Photometric Redshift Estimation of Quasars: Local versus Global Regression

Fabian Gieseke[1] f.gieseke@uni-oldenburg.de

Kai Lars Polsterer[2] polsterer@astro.ruhr-uni-bochum.de

Peter-Christian Zinn[2] zinn@astro.rub.de

[1]Computational Intelligence, Computer Science Department, University of Oldenburg, Germany [2]Astronomisches Institut, Department of Physics and Astronomy, Ruhr-University of Bochum, Germany

The Task: Photometric Redshift Estimation

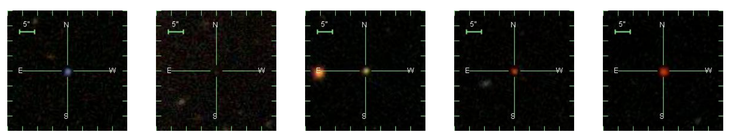
Abstract

The task of estimating an object's redshift based on photometric data is one of the most important ones in astronomy. This is especially the case for quasars. Common approaches for this regression task in the field of astronomy are based on nearest neighbor search, template fitting schemes, or combinations of, e.g, standard clustering and regression techniques. As we show in this work, simple frameworks like the k-nearest neighbor regression scheme work extremely well if one considers the overall feature space (containing patterns of all objects with low, middle, and high redshifts). However, such methods seem to fail as soon as only few or even no training patterns are given in the appropriate region of the feature space. In the literature, a wide range of other regression techniques can be found. Among the most popular ones are regularized regression schemes like ridge regression or support vector regression. In this work, we show that an out-of-the-box application of this type of schemes for the whole feature space is difficult due to the involved computational requirements and the specific properties of the data at hand. However, in contrast to nearest neighbor search schemes, such methods can be employed to extrapolate, i.e, to predict redshifts for patterns in new, unseen regions of the feature space.

Data

We describe the use of machine learning regression models to estimate the *redshift* of *quasi-stellar radio sources (quasars)* based on *photometric data*. Our data set is based on the *Sloan Digital Sky Survey*, which is said to be "one of the most ambitious and influential surveys in the history of astronomy" [4]. The data for this survey has been obtained via a 2.5 meter telescope at the Apache Point Observatory which is equipped with two special-purpose instruments: a 120 mega pixel camera and a pair of spectrographs that collect photometric and spectroscopic data, respectively.

Distant Quasars (RBG images)



Regression Models

For regression problems, one is given a *training set* $T = \{(\mathbf{x}'_1, y'_1), \ldots, (\mathbf{x}'_n, y'_n)\} \subset \mathcal{X} \times \mathbb{R}$ with *patterns* $\mathbf{x}'_i \in \mathcal{X}$ and associated *labels* $y'_i \in \mathbb{R}$. The goal of the learning process is to generate a *model* that can predict reasonable labels for unseen patterns.

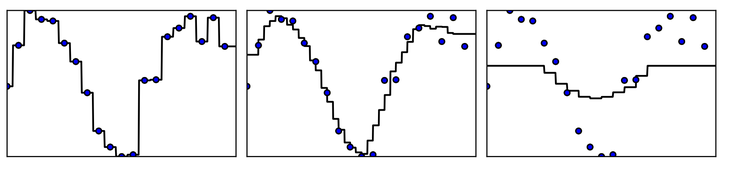
k-Nearest Neighbor Regression

The k-*nearest neighbor* (kNN) regression model uses the k closest objects from the given set of objects to assign a label to a new object [1]. More precisely, the regression model is given by the function

$$f(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}'_i \in N_k(\mathbf{x})} y'_i, \qquad (1)$$

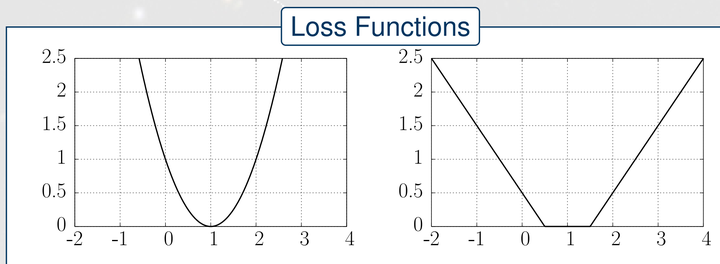
where $N_k(\mathbf{x})$ denotes the k-nearest neighbors of $\mathbf{x} \in \mathcal{X}$ in the training set T. To define closeness, arbitrary metrics can be used. A popular choice is the Euclidean metric. The parameter k determines the trade-off between local and global influence of the patterns.

Influence of Parameter k



Support Vector Regression

Loss Functions



Support vector regression (SVR) [1] models are of the form

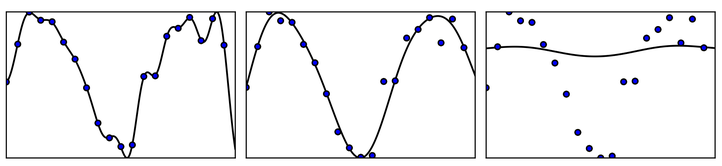
$$\inf_{f \in \mathcal{H}_k, b \in \mathbb{R}} \frac{1}{l} \sum_{i=1}^{l} \max\left(0, |y'_i - (f(\mathbf{x}'_i) + b)| - \varepsilon\right) + \lambda \|f\|^2_{\mathcal{H}_k}, \qquad (2)$$

where $\mathcal{L}(y,t) = \max(0, |y-t| - \varepsilon)$ with $\varepsilon \in \mathbb{R}^+$ is called the ε-*insensitive loss*. The space $\mathcal{H}_k$ is an appropriate *hypothesis space* containing functions of the form

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, \mathbf{x}'_i) \qquad (3)$$

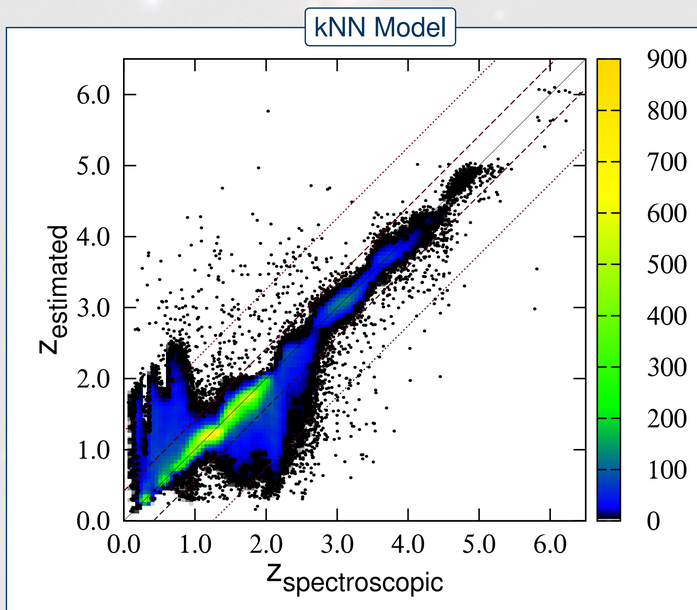
with coefficients $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Here, the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *positive semidefinite kernel function* which can be seen as similarity measure for the patterns. The first term of the above objective measures how well the function f can predict the (real-valued) labels and the second term measures the complexity of the model. The parameter λ determines the trade-off between both objectives. Given $\mathcal{X} = \mathbb{R}^d$, common choices for the kernel function are the *linear kernel* $k(\mathbf{x}'_i, \mathbf{x}'_j) = \langle \mathbf{x}'_i, \mathbf{x}'_j \rangle$ and the *RBF kernel* $k(\mathbf{x}'_i, \mathbf{x}'_j) = \exp(-\frac{\|\mathbf{x}'_i - \mathbf{x}'_j\|^2}{2\sigma^2})$.

From Complex to Simple Models

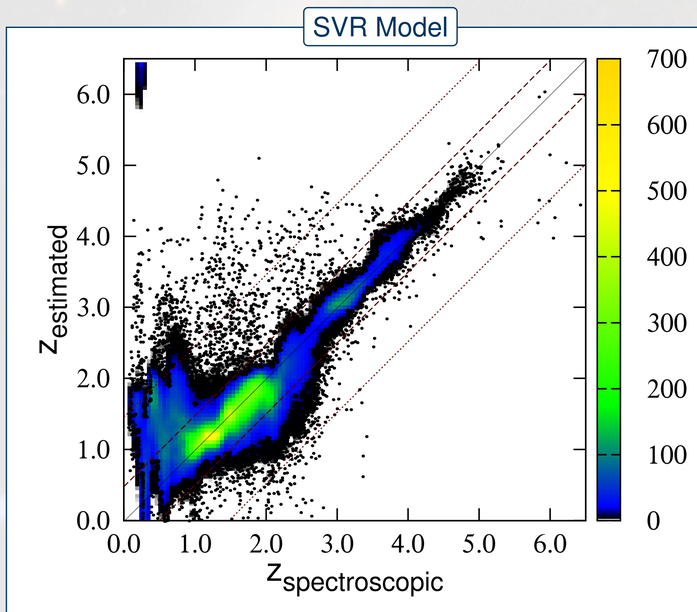


Results of Standard Redshift Estimation

kNN Model

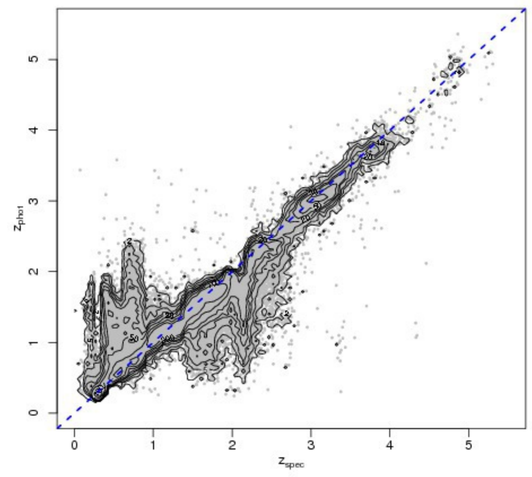


SVR Model



Experimental Setup: The labels are based on the SDSS quasar catalog [3]. As features, we consider all colors from neighbored bands for each of the objects (`u-g`, `g-r`, `r-i`, `i-z`). This results in a data set consisting of $104,440$ patterns in $\mathbb{R}^4$. The kNN model is generated based on the whole data set. The SVR model is trained on a selected subset of patterns (due to the cubic runtime needed to train a model). Both models are tested on the (remaining) patterns.

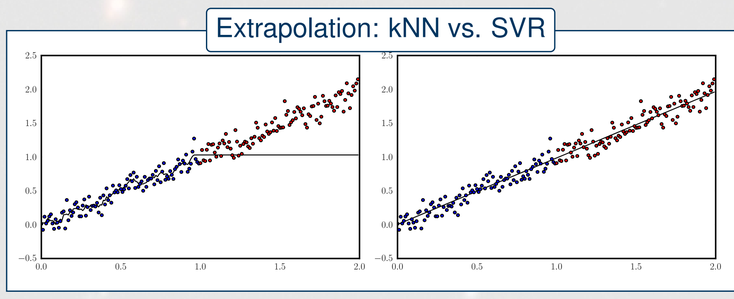
Weak-Gated Experts (Laurino et al.)



Looking over the Tea Cup's Rim

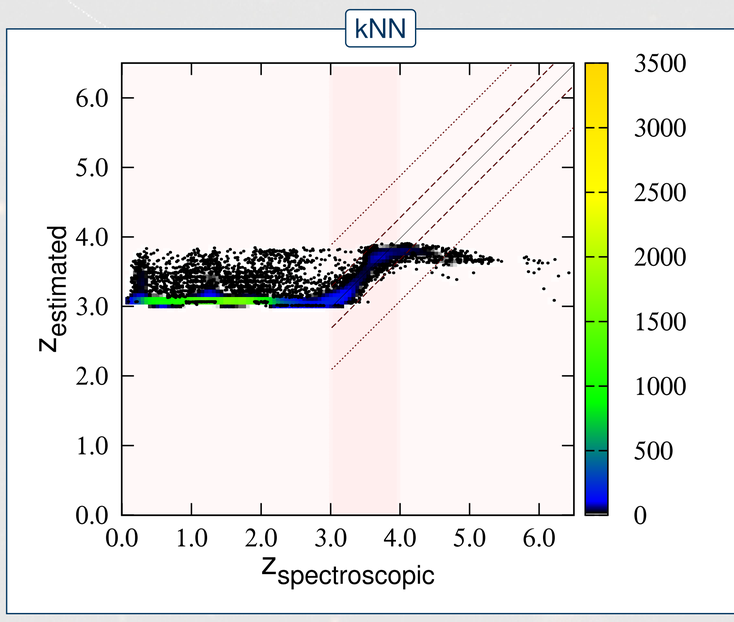
The k-nearest neighbor model is well-suited for densely populated regions of the feature space. However, it cannot predict any *trends* for unseen data. In contrast, support vector regression might yield reasonable models (depending on the data).

Extrapolation: kNN vs. SVR

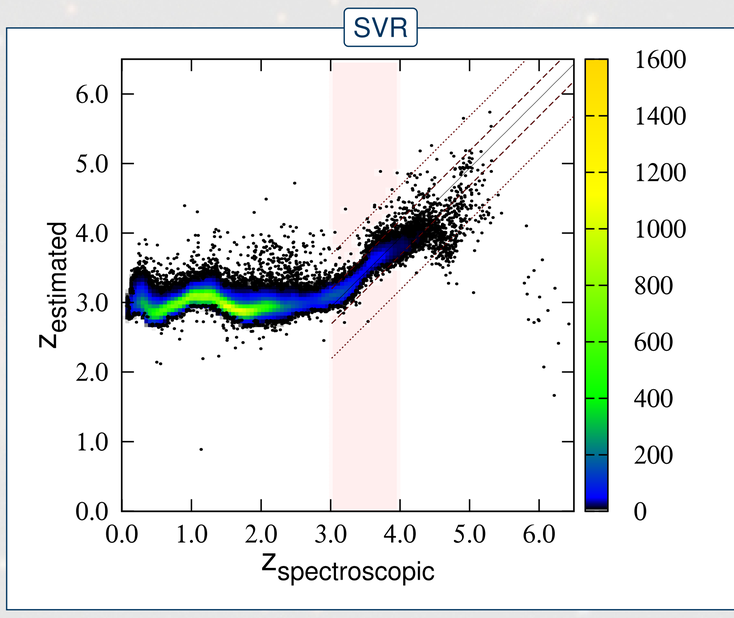


Experimental Setup: Only quasars with $z \in [3,4]$ are used for training the models (again, a selected subset is used for SVR). The remaining patterns are used for testing.

kNN



SVR



References

[1] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning*, Springer, 2009.

[2] Omar Laurino, Raffaele D'Abrusco, Giuseppe Longo, and Giuseppe Riccio. *Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation*, MNRAS, 2011.

[3] Schneider *et al. The Sloan Digital Sky Survey Quasar Catalog. V. Seventh Data Release*, AJ, 139(6): 2360-2373, 2010.

[4] Sloan Digital Sky Survey. `http://www.sdss.org`, October, 2011.

Acknowledgements This work is based on data of the SDSS project [4]. We thank Anna Amelung for the cartoons.

ADASS 2011, Paris, France, November 2011.

Fabian Gieseke

Kai Lars Polsterer

http://www.astro.rub.de/polsterer/ADASS2011a.pdf