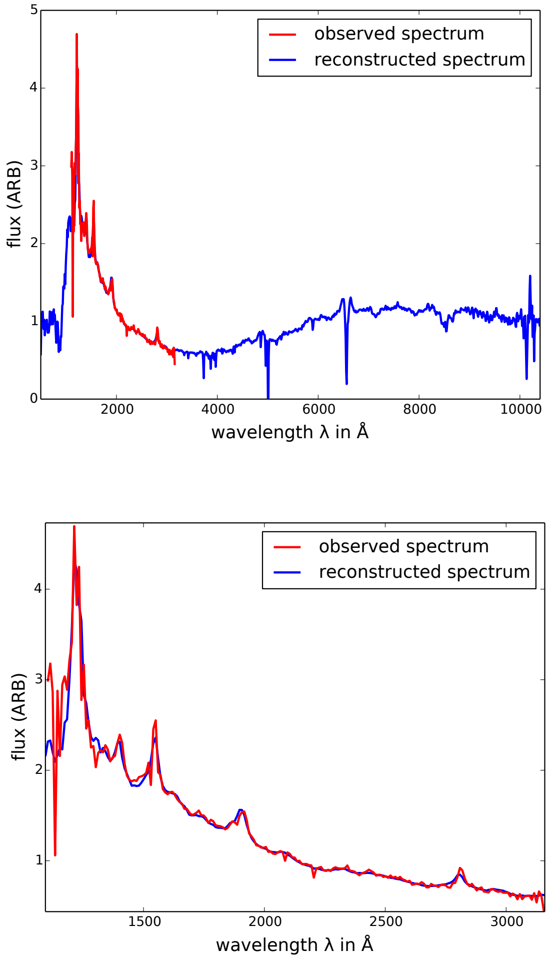

Fig. 1: Above: observed "incomplete" spectrum $\boldsymbol{\xi}$ plotted in red, reconstructed and "complete" spectrum $\hat{\boldsymbol{\xi}}$ obtained by PPCA plotted in blue. Below: same as before, but we zoom in the region where observed and reconstructed spectrum overlap.

$$\frac{d\tilde{\lambda}}{d\lambda} = 1/(z+1) \Rightarrow d\lambda = (z+1)d\tilde{\lambda}$$

and thus

$$\begin{aligned} \mathcal{I}_b(\boldsymbol{\theta}, z) &= \frac{\int_0^\infty (z+1)\tilde{\lambda}\hat{\boldsymbol{\xi}}(\boldsymbol{\theta}) f_b((z+1)\tilde{\lambda})(z+1)d\tilde{\lambda}}{\int_0^\infty (z+1)\tilde{\lambda} f_b((z+1)\tilde{\lambda})(z+1)d\tilde{\lambda}} \\ &\overset{\tilde{\lambda} \to \lambda}{=} \frac{\int_0^\infty \lambda \hat{\boldsymbol{\xi}}(\boldsymbol{\theta}) f_b((z+1)\lambda) d\lambda}{\int_0^\infty \lambda f_b((z+1)\lambda) d\lambda} . \end{aligned} \tag{9}$$

Effectively, we have now pushed the redshift from the discontinuous spectrum to the continuous filter bands by replacing $\tilde{\lambda} \to \lambda$. However, these can be easily approximated with an analytical function, here with a linear regression model with basis functions

$$f(\lambda) = \sum_{c=1}^{N_{comp}} v_c \exp\left(-0.5\left(\frac{\mu_c - \lambda}{\sigma_c}\right)^2\right) , \tag{10}$$

[1] Just like in the original PPCA, $\boldsymbol{\mu}$ is set equal to the sample mean of the data.

with $v_c$, $\mu_c$, $\sigma_c$ being the weights, the means and the widths of each of the $N_{comp}$ RBF basis functions.

In order to compute now the expected flux from our model, we approximate this integral as a regular Riemann sum, where the bin width $\Delta$ is given by the distance between two regularly sampled grid points, as described in the preprocessing (see Section III-B). Finally, the flux in band $b$ is computed as

$$\begin{aligned}\mathcal{I}_b(\theta, z) &\approx \frac{\Delta \sum_d^D \lambda_d \hat{\boldsymbol{\xi}}(\boldsymbol{\theta}) f_b((z+1)\lambda_d)}{\Delta \sum_d^D \lambda_d f_b((z+1)\lambda_d)} \\ &= \frac{\sum_d^D \lambda_d \hat{\boldsymbol{\xi}}(\boldsymbol{\theta}) f_b((z+1)\lambda_d)}{\sum_d^D \lambda_d f_b((z+1)\lambda_d)}.\end{aligned} \tag{11}$$

In summary, we know how the flux in a band $b$ for a spectrum generated by PPCA coordinates $\boldsymbol{\theta}$ and redshift $z$ can be computed. Now, all we have to do is to convert the observed magnitudes to equivalent fluxes in the spectra[2], $10^{-0.4(T_b - ZP_b)}$, where $M_b$ denotes the magnitude and $ZP_b$ is the zero-point[3] for band $b$. Lastly, we need to multiply the flux with an arbitrary scaling constant $s$, in order to accommodate for the difference in average flux, i.e. $s\mathcal{I}_b$.

The free model parameters are $s$, $\boldsymbol{\theta}$ and $z$. Though in principle coordinate vector $\boldsymbol{\theta}$ is continuous, we found out in preliminary numerical experiments that it was very easy to overfit it. The reason of overfitting is because, without imposing any control on it, coordinate $\boldsymbol{\theta}$ can move away from the region in $\mathbb{R}^Q$ occupied by the PPCA-projections of the spectra. This has as a consequence that $\boldsymbol{\theta}$ is mapped to arbitrary (physically implausible) spectra in its attempt to explain the observed fluxes and hence $\boldsymbol{\theta}$ is overfitted. In order to circumvent this problem[4], we choose to model parameter $\boldsymbol{\theta}$ as a discrete parameter that takes values in the set $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N\}$, the low-dimensional projections of the observed spectra $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N$. The low-dimensional projections $\boldsymbol{\theta}_n$ can be interpreted as a discrete set of low-dimensional coordinates that give rise to a set of spectra $\hat{\boldsymbol{\xi}}_n$.

Similarly, we also noticed that gradient optimisation of $z$ is not practical. The objective function (likelihood in the next section) is plagued by multiple local optima hence gradient optimisation gets easily trapped. Hence, we also optimise $z$ by searching on a regular grid, see Table I.

### C. Model likelihood

Assuming Gaussian noise on the observed data, we define the following likelihood function for our model:

$$p(\boldsymbol{M}|\boldsymbol{\theta}, z, s) = \prod_b \mathcal{N}\left(10^{-0.4(M_b - ZP_b)} | s\mathcal{I}_b\left(\boldsymbol{\theta}, z\right), {\sigma_b}^2\right) \;, \tag{12}$$

where $\boldsymbol{M}$ is the vector of observed magnitudes, and $\sigma_b^2$ is respective variance in band $b$.

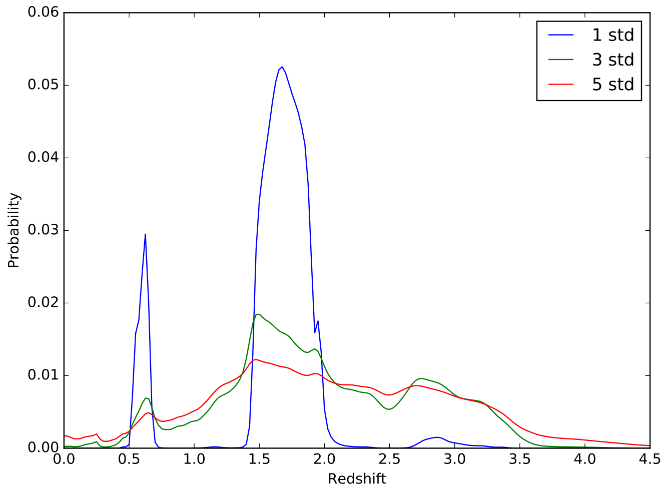

Fig. 2: The figure illustrates the multimodal posterior distribution calculated by the proposed model, and how this posterior changes in the presence of increasing noise $\sigma_b$ in the observed magnitudes. The plotting range of the redshift is limited for illustration purposes.

We complete the model by defining priors on the free parameters. For scaling[5], we impose the uninformative prior:

$$p(s) = \frac{1}{\log s_{max} - \log s_{min}} \frac{1}{s} \;, \tag{13}$$

while for the redshift and coordinates a uniform prior is assumed respectively. Given observed magnitude data $\boldsymbol{M}$, we compute the posterior of $z$ by integrating over the discrete set of coordinates $\boldsymbol{\theta}_n$ and normalising:

$$\begin{aligned}p(z_i|\boldsymbol{M}) &= \frac{\int \sum_n p(\boldsymbol{M}|\boldsymbol{\theta}_n, z_i, s) p(\boldsymbol{\theta}_n) p(z_i) p(s) ds}{\sum_{z_j} \int \sum_n p(\boldsymbol{M}|\boldsymbol{\theta}_n, z_j, s) p(\boldsymbol{\theta}_n) p(z_j) p(s) ds} \\ &\approx \frac{\sum_{s_j} \sum_n p(\boldsymbol{M}|\boldsymbol{\theta}_n, z_i, s_j) p(\boldsymbol{\theta}_n) p(z_i) p(s_j)}{\sum_{z_j} \sum_{s_j} \sum_n p(\boldsymbol{M}|\boldsymbol{\theta}_n, z_j, s_j) p(\boldsymbol{\theta}_n) p(z_j) p(s_j)} \;,\end{aligned} \tag{14}$$

where the integration over the scaling parameter $s$ is approximated by a sum on a regular grid. We summarise the evaluation grid of the discretised model parameters in Tab. I.

We note that the posterior in Eq. (14) is a probability mass function (PMF) defined on a discrete support, while the target redshifts $z$ in the dataset take their values in a continuous interval. Hence, we cannot evaluate the posterior in Eq. (14) on arbitrary values. We therefore convert the PMF posterior into a piecewise uniform distribution. That is, we define a uniform distribution for each interval between two grid points and scale it by $p(z_i|\boldsymbol{M})$. The new probability density distribution of our model simply reads:

$$p_{spec}(z|\boldsymbol{M}) = \frac{1}{0.025} \frac{p(z_i|\boldsymbol{M})}{\sum_{z_j} p(z_j|\boldsymbol{M})} \;, \tag{15}$$

[2] Note that we prefer to work in flux space, as the PCA might well return also negative spectra, which are non-physical, but can still occur as part of the optimization process.

[3] In our numerical experiments zero-points are arbitrary. In general, for other data, we need to gauge them correctly.

[4] Alternatively, we could have imposed a penalty on the continuous parameters $\boldsymbol{\theta}$ that penalises distance from the data populated region in $\mathbb{R}^Q$.

[5] Note that scaling can be omitted by optimizing colors instead of bands, then of course the input dimension would decrease by one accordingly.

TABLE I: Evaluation parameters

| Parameter | Regular grid | Grid points |
|---|---|---|
| scaling $s$ | $\{0.5, 0.525, \ldots, 2.0\}$ | 60 |
| coordinate | $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ | 5000 |
| redshift $z$ | $\{0.0, 0.025, \ldots, 5.5\}$ | 220 |

for $z \in [0.025(i-1), 0.025i]$, i.e. $z$ belongs to the $i-$th interval of the grid in Table I. The normalisation constant results from the fact that the distribution $p_{spec}$ consists of uniform distributions all of interval length 0.025; each uniform distribution is scaled by $p(z_i|\boldsymbol{M})$ and thus we also need to normalise by the sum $\sum_{z_j} p(z_j|\boldsymbol{M})$.

Finally, if a point estimate is required from the proposed model, we use to that end the highest mode $\arg\max_z p_{spec}(z)$.

### D. Demonstration of model behaviour

The posterior calculated in Eq. (15) has the following benefits: first, the model expresses a multimodal distribution over the possible redshifts $z$ that explain the observed magnitudes $\boldsymbol{M}$. Secondly, the posterior changes in response to the presence of noise in the observations $\boldsymbol{M}$. This is illustrated in Fig. 2 on a data item from the dataset described in Section III-A. Finally, we have control over the prior on $z$.

In Fig. 3, we pick a test object (from the dataset described in Section III-A) whose colour admits multiple redshifts as possible solutions. Though these multiple alternative redshifts are not directly available, we make the assumption that they can be approximately recovered as the redshifts that belong to objects that are closest (in Euclidean sense) to the test object in terms of colour. These retrieved alternative redshifts are plotted as grey lines. Hence, the first thing to note is that the grey lines are clustered around two locations. This tells us that that objects with very similar colours may correspond to really different redshifts, i.e. a colour can be associated with multiple distinct redshifts. We also plot the true redshift as a red line, which is known in this case as spectral data are available for the test object. The posterior distribution $p_{spec}(z)$ obtained from our model, given in Eq. 15, is plotted as a blue line. It is very pleasing to see that both dominating modes of the model posterior overlap with the alternative redshifts (grey lines). The fact that the true redshift appears closer to the smaller mode as opposed to the larger one does not mean that our prediction is wrong; it merely means that our model assigns less probability mass than it would assign to other redshifts. Indeed, given that the number of alternative redshifts (grey lines) is less at the smaller than at the larger mode justifies this behaviour. The plot clearly shows that both dominating modes of the model posterior are justified as grey lines appear close to either of them. We also plot the prediction of the random forest (trained on data described in Section III-A) as a line in cyan. As previously explained, the random forest cannot cope with multimodality and hence its prediction is a compromise of the multiple modes. In this particular case, it leads to a prediction located in a region where the probability density is low, i.e. in a region where no grey lines are present.

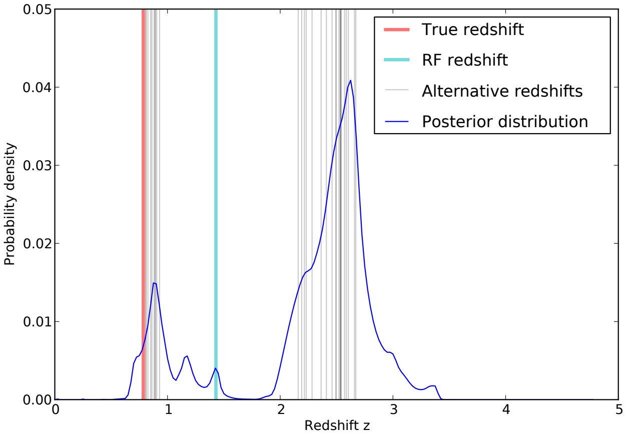

Fig. 3: For a given data item the actual redshift is shown in red. The redshift obtained by random forest regression is shown in cyan. In addition, alternative redshifts, of objects similar in terms of colour, are shown as grey lines. In blue we plot the density of the proposed model. The predicted modes appear justified: one mode explains a smaller cluster of alternative redshfits (that includes the true redshift), while the other explains a larger, distinct group of alternatives redshifts.

## III. Numerical experiments

### A. Data

We demonstrate our proposed approach on a small subset of quasars contained in the BOSS catalogue. First, we extract 7506 randomly selected quasar spectra from BOSS which we divide into a training (5000) and test set (2506). The redshift distributions are shown in Fig. 4. The idea is now to extract the photometry directly from the spectra instead of using their observed direct photometric counterparts. This way of approaching the problem has the following advantages:

- no calibration of the zero points needed
- no uncertainties in the observables (spectra are considered noiseless)
- full control over how data have been generated.

One of the downsides of using the spectra is that only part of the $u$ band is covered and thus we have only 3 colors at our disposal for inferring redshift (in the presented case these will be the three independent colors $g-r$, $g-i$, $g-z$). Note that in our methodology the fluxes themselves are used instead of colours and therefore the data presented to our algorithm are four dimensional. This is not an advantage as our model contains an additional scaling that has to be optimized.

The data are randomly split into a training and testing set of $N = 5000$ and $N_{test} = 2506$ objects respectively. In our approach, the training set is used to train the PPCA and project the 5000 spectra $\boldsymbol{\xi}_n$ to the the low-dimensional space $\mathbb{R}^Q$ and in order obtain low-dimensional coordinates $\boldsymbol{\theta}_n$.

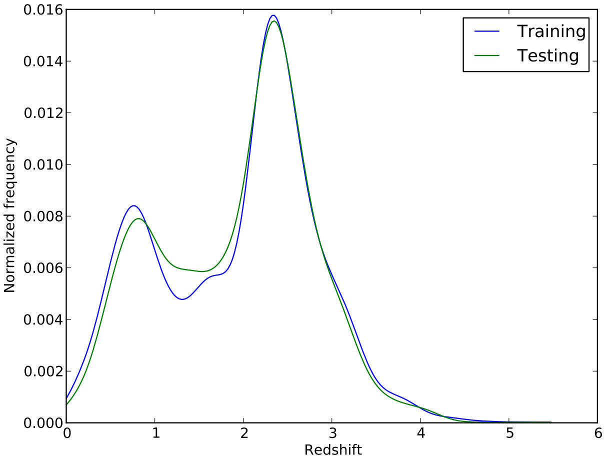

Fig. 4: Redshift distribution of training and test data.

### B. Preprocessing

All the required spectra are downloaded from the SDSS server. In a first step, all spectra are binned with a binning factor of 10 according to the following rules:

$$\lambda^j_{new} = \frac{1}{10}\sum_{i=10j}^{10(j+1)} \lambda^i_{old}$$

$$f^j_{new} = \frac{\sum_{i=10j}^{10(j+1)} f^i_{old}(\Delta f^i_{old})^{-2}}{\sum_{i=10j}^{10(j+1)} (\Delta f^i_{old})^{-2}}$$

$$\Delta f^j_{new} = \frac{1}{\sum_{i=10j}^{10(j+1)} (\Delta f^i_{old})^{-2}}$$

where $\lambda$, $f$ and $\Delta f$ are the wavelength, the spectral flux and the error of the spectral flux respectively. Subsequently, the spectra are shifted into their rest-frame and the flux values are extracted on a fixed grid ($\lambda \in [500, 10400]$ in 1000 equally spaced steps) using spline interpolation. As aforementioned, missing parts are given a value of $\bar{\xi}_n$ and a standard deviation of $\sigma_{high}$.

### C. Performance Criteria

We compare the algorithms using the following criteria:

- Root mean squared error, $RMSE$. A way of measuring error in redshift regression problems is the normalized redshift deviation $\Delta z_{norm} = \frac{z_{reg} - z_{true}}{1+z_{true}}$. Hence:

$$RMSE = \sqrt{\frac{1}{N_{test}}\sum_{n=1}^{N_{test}} \Delta z^2_{norm,n}} \ . \qquad (16)$$

- Median absolute deviation, $MAD$:

$$\frac{1}{N_{test}}\sum_{n=1}^{N_{test}} |\Delta z_{norm,n} - median(\Delta z_{norm})| \ . \qquad (17)$$

  The MAD is less susceptible to highly deviating objects.

- Likelihood, $\frac{1}{N}\log\mathrm{L_{True}}$. This measure expresses how well the data are explained under a model, i.e. the model learnt by each of the three candidate algorithms (see Section III-D). We compute under each model the likelihood averaged over all data items in the test set:

$$\frac{1}{N}\log\mathrm{L_{True}} = \frac{1}{N_{test}}\sum_{n=1}^{N_{test}} p_{model}\left(z_{n;\,true}\right) \ , \qquad (18)$$

  where $p_{model} \in \{p_{GP}, p_{RF}, p_{spec}\}$.

### D. Algorithms

The following algorithms are put to the test:

- Proposed approach. We set the number of embedding dimensions for PPCA to $Q = 10$. Concerning the treatment of missing values in the observed spectra, we set $\bar{\xi}_n$ equal to the average of observed values of spectrum $\boldsymbol{\xi}_n$, and use a fixed $\sigma_{high} = 1000$ for all spectra.
  The $RMSE$ and $MAD$ performance criteria described in Section III-C, require that models deliver a point prediction. For the proposed model, we take its point prediction to be the highest mode of the posterior $\arg\max_z p_{spec}(z)$.
- Random forest. Due to its popularity and success in the astronomical community, we include the random forest (RF) [8] as a candidate algorithm. The number of trees is set to 1000. For comparison purposes, we derive a likelihood function for the RF. The likelihood is simply defined as a Gaussian distribution with a mean given by the prediction $z_{RF}$ of the RF and the standard deviation calculated by the residuals (here $\sigma = 0.30$). Thus we obtain, $p_{RF}(z) = \mathcal{N}\left(z|z_{RF}, \sigma^2\right)$. Alternatively, we also optimise $\sigma$ so that criterion $\frac{1}{N}\log\mathrm{L_{True}}$ is optimized ($\sigma = 0.76$), please see Section III-C.
- Gaussian Process. We include Gaussian process (GPs) [14] in our comparison since it is a flexible model that enjoys automatic regularisation and outputs a Gaussian predictive density $p_{GP}(z)$. We employ the standard RBF kernel.

### E. Results

In Table II, we show how the three algorithms fare according to the performance criteria. For the commonly used $RMSE$, we can see that the GP and RF perform very similar and significantly better than the proposed approach. However, in the other two criteria the presented algorithm performs much better than its competitors. We report in detail the results for each criterion.

In Fig. 5, the regressed redshift is plotted against the true one and additionally a histogram over $\Delta z_{norm}$ is shown for each of the algorithms. On a first look, we clearly see why the $RMSE$ is much worse for the proposed model. While for the GP and RF the points are closer to the diagonal line (left column of plots in Fig. 5), there are some very drastic deviations apparent in our algorithm. This behaviour can be explained by the fact that the GP and RF adapt to the distribution of $z$ in the dataset, i.e. they learn to a certain degree that most redshifts fall in

TABLE II: Summary of performance criteria.

| | Our approach | RF<br>($\sigma = 0.30$) | RF<br>($\sigma = 0.76$) | GP |
|---|---|---|---|---|
| $RMSE$ | 0.476 | 0.344 | 0.344 | **0.326** |
| $MAD$ | **0.078** | 0.123 | 0.123 | 0.111 |
| $\frac{1}{N}\mathrm{logL}_{\mathrm{True}}$ | **-0.514** | -2.979 | -1.138 | -77.875 |

the interval. This is reasonable, if the redshift distribution is *similar* for training and testing. If this cannot be guaranteed (as in most realistic settings, since the observational biases between surveys can be different), this will effectively lead to an amplification of this bias and thus even worse predictions will be produced, as shown later on. Our model adopts as a prior the uniform distribution over redshifts, but of course we can influence this behaviour by choosing as a prior the distributions of redshifts in the dataset. If we do so, the predictions get closer to the diagonal (not shown in figures), and we obtain an $RMSE$ of 0.400.

The $MAD$ criterion is less sensitive to large deviations.As a consequence, the $MAD$ measures rather the width of the central distribution than the width of the full distribution. As seen in Fig. 5, the predictions obtained from our model are much precise than the ones by the RF or GP. This becomes even clearer if we consider the fraction of objects that deviate more than a certain value, cf. Fig. 6. For the vast majority of the objects ($\approx$ 70%) the deviation from the true value is considerably lower than for the RF and GP predictions.

However, given that we are dealing with a problem where every set of magnitudes (or colours) admits multiple solutions, the use of the $RMSE$ and $MAD$ is not appropriate as they focus on comparing a point prediction to a single target solution[6] $z$, i.e. *they are both inherently unable to measure how well a multimodal prediction does*. Hence, $\frac{1}{N} logL_{\mathrm{True}}$ is better at quantifying performance as it correctly takes into account the multimodal predictive density of our model. We therefore see that the proposed model displays a considerable higher likelihood than the RF and the GP.

A further important consequence of the multimodal nature of redshift prediction is that plots of the type on the left column of Fig. 5 are not appropriate for displaying performance. For problems where each prediction seeks to match a unique target solution, these plots state that predictions and targets should meet on the diagonal. *However, for problems where multiple solutions are possible we can no longer demand that the multiple solutions lie on the diagonal since they are distinct solutions*. Hence, such plots are actually inappropriate and are shown here just for the sake of aligning with previous works.

## IV. Discussion and Conclusion

The presented approach can be developed further in two aspects: the methodological one and the astronomical one. Currently, as aforementioned, the coordinates $\boldsymbol{\theta}$ are discretised in order to prevent overfitting. It would be desirable to provide some control mechanism in order to constrain the coordinates $\boldsymbol{\theta}$ in regions of the lower dimensional space populated by data. From an astronomical point of view, there is more work to be done. So far we have just demonstrated the concept on a small dataset where the magnitudes were extracted with the provided filter curves and (known and noise-free) zero points were added. We chose this setting as we wanted to have full control on the model and not to be distracted by erroneous and noisy calibrations. It is important to notice that a purely data-driven approach can deal with this quite naturally, while the presented algorithm depends heavily on the correctness of these calibrations. On the other hand, it is of course also possible to include a given uncertainty of the zero points into the model and also this can be cross-validated on a hold-out

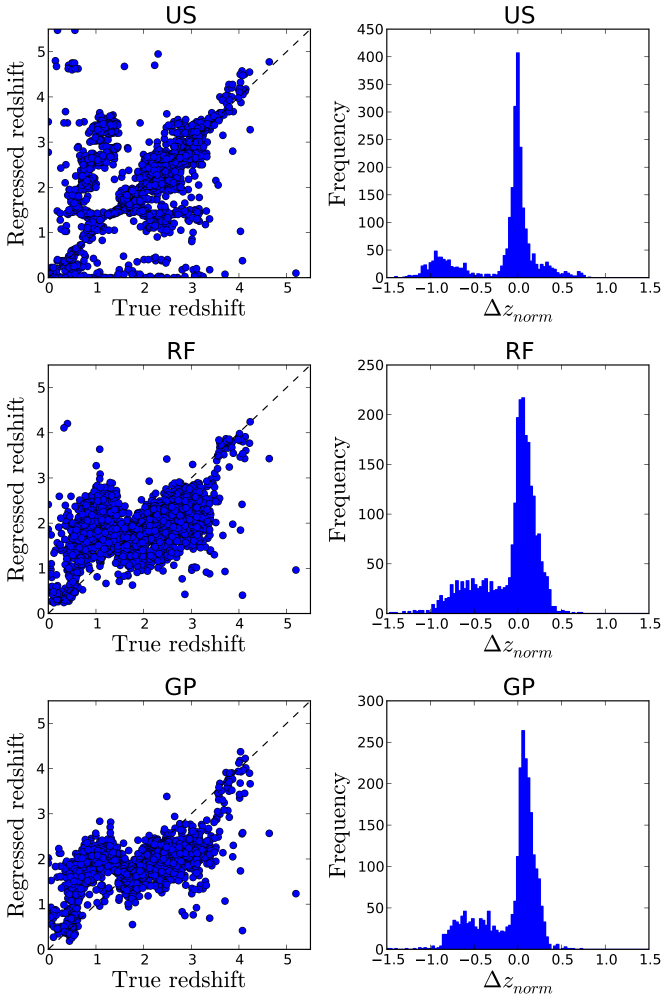

Fig. 5: Each row shows the results of the experiment noted on top of each plot (our model, random forest and Gaussian process). On the left side the value obtained by the regression algorithm is plotted against the actual value. On the right one can see a histogram of $\Delta z_{norm} = \frac{z_{reg} - z_{true}}{1+z_{true}}$. One can clearly see that for the proposed model the peak is much sharper, but the left wing is much more pronounced than for the other two algorithms.

[6] Which is fine for regression problems where an input is associated with a single target.

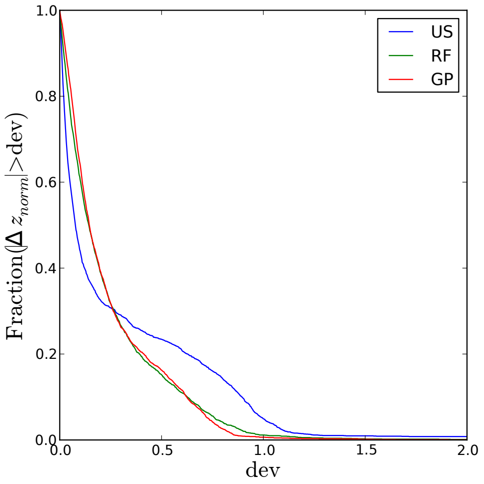

Fig. 6: Fraction of objects deviating more than $|\Delta z_{norm}| > dev$. Here, we only take a point prediction for our model $\arg\max_z p_{spec}(z)$. While the RF and GP show very similar behaviour, the proposed algorithm predicts the redshift for 60% of the objects much more precisely but is heavily influenced from redshifts deviating more than 0.3.

set. In summary, a much more detailed understanding of how photometric measurements relate to the spectra is required.

An advantage of our model is, that we can include uncertainty of the photometric measurements. This includes also *missing* values which are a common struggle in astronomy due to the different coverage and depth of the surveys. An interesting prospect is to extend the model towards the infrared. At the moment, our coverage above $1\mu m$ is very shallow and thus it would be desirable to retrieve near-infrared to mid-infrared spectra of low-redshifted quasars (as otherwise the rest-frame would be in the optical again). This would allow us to include also infrared data as then the coverage of the coordinates would reach into the near infrared. It is important to notice that it does not matter whether the infrared spectra are the same objects as the optical ones, it only has to be guaranteed that there is considerable overlap with the coordinates as they are now.

Another issue is that standard regression models do not have control on the prior of $z$ and are thus implicitly biased by how $z$ is distributed in a given dataset. While this might be of advantage in some cases, it is a generally an unwanted side-effect of the training procedure. In contrast, the prior on $z$ is easily controlled in our model. In the presence of physical apriori information an informed prior may be employed.

In conclusion, based on first principles, we have formulated a simple probabilistic model which expresses a multimodal predictive density for photometric redshifts. Numerical comparisons support our line of work. We point out that by design, in their standard formulation, the RF and GP cannot fully address the problem of redshift estimation as they cannot predict multiple distinct solutions.

## Acknowledgement

The authors gratefully acknowledge the support of the Klaus Tschira Foundation. The authors thank Pavlos Protopapas for his input and also acknowledge SDSS for their valuable service in making data access possible. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

## References

[1] F. B. Abdalla, A. Amara, P. Capak, E. S. Cypriano, O. Lahav, and J. Rhodes. Photometric redshifts for weak lensing tomography from space: the role of optical and near infrared photometry. *Monthly Notices of the Royal Astronomical Society*, 387:969–986, 2008.

[2] R. Antonucci. Unified models for active galactic nuclei and quasars. *Annual review of astronomy and astrophysics*, 31:473–521, 1993.

[3] Pablo Arnalte-Mur, Alberto Fernndez-Soto, Vicent J. Martnez, Enn Saar, Pekka Heinmki, and Ivan Suhhonenko. Recovering the real-space correlation function from photometric redshift surveys. *Monthly Notices of the Royal Astronomical Society*, 394(3):1631–1639, 2009.

[4] Narciso Bentez. Bayesian photometric redshift estimation. *The Astrophysical Journal*, 536(2):571, 2000.

[5] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[6] C. M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, 1994.

[7] M. Bolzonella, J.-M. Miralles, and R. Pelló. Photometric redshifts based on standard SED fitting procedures. *Astronomy and Astrophysics*, 363:476–492, 2000.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5, 2001.

[9] T. Budavári, A. S. Szalay, A. J. Connolly, I. Csabai, and M. Dickinson. Creating Spectral Templates from Multicolor Redshift Surveys. *The Astronomical Journal*, 120:1588–1598, 2000.

[10] Michael I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[11] G. Kauffmann and M. Haehnelt. A unified model for the evolution of galaxies and quasars. *Monthly Notices of the Royal Astronomical Society*, 311:576–588, 2000.

[12] O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio. Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation. *Monthly Notices of the Royal Astronomical Society*, 418:2165–2195, 2011.

[13] A. L. O'Mill, F. Duplancic, D. García Lambas, and L. Sodré, Jr. Photometric redshifts and k-corrections for the Sloan Digital Sky Survey Data Release 7. *Monthly Notices of the Royal Astronomical Society*, 413:1395–1408, 2011.

[14] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[15] M. U. Subbarao, A. J. Connolly, A. S. Szalay, and D. C. Koo. Luminosity Functions From Photometric Redshifts. I. Techniques. *Astronomical Journal*, 112:929, 1996.

[16] A. Tchekhovskoy, R. Narayan, and J. C. McKinney. Black Hole Spin and The Radio Loud/Quiet Dichotomy of Active Galactic Nuclei. *The Astrophysical Journal*, 711:50–63, 2010.

[17] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[18] C. M. Urry and P. Padovani. Unified Schemes for Radio-Loud Active Galactic Nuclei. *Publications of the Astronomical Society of the Pacific*, 107:803, 1995.

[19] X. B. Wu, Guoqiang Hao, Zhendong Jia, Yanxia Zhang, and Nanbo Peng. SDSS quasars in the WISE preliminary data release and quasar candidate selection with optical/infrared colors. *The Astronomical Journal*, 144:49, 2012.

[20] X. B. Wu and Z. Jia. Quasar candidate selection and photometric redshift estimation based on SDSS and UKIDSS data. *Monthly Notices of the Royal Astronomical Society*, 406:1583–1594, 2010.