Accepted by ApJ in May 2013

Photometric redshifts for Quasars in multi band Surveys

M. Brescia[1,2]

INAF-Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy

brescia@oacn.inaf.it

S. Cavuoti[2]

Department of Physics, University Federico II, via Cinthia 6, I-80126 Napoli, Italy

R. D'Abrusco[3]

Harvard Smithsonian Center for Astrophysics, Cambridge, MA, USA

G. Longo[2,4]

Department of Physics, University Federico II, via Cinthia 6, I-80126 Napoli, Italy

A. Mercurio[1]

INAF-Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy

ABSTRACT

MLPQNA stands for Multi Layer Perceptron with Quasi Newton Algorithm and it is a machine learning method which can be used to cope with regression and classification problems on complex and massive data sets. In this paper we give the formal description of the method and present the results of its application to the evaluation of photometric redshifts for quasars. The data set used for the experiment was obtained by merging four different surveys (SDSS, GALEX, UKIDSS and WISE), thus covering a wide range of wavelengths from the UV

[1] INAF-Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy

[2] Department of Physics, University Federico II, via Cinthia 6, I-80126 Napoli, Italy

[3] Harvard Smithsonian Center for Astrophysics, Cambridge, MA, USA

[4] Visiting Associate, California Institute of Technology, Pasadena, CA, USA

to the mid-infrared. The method is able i) to achieve a very high accuracy; ii) to drastically reduce the number of outliers and catastrophic objects; iii) to discriminate among parameters (or features) on the basis of their significance, so that the number of features used for training and analysis can be optimized in order to reduce both the computational demands and the effects of degeneracy. The best experiment, which makes use of a selected combination of parameters drawn from the four surveys, leads, in terms of Δz_{norm} (i.e. $(z_{spec} - z_{phot})/(1 + z_{spec})$), to an average of $\Delta z_{norm} = 0.004$, a standard deviation $\sigma = 0.069$ and a Median Absolute Deviation $MAD = 0.02$ over the whole redshift range (i.e. $z_{spec} \leq 3.6$), defined by the 4-survey cross-matched spectroscopic sample. The fraction of catastrophic outliers, i.e. of objects with photo-z deviating more than 2σ from the spectroscopic value is $< 3\%$, leading to a $\sigma = 0.035$ after their removal, over the same redshift range. The method is made available to the community through the DAMEWARE web application.

Subject headings: methods: data analysis, methods: machine learning, catalogues, surveys, quasars: general, distances and redshifts

1. Introduction

Photometric redshifts (hereinafter photo-z) provide an estimate of the redshift of sources obtained using photometry instead of spectroscopy. They are in fact driven by: (i) the shape of the broadband continuum of the object's spectroscopic emission, and (ii) by a limited number of strong spectral features (i.e. the one at 4000 Å, the Lyα forest and the Lyman limit), which are still recognizable after the integration of the Spectral Energy Distribution (SED) sampled by the filter's transmission function.

At the price of lower accuracy, photo-z offer several advantages with respect to their spectroscopic counterparts: (i) being derived from intermediate/broad band imaging, photo-z are much more effective in terms of observing time; (ii) they may allow to probe objects much fainter than the spectroscopic flux limit and (iii) under specific conditions, they allow to correct some biases, such as those encountered at high redshift where, as it has been noticed (Fernandez-Soto et al. 2001), spectroscopy is pushed to its limits both by the low signal-to-noise ratio (SNR) in the spectra and by the fact that, in many cases, even when a good signal-to-noise ratio is achieved, the lack of features in the observed spectral range may undermine the estimation of a trustworthy redshift (Lanzetta et al. 1998).

The latter aspect becomes crucial when photometric redshifts methods are applied to quasars (QSO) and, in particular to the construction and characterization of the large, complete samples which are required by modern cosmology. In fact, quasar samples have always been, and still are, constructed either by compiling lists of more or less serendipitous discoveries obtained with different techniques and selection criteria (Veron & Veron 2000), or via a two-step process where the first one consists in the identification of QSO candidates from multi-wavelength surveys, and the second requires the spectroscopic validation of the candidates. In practice, due to the large amount of observing time required by spectroscopy, the latter step is usually optimized by applying the spectroscopic validation procedure just to a more or less significant subsample of the candidates, and then by extrapolating the resulting statistics to the whole sample. Modern surveys are usually so deep and extensive that the number of candidates rapidly becomes too large to be handled with the latter approach. On the other hand, modern multi-wavelength digital surveys also provide such a wealth of information (multi-band high accuracy photometry) that it becomes feasible to approximate the SED of objects over a quite large range of redshifts (Richards et al. 2001a,b; Budavari et al. 2001; Wolf et al. 2004), thus minimizing the need for spectroscopic follow-up.

In the last few years it has in fact been demonstrated that, after having provided an accurate enough photometry and significant wavelength coverage, it is possible to obtain samples of photometrically selected quasars matching the low contamination and high completeness (D'Abrusco et al. 2009; Bovy et al. 2012) required by many fields of modern cosmology. The relevance of these photometric samples will increase more and more in the near future, when the new generation of deeper and more accurate surveys will allow to access larger and more complete samples of QSOs. These *photometric* samples are in fact already being used for a variety of applications such as the measurement of the integrated Sachs–Wolfe effect (Giannantonio et al. 2008), the cosmic magnification bias (Scranton et al. 2005), the clustering of quasars on large (Myers et al. 2006) and small (Hennawi et al. 2006) scales, to quote just a few. Since both candidate selection and photometry redshift estimates are performed on the same data (colors in many bands), it is also apparent that for the same samples, photometric data alone should carry enough information to characterize in an almost univocal way the SED and therefore also to derive accurate estimates of photometric redshifts (D'Abrusco et al. 2009; Laurino et al. 2011; Bovy et al. 2012).

It goes without saying that the utility of the photometric samples goes hand in hand with the development of photo-z methods capable to provide accurate enough estimates of the redshifts.

In this paper we use a new empirical method, named Multi Layer Perceptron with Quasi Newton Algorithm or MLPQNA, and apply it to the evaluation of photometric redshift of quasars. In section 2 we discuss the datasets used for the experiments and in section 3 we present both a detailed description of the MLPQNA method and the statistical indicators used throughout the paper. We wish to stress that the lack of a common agreement on such indicators is among the main obstacles in comparing the performances of different methods. In section 4 we describe the experiments performed in order to select the best combination of input parameters, bands and network topology. The results of these experiments are summarized and discussed in section 5, where we also present the final performances of the best experiments. Finally, we compare our results with those available in literature and draw some general conclusions.

A short appendix provides the reader with the math behind the Quasi Newton Algorithm.

2. The Dataset

The sample of quasars, used in the experiments described in this paper, is based on the spectroscopically selected quasars from the SDSS-DR7 database (table *Star* of the SDSS database). According to the spectroscopic classification index (*index SP* or *specClass*) provided in the SDSS-DR7 release (Schneider et al. 2010), we selected quasars, for which a reliable measure of the spectroscopic redshifts (with $zConf > 0.90$) is available.

We then cross-matched the SDSS quasars sample identified as point sources with clean measured photometry in all filters (*ugriz*), with the latest versions of the datasets from: GALEX (Martin et al. 2005), UKIDSS (Lawrence et al. 2007) and WISE (Wright et al. 2010). These three surveys observed large fractions of the sky in the ultraviolet, near infrared and middle infrared spectral intervals, respectively. After the cross matching we obtained a series of multi-band catalogues, defined as it follows.

SDSS - (DR7) (Aihara et al. 2011) has observed $\sim 1.4 \times 10^4$ deg^2 of the sky in 5 bands (*ugriz*) covering the [3551, 8931] Å wavelengths range. Photometric SDSS observations reach the limiting magnitude of 22.2 in the r band (95% completeness for point sources; Abazajian et al. 2009).

GALEX - (DR6/7) (Martin et al. 2005) is a 2-band survey (*nuv, fuv* for near and far ultraviolet respectively) covering the [1300,3000] Å wavelength interval. The GALEX photometric survey has observed the whole sky to the near ultraviolet limiting magnitude nuv=20.5.

UKIDSS - (DR9) (Lawrence et al. 2007) has been designed to be the SDSS infrared counterpart and covers ∼7000 deg^2 of the sky in the $YJHK$ near-infrared bands covering the $\sim$ 0.9 to 2.4 μm spectral range down to the limiting magnitude K=18.3. The Large Area Survey (LAS) has imaged $\sim$ 4000 deg^2 (overlapping with the SDSS), with the additional Y band down to the limiting magnitude of 20.5.

The **WISE** mission (Wright et al. 2010) has observed the entire sky in the mid-infrared spectral interval at 3.4, 4.6, 12, and 22 μm with an angular resolution of 6.1″, 6.4″, 6.5″ and 12.0″ in the four bands, achieving 5σ point source sensitivities of 0.08, 0.11, 1 and 6 mJy in unconfused regions on the ecliptic, respectively. The astrometric accuracy of WISE is $\sim 0.50'', 0.26'', 0.26''$, and 1.4″ for the four WISE bands, respectively.

The transmission curves of all filters related with the four surveys are shown in Fig. 1. All these surveys present a large common overlap region and overall good astrometry with comparable astrometric accuracy. In order to cross-match the catalogues we used a maximum radius $r = 1.5''$ to associate the optical quasars to counterparts in each of the three catalogs. Afterwards we rejected all sources containing one or more missing data in any of their photometric parameters. In this case with the term *missing data* we mean undefined numerical values underlying either not detected or contaminated magnitude measurements. This last step is crucial in empirical methods since the presence of missing data might affect their generalization capabilities (Marlin 2008).

The resulting number of objects in the datasets used for the experiments are:

- SDSS: $\sim 1.1 \times 10^5$;
- SDSS ∩ GALEX: $\sim 4.5 \times 10^4$;
- SDSS ∩ UKIDSS: $\sim 3.1 \times 10^4$;
- SDSS ∩ GALEX ∩ UKIDSS: $\sim 1.5 \times 10^4$;
- SDSS ∩ GALEX ∩ UKIDSS ∩ WISE: $\sim 1.4 \times 10^4$;

An additional dataset was produced by decimating the final *four-surveys* cross-matched catalogue. This dataset was used to perform the preliminary feature-selection or *pruning* phase (see Sec. 4.2) and consisted of $\sim 3.8 \times 10^3$ objects, each observed in 15 bands (4 UKIDSS, 2 GALEX, 5 SDSS and 4 WISE) and with accurate spectroscopic redshift estimates. The decimation was needed to reduce the computational time needed to perform the large number of experiments described in what follows. For some bands there were multiple measurements (i.e. magnitude measured accordingly to different definitions) and therefore we are left with a total of 43 different features.

Finally, in producing training and test sets we made sure that they had compatible spectroscopic redshifts distributions (see Fig. 2).

3. The Method

This section is dedicated to the description of the machine learning method used for the experiments. All mathematical details are reported in the Appendix.

3.1. Multi Layer Perceptron (MLP)

From a technical point of view, the MLPQNA method, is a Multi Layer Perceptron (MLP; Bishop 2006) neural network trained by a learning rule based on the Quasi Newton Algorithm (QNA); in other words and as it is synthesized in the acronym, MLPQNA differs from more traditional MLP's implementations in the way the optimal solution of the regression problem is found. In previous papers, most of the characteristics of the method have been described in the contexts of both classification (Brescia et al. 2012) and regression (Cavuoti et al. 2012a).

According to Bishop (2006), feed forward neural networks (in their various implementations) provide a general framework for representing non linear functional mappings between a set of input variables (also called features) and a set of output variables (the targets). The training of a neural network can be in fact seen as the search for the function which minimizes the errors of the predicted values with respect to the true values available for a small but significant subsample of objects in the same parameter space. This subset is also called *training set* or *Knowledge Base* (KB). The final performances of a specific Neural Network (NN) depend on many factors, such as topology, the way the minimum of the error function is searched and found, the way errors are computed, as well as the intrinsic quality of training data.

The formal description of a feed-forward neural network with two computational layers is given in the Eq. 1:

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} g \left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i \right) \tag{1}$$

Equation 1 can be better understood by using a graph like the one shown in Fig. 3. The input layer (x_i) is made of a number of neurons (also known as perceptrons) equal to the number of input variables (d); the output layer, on the other hand, will have as many neurons as the output variables (k).

In the general case, the network may have an arbitrary number of hidden layers (in the depicted case there is just one hidden layer with three neurons), each of one can be formed by an arbitrary number of neurons (M). In a fully connected feed-forward network each node of a layer is connected to all the nodes in the adjacent layers. Each connection is represented by an adaptive weight $\left(w^l_{kj}\right)$ which can be regarded as the strength of the synaptic connection between neurons k and j, while the response of each perceptron to the inputs is represented by a non-linear function g, referred to as the *activation function.* All the above characteristics, the topology of the network and the weight matrix of its connections, define a specific implementation and are usually called *model.*

The model, however, is only part of the story. In fact, in order to find the model that best fits the data in a specific problem, one has to provide the network with a set of examples, *id est* of objects for which the final output is known by independent means. These data, already defined as *training set* or *Knowledge Base*, are used by the network to find the optimal model.

In our implementation we choose as learning rule the QNA, which differs from the Newton Algorithm in how the Hessian of the error function is computed. Newtonian models are variable metric methods used to find local maxima and minima of functions (Davidon 1968) and, in the case of MLPs, they can be used to find the stationary (i.e. the zero gradient) point of the learning function. The complete mathematical description of the MLP with QNA model is reported in the appendix A.

The model has been made available to the community through the DAta Mining & Exploration Web Application REsource (DAMEWARE[1]; Cavuoti et al. 2012b).

3.2. The implementation of MLPQNA

In this work we use our implementation of the QNA based on the limited-memory BFGS (L-BFGS; Byrd et al. 1994), where BFGS is the acronym composed of the names of the four inventors (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970).

Summarising, the algorithm for MLP with QNA is the following. Let us consider a generic MLP with $w^{(t)}$ being the weight vector at time (t).

[1] *http://dame.dsf.unina.it/beta_info.html*

1. Initialize all weights $w^{(0)}$ with small random values (typically normalized in $[-1, 1]$), set the constant error tolerance ε and $t = 0$;

2. present to the network all training set and calculate $E(w^{(t)})$ as the error function for the current weight configuration;

3. if $t = 0$ then $d^{(t)} = -\nabla E^{(t)}$

4. else $d^{(t)} = -\nabla E^{(t-1)} + Ap + B\nu$, where $p = w^{(t+1)} - w^{(t)}$ and $\nu = g^{(t+1)} - g^{(t)}$;

5. calculate $w^{(t+1)} = w^{(t)} - \alpha d^{(t)}$, where α is obtained by line search equation (see Eq. A6 in the Appendix);

6. calculate A and B for the next iteration, as reported in Eq. A19;

7. if $E(w^{(t+1)}) > \varepsilon$ then $t = t + 1$ and goto (ii), else STOP.

As it is known, all *line search* methods, being based on techniques which search for the minimum error by exploring the error function surface, are likely to get stuck in a local minimum and many solutions to this problem have been proposed (Floudas & Jongen 2005). In order to optimize the convergence of the Gradient Descent Analysis (GDA, see Appendix), Newton's method uses the information on the second-order derivatives. By having information on the second derivatives, QNA is more effective in avoiding local minima of the error function and more accurate in the error function trend follow-up, thus revealing a *natural* capability to find the absolute minimum error of the optimization problem (Shanno 1990).

In the L-BFGS version of the algorithm, in the case of high dimensionality (i.e. input data with many parameters), the amount of memory required to store the Hessian is too large, along with the machine time required to process it. Therefore, instead of using a complete number of gradient values to generate the Hessian, we can use a smaller number of values to approximate it.

By the way, if the convergence slows down, performances may even increase. A statement which only a first sight might seem paradoxical but, while the convergence is measured by the number of iterations, the performances depend on the number of processor's time units spent to calculate the result.

Related to the computational cost there is also the strategy adopted in terms of stopping criteria for the method. As it is known, the process of adjusting the weights based on the gradients is repeated until a minimum is reached. In practice, one has to decide the stopping condition of the algorithm. Among the possible criteria, the algorithm could be terminated when: (i) the Hessian approximation error becomes sufficiently small (by definition the gradient will be zero at a minimum); (ii) the maximum number of iterations is reached, in terms of a fixed threshold; (iii) based on the cross validation.

The cross validation can be used to monitor generalization performance during training and to terminate the algorithm when there is no more improvement. Statistically significant results come out by trying multiple independent data partitions and then averaging the performances. There are several variants of cross validation methods (Sylvain & Celisse 2010). We, in particular, have chosen the k-fold cross validation, particularly suited in presence of a scarcity of known data samples (Geisser 1975). The mechanism, also known as *leave-one-out*, is quite simple, since it consists in dividing the training set of N samples into k subsets ($k > 1$). The model is then trained on $k - 1$ subsets and validated by testing it on the left out subset. This procedure is then iterated leaving out each time a different subset for validation and its mean squared error is averaged on all cycles.

For what the MLP topology is concerned, a significant contribution came from the seminal paper by Bengio & LeCun (Bengio & LeCun 2007). They in fact re-analysed the implications of the Haykin pseudo-theorem (Haykin 1998), proving that complex problems, in which the mapping function is highly non linear and the local density of data in the parameter space is very variable, are better matched by *deep* networks with more than one hidden computational layer.

3.3. Statistical Indicators

In order to evaluate and reciprocally compare the experiments described in the next section we adopted the following definitions:

$$\mathrm{bias(x)} = \frac{\sum\limits_{i=1}^{N} x_i}{N} \tag{2}$$

$$\sigma(x) = \sqrt{\frac{\sum\limits_{i=1}^{N} \left[x_i - \left(\frac{\sum_{i=1}^{N} x_i}{N} \right) \right]^2}{N}} \tag{3}$$

$$\mathrm{MAD(x)} = Median\left(|x|\right) \tag{4}$$

$$\text{NMAD(x)} = 1.48 \times Median\left(|x|\right) \tag{5}$$

$$\text{RMS(x)} = \sqrt{\frac{\sum\limits_{i=1}^{N} x_i^2}{N}} \tag{6}$$

where σ is the standard deviation, MAD is the Median Absolute Deviation, NMAD the normalized MAD and RMS is the Root Mean Square. The term x in all above expressions may be either Δz defined as:

$$\Delta z = (z_{spec} - z_{phot}) \tag{7}$$

or the normalized residuals Δz_{norm} defined as:

$$\Delta z_{norm} = (z_{spec} - z_{phot})/(1 + z_{spec}) \tag{8}$$

4. The experiments

Our approach is based on machine learning methods and therefore, it needs to be as automatic as possible, in order to optimize the decisional support to the user (in this case the astronomer). Therefore, the results of the individual experiments as well as their comparison with others, have to be evaluated in a consistent and objective manner through an homogeneous set of statistical indicators.

In what follows we shall discuss the general experiment workflow and the outcome of the experiment phases.

4.1. The knowledge base and model setup

For machine learning supervised methods it is common practice to use the available KB to obtain at least three disjoint subsets for every experiment: one (training set) for training purposes, i.e. to train the method in order to acquire the hidden correlation among the input features, which is needed to perform the regression; the second one (validation set) to check the training, in particular against loss of generalization capabilities (a phenomenon also known as overfitting); and the third one (test set) to evaluate the overall performances of the model. As a rule of thumb in case of machine learning methods, these sets should be populated with respectively 60%, 20% and 20% of the objects in the KB (Kearns 1996). In our case, however, we reduced the training+validation data amount (from 80% to 60%), driven by the past experience with the very accurate regression capabilities of the model also in case of smaller knowledge bases (Brescia et al. 2012; Cavuoti et al. 2012a), obtaining implicitly the possibility to verify its prediction performance on a larger test set, as well as a faster execution of the training phase. Furthermore, in order to ensure a proper coverage of the KB in the Parameter Space (PS), the data objects were indeed divided up among the three datasets by random extraction, and usually this process is iterated several times to minimize the possible biases induced by fluctuations in the coverage of the PS, namely small differences in the redshift distribution of training and test samples used in the experiments.

The first two criteria used to decide the stopping condition of the algorithm, as mentioned at the end of Sec. 3.2, are mainly sensitive to the choice of specific parameters and may lead to poor results if the parameters are improperly set. The cross validation does not suffer of such drawback; it can avoid overfitting the data and is able to improve the generalization performance of the model. However, if compared to the traditional training procedures, the cross validation is much more computationally expensive. Therefore, by exploiting the cross validation technique (see Sec. 3.2), training and validation were indeed performed together using $\sim 60\%$ of the objects as a training + validation set, and the remaining $\sim 40\%$ as test set.

The automatized process of the cross-validation was done by performing ten different training runs with the following procedure: (i) splitting of the training/validation set into ten random subsets, each one composed by 10% of the dataset; (ii) at each training run we applied the 90% of the dataset and the excluded 10% for validation.

As remarked in Sec. 3.2, the k-fold cross validation is able to avoid overfitting on the training set (Bishop 2006), with an increase of the execution time estimable around $k - 1$ times the total number of runs (Cavuoti et al. 2012a).

In terms of the internal parameter setup of the MLPQNA, we need to consider the following topological parameters:

- input layer: a variable number of neurons, corresponding to the pruned number of survey parameters used in all experiments, up to a maximum number of 43 nodes (all available features);
- neurons on the first hidden layer: a variable number of hidden neurons, depending on the number N of input neurons (features in the dataset), equal to $2N + 1$ as rule of

thumb;

- neurons on the second hidden layer: a variable number of hidden neurons, ranging from 0 (to ignore the second layer) to $N - 1$;
- output layer: one neuron, returning the reconstructed photo-z value.

For the QNA learning rule, we heuristically fixed the following values as best parameters for the final experiments:

- step: 0.0001 (one of the two stopping criteria. The algorithm stops if the approximation error step size is less than this value. A step value equal to zero means to use the parameter MaxIt as the unique stopping criterion);
- res : 40 (number of restarts of Hessian approximation from random positions, performed at each iteration);
- dec : 0.1 (regularization factor for weight decay. The term $dec * ||networkweights||^2$ is added to the error function, where *networkweights* is the total number of weights in the network, directly depending on the total number of neurons inside. When properly chosen, the generalization performances of the network are highly improved);
- MaxIt: 8000 (max number of iterations of Hessian approximation. If zero the step parameter is used as stopping criterion);
- CV (k): 10 (k-fold cross validation, with $k = 10$);
- Error evaluation: Mean Square Error (between target and network output).

With these parameters, we obtained the statistical results reported in Sec. 4.4.

4.2. Selection of features

As it is known, supervised machine learning models are powerful methods for learning the hidden correlation between input and output features from training data. Of course, their generalization and prediction capabilities strongly depend on: the intrinsic quality of data (signal-to-noise ratio), the level of correlation among different features; the amount of missing data present in the dataset (Ripley 1996). It is obvious that some, possibly many, of the 43 parameters listed in Tab. 1 may not be independent and that their number needs to be reduced in order to speed up the computation (which scales badly with the number of features). This is a common problem in data mining and there is a wide literature on how to optimize the selection of features which are most relevant for a given purpose (Lindeberg 1998; Guyon & Elisseeff 2003, 2006; Brescia 2012). This process is usually called *Feature selection* or *pruning*, and consists in finding a compromise between the number of features (and therefore the computational time) and the required accuracy of the final results. In order to do so, we extracted from the main catalogue several subsets containing different groups of variables (features). Each one of these subsets was then analyzed separately in specific runs of the method (runs which in the data mining community are usually called experiments), in order to allow the comparison and evaluation. We wish to stress that our main concern was not only to disentangle which bands carry the most information but also, for a given band, which type of measurements (e.g. Point Spread Function, petrosian or isophotal magnitude) are more effective.

We performed a series of regression experiments to evaluate the performances obtained by the pruning of photometric quantities on the small dataset described in Sec. 2. The pruning experiments consisted into several combinations of surveys and their features:

- a *full* features experiment to be used as a benchmark for all the other experiments;
- some *service* experiments used to select the best combination of input features in order to eliminate redundancies in the flux measurements (i.e., petrosian magnitudes against isophotal magnitudes);
- *three-survey* experiments for all possible combinations of three (out of four) surveys;
- *two-survey* experiments with all possible combinations of two (out of four) surveys;
- *single-survey* experiments.

The output of the experiments consisted of lists of photometric redshift estimates for all objects in the KB. All pruning experiments were performed using ~ 3000 objects in the training set and ~ 800 in the test set. In Tab. 2, we list the outcome of the experiments for the feature selection. Both *bias* (Δz) and $\sigma\,(\Delta z)$ were computed using the objects in the test set alone. As it can be seen, among the various types of magnitudes available for GALEX and UKIDSS, the best combination is obtained using the isophotal magnitudes for GALEX and the calibrated magnitudes ($HallMag$) for UKIDSS.

Therefore at the end of the pruning phase the best combination of features turned out to be: the five SDSS $psfMag$, the two isophotal magnitudes of GALEX, the four $HallMag$

for UKIDSS and the four magnitudes for WISE.

4.3. Magnitudes vs Colors

Once the most significant features had been identified, we had to check which types of flux combinations were more effective, in terms of magnitudes or related colors. Experiments were performed on all five cross-matched datasets listed in section 2.

As it could be expected, the optimal combination turned out to be always the mixed one, i.e the one including colors and one reference magnitude for each of the included surveys (r for SDSS, nuv for GALEX, K for UKIDSS and W4 for WISE). From the data mining point of view this is rather surprising since the amount of information should not change by applying linear combinations between features. But from the physical point of view this can be easily understood by noticing that even though colors are derived as a subtraction of magnitudes, the content of information is quite different, since an ordering relationship is implicitly assumed, thus increasing the amount of information in the final output (gradients instead of fluxes). The additional reference magnitude instead removes the degeneracy in the luminosity class for a specific galaxy type.

4.4. MLPQNA Network Topology

The final check was about the hierarchical complexity of the network in terms of number of internal layers, whether *shallow* or *deep* according the definitions in Bengio & LeCun (2007), where *deep* is referred to a feed-forward network with more than one hidden layer. The above quoted cross-matched datasets were therefore processed through both a three-layers (input + hidden + output) and a four-layers (input + 2 hidden layers + output) network. In all cases the four-layers network performed significantly better, thus confirming the performance enhancement with *deep* networks in case of a particularly complex non-linear regression cases, i.e. in case of a highly multi-variate distributions of the input parameter space.

The experiments with best results have been obtained using a four-layers network, trained on the mixed (colors + reference magnitudes) datasets and their statistics are reported in tables 3, 4, 5 and 6.

5. Discussion and conclusions

In 2002 we begun to explore the usage of MLP's for the evaluation of photo-z both for *normal* galaxies and quasars (Tagliaferri et al. 2002). Several years later, D'Abrusco et al. 2007 used a combination of two MLP's to correct for the degeneracy introduced by the inhomogeneities in the knowledge base. Then Laurino et al. 2011 demonstrated that the subtleties in the mapping function could be more easily captured using the so-called *WGE (Weak Gated Experts)* method, a hierarchical combination of MLP's each specialized in a specific partition of the parameter space, whose individual outputs were combined by an additional MLP.

Furthermore, Bengio & LeCun 2007 published a seminal paper which somehow has disproved the Haykin-pseudo theorem (Haykin 1998), pointing out that problems with a large amount of distribution irregularities in the parameter space, are better treated by what they defined as *deep* networks, i.e. networks with more than one computational (hidden) layer. In this paper we exploited Bengio & LeCun 2007 findings, by using the supervised machine learning based method MLPQNA to evaluate photometric redshifts of quasars using multi-band data obtained from the cross-matching of the GALEX, SDSS, UKIDSS and WISE surveys.

In the tables 3, 4 and 5 we compare our best results to those presented by other authors (Ball et al. 2008; Richards et al. 2009; Laurino et al. 2011; Bovy et al. 2012), in terms of an homogeneous set of statistical indicators, defined in Sec. 3.3. Unfortunately, the whole set of indicators was not available for all bibliographical sources and in several cases we could only use a few quantities. Results are listed according to the combinations of surveys used in the experiment.

The best experiment, which makes use of a selected combination of parameters drawn from the four cross-matched surveys, leads to a $bias = 0.004$ and a Median Absolute Deviation $MAD = 0.02$. The fraction of catastrophic outliers, i.e. of objects with photo-z deviating more than 2σ from the spectroscopic value is $< 3\%$, leading to a $\sigma(\Delta z_{norm}) = 0.035$ after their removal (as reported in Tab. 6). The larger the number of surveys (bands) used, the more accurate are the results. This result, which might seem evident, is not obvious at all, since the higher amount of information carried by the additional bands implies also a strong decrease in the number of objects which are contained in the training set and should therefore cause a decrease in the generalization performances of the network.

This result, together with the fact that MLPQNA performs well also with small KB's (Cavuoti et al. 2012a), seems particularly interesting, since it has far reaching implications on ongoing and future surveys: a better photometric coverage is much more relevant than an increase of spectroscopic templates in the KB.

Concerning the performance evaluation in terms of photometric redshift reconstruction, all statistical results reported throughout this paper are referred to test data sets only. In fact, it is good practice to evaluate the results on data (i.e. the test set) which have never been presented to the network during the training and/or validation phases. The usage of *test plus training* data might introduce an obvious positive systematic bias which could mask reality.

More in general, empirical methods, such as MLPQNA, have the advantage that the training set is made up of real sky objects. Hence they do not suffer from the uncertainty of having accurate templates. In this sense any empirical method intrinsically includes effects such as the filter band-pass and flux calibrations. In fact, as deeply discussed by Collister & Lahav (2004), one of the main drawbacks of these methods is the difficulty in extrapolating to regions of the input parameter space that are not well sampled by the training data. Therefore the efficiency of empirical methods degrades for objects at fainter magnitudes than those included in the training set, as this would require an extrapolation capability on data having properties, such as redshift and photometry, not included in the learned sample. In fact, another strong requirement of such methods is that the training set must be large enough to cover properly the parameter space in terms of colors, magnitudes, object types and redshift. In this case the calibrations and corresponding uncertainties are well known and only limited extrapolations beyond the observed locus in color-magnitude space are required. In conclusion, under the conditions described above about the consistency of the training set, a realistic way to measure photometric uncertainties is to compare the photometric redshifts estimation with spectroscopic measures in the test samples.

As it can be seen in the tables 3, 4 and 5, in all cases MLPQNA obtains very relevant results. Only in the SDSS+GALEX case, the non-normalized quantities (i.e. those referred to the error $\Delta z = z_{spec} - z_{phot}$) show a substantial agreement between our results and those by Laurino et al. 2011. The better performances of MLPQNA in the normalized indicators (i.e. those referred to the error $\Delta z_{norm} = (z_{spec} - z_{phot})/(1 + z_{spec})$), is a consequence of the better performances of the MLPQNA method in terms of fraction of catastrophic outliers.

We wish to stress that both our four-layers MLPQNA and the *WGE* method discussed in Laurino et al. 2011 take advantage of a substantial improvement in complexity with respect to the traditional three-layers MLP networks used in the literature, and therefore deal better with the complexity of the multi-color parameter space. Average statistical indicators such as bias and standard deviation, however, provide only part of the information which allows to correctly evaluate the performances of a method and, for instance, they provide only very little evidence of the systematic trends which are observed as a sudden increase in the residuals spread over specific regions of the redshift space (Laurino et al. 2011). In the worst cases, these regions correspond to degeneracies in the parameter space and, as it could be expected, the relevance of such degeneracies decreases for increasing number of bands.

For what the analysis of the catastrophic outliers is concerned, according to Mobasher et al. (2007), the parameter $D_{95} \equiv \Delta_{95} / \left(1 + z_{phot}\right)$ enables the identification of outliers in photometric redshifts derived through SED fitting methods (usually evaluated through numerical simulations based on mock catalogues). In fact, in the hypothesis that the redshift error $\Delta z_{norm} = \left(z_{spec} - z_{phot}\right) / \left(1 + z_{spec}\right)$ is Gaussian, the catastrophic redshift error limit would be constrained by the width of the redshift probability distribution, corresponding to the 95% confidence interval, i.e. with $\Delta_{95} = 2\sigma\left(\Delta z_{norm}\right)$. In our case, however, photo-z are empirical, i.e. not based on any specific fitting model and it is preferable to use the standard deviation value $\sigma\left(\Delta z_{norm}\right)$ derived from the photometric cross matched samples, although it could overestimate the theoretical Gaussian σ, due to the residual spectroscopic uncertainty as well as to the method training error. Therefore, we consider as catastrophic outliers the objects with $\left|\Delta z_{norm}\right| > 2\sigma\left(\Delta z_{norm}\right)$. It is also important to notice that for empirical methods it is useful to analyze the correlation between the $NMAD\left(\Delta z_{norm}\right) = 1.48 \times median\left(\left|\Delta z_{norm}\right|\right)$ and the standard deviation $\sigma_{clean}(\Delta z_{norm})$ calculated on the data sample for which $\left|\Delta z_{norm}\right| \leq 2\sigma\left(\Delta z_{norm}\right)$. In fact, the quantity $NMAD$ would be comparable to the value of the σ_{clean}.

As it is shown in Tab. 6, in our data the $\sigma_{clean}(\Delta z_{norm})$ is always slightly larger than the corresponding $NMAD(\Delta z_{norm})$, which is exactly what is expected due to the overestimate induced by the above considerations (see also Fig. 4).

Finally, we would like to stress that the difficulties encountered by us and by other teams in comparing different methods, especially in light of the crucial role that photo-z play in the scientific exploitation of present and future large surveys (cf. The Dark Energy Survey Collaboration 2005, Chambers 2011, Refregier et al. 2010), confirm that it would be desirable to repropose an upgraded version of the extremely useful PHAT contest (Hildebrandt et al. 2010, Cavuoti et al. 2012a), which allowed a direct, effective and non ambiguous comparison of different methods applied on the same datasets and through the same set of statistical indicators. This new contest should be applied to a much larger dataset, with a more practical selection of photometric bands, and should take into account also other parameters such as scalability and robustness of the algorithms, as well as the degeneracy characterization.

Acknowledgments

The authors would like to thank the anonymous referee for the comments and suggestions which helped us to improve the paper. The authors wish to thank the whole DAMEWARE team[2], for the many useful discussions.

The authors also wish to thank the financial support of Project F.A.R.O., 3^{rd} call by the University Federico II of Naples, and of the PRIN-MIUR 2011 for Euclid Mission.

One of us (GL) wishes to thank Prof G.S. Djorgovski and the whole Department of Astronomy at the California Institute of Technology in Pasadena, for hospitality.

AM and MB wish to thank the financial support of PRIN-INAF 2010, *Architecture and Tomography of Galaxy Clusters.*

R. D'A. acknowledges the financial support of the US Virtual Astronomical Observatory, which is sponsored by the National Science Foundation and the National Aeronautics and Space Administration.

[2] *http://dame.dsf.unina.it/project_members.html*

Table 1. Summary of the data extracted from the databases of the four surveys and merged to form our final catalogue. Even though most names of the parameters are self explanatory, we wish to remind that the various *psfMag* are magnitudes derived by integrating fluxes over the best fitting point spread function. The aperture sizes refer to the radii.

Survey	Bands	Name of feature	Synthetic description
GALEX	nuv, fuv	mag, mag_iso	Near and Far UV total and isophotal mags
		mag_Aper_1 mag_Aper_2 mag_Aper_3	phot. through 3, 4.5 and 7.5 arcsec apertures
		mag_auto and kron_radius	magnitudes and Kron radius in units of A or B
SDSS	u, g, r, i, z	psfMag	PSF fitting magnitude in the u g, r, i, z bands.
UKIDSS	Y, J, H, K	PsfMag	PSF fitting magnitude in Y, J, H, K bands
		AperMag3, AperMag4, AperMag6	aperture photometry through 2, 2.8 & 5.7$''$ circular aperture in each band
		HallMag, PetroMag	Calibrated magnitude within circular aperture r_hall and Petrosian magnitude in Y, J, H, K bands

WISE	W1, W2, W3, W4	W1mpro, W2mpro, W3mpro, W4mpro	W1: 3.4 μm and 6.1$''$ angular resolution;
			W2: 4.6 μm and 6.4$''$ angular resolution;
			W3: 12 μm and 6.5$''$ angular resolution;
			W4: 22 μm and 12$''$ angular resolution.
			Magnitudes measured with profile-fitting photometry at the 95% level. Brightness upper limit if the flux measurement has SNR$<$ 2

SDSS	-	z_{spec}	Spectroscopic redshift

Table 2. Experiments for the feature selection phase. Col.s 1-4: surveys used for the experiment, where superscript index indicates the used magnitudes: ^{1} *mag*; ^{2} *mag_iso*; ^{3} *magnitudes through 3, 4.5 and 7.5 arcsec apertures*; ^{4} *mag_auto*; ^{5} *kron_radius*; ^{6} *HallMag*; ^{7} *PetroMag*. A cross in a column means that the survey corresponding to that column was used for the experiment.

GALEX	SDSS	UKIDSS	WISE	*bias* (Δz)	σ (Δz)
Service Experiments					
X	X	X	X	0.0033	0.174
$X^{1,2}$	X	X^{6}	X	-0.0001	0.152
X^{3}	X	X^{6}	X	-0.0016	0.165
X^{1}	X	X^{6}	X	0.0054	0.151
X^{2}	X	X^{6}	X	-0.0026	0.151
$X^{4,5}$	X	X^{6}	X	-0.0008	0.152
$X^{1,2,3}$	X	X^{6}	X	0.0041	0.163
$X^{2,3}$	X	X^{6}	X	-0.0033	0.155
		$X^{6,7}$		-0.0091	0.299
		X^{7}		0.0111	0.465
		X^{6}		-0.0081	0.294

Table 2. continue

GALEX	SDSS	UKIDSS	WISE	*bias* (Δz)	$\sigma\,(\Delta z)$
Four Survey Experiment					
X[2]	X	X[6]	X	-0.0026	0.151
Three Survey Experiment					
X[2]	X	X[6]		-0.0046	0.152
X[2]	X		X	0.0025	0.162
	X	X[6]	X	-0.0032	0.179
X[2]		X[6]	X	0.0110	0.203
Two Survey Experiment					
		X[6]	X	0.0045	0.236
X[2]			X	0.0175	0.288
	X	X[6]		-0.0027	0.210
	X		X	-0.0039	0.197
X[2]	X			-0.0055	0.240
X[2]		X[6]		0.0133	0.238
One Survey Experiment					
			X	0.0165	0.297
	X			-0.0162	0.338
X[1,2]				0.0550	0.419
		X[6]		-0.0081	0.294

[1] *mag*

[2] *mag_iso*

[3] *mag_Aper 1, 2 and 3*

[4] *mag_auto*

[5] *kron_radius*

[6] *HallMag*

[7] *PetroMag*

Table 3. Comparison among the performances of the different references. MLPQNA is related to our experiments, based on a four-layers network, trained on the mixed (colors + reference magnitudes) datasets. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; Column 2-5, respectively: bias, standard deviation, MAD, RMS, calculated on $\Delta z = (z_{spec} - z_{phot})$ related to the test sets. For the definition of the parameters and for discussion see text.

Exp	$BIAS(\Delta z)$	$\sigma(\Delta z)$	$MAD(\Delta z)$	$RMS(\Delta z)$
SDSS				
MLPQNA	0.007	0.25	0.102	0.26
Bovy et al.	-	0.46	-	-
Laurino et al.	0.210	0.28	0.110	0.35
Ball et al.	-	0.35	-	-
Richards et al.	-	0.52	-	-
SDSS + GALEX				
MLPQNA	0.003	0.21	0.060	0.22
Bovy et al.	-	0.26	-	-
Laurino et al.	0.13	0.21	0.061	0.25
Ball et al.	-	0.23	-	-
Richards et al.	-	0.37	-	-
SDSS + UKIDSS				
MLPQNA	0.001	0.25	0.066	0.26
Bovy et al.	-	0.28	-	-
SDSS + GALEX + UKIDSS				
MLPQNA	0.0009	0.18	0.043	0.19
Bovy et al.	-	0.21	-	-
SDSS + GALEX + UKIDSS + WISE				
MLPQNA	0.002	0.15	0.040	0.15

Table 4. Comparison among the performances of the different references. MLPQNA is related to our experiments, based on a four-layers network, trained on the mixed (colors + reference magnitudes) datasets. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; columns 2-6, respectively: bias, standard deviation, MAD, RMS and NMAD calculated on $\Delta z_{norm} = (z_{spec} - z_{phot}) \, / \, (1 + z_{spec})$ related to the test sets. For the definition of the parameters and for discussion see text.

Exp	$BIAS(\Delta z_{norm})$	$\sigma(\Delta z_{norm})$	$MAD(\Delta z_{norm})$	$RMS(\Delta z_{norm})$	$NMAD(\Delta z_{norm})$
SDSS					
MLPQNA	0.032	0.15	0.039	0.17	0.058
Laurino et al.	0.095	0.16	0.041	0.19	-
Ball et al.	0.095	0.18	-	-	-
Richards et al.	0.115	0.28	-	-	-
SDSS + GALEX					
MLPQNA	0.012	0.11	0.029	0.11	0.043
Laurino et al.	0.058	0.29	0.029	0.11	-
Ball et al.	0.06	0.12	-	-	-
Richards et al.	0.071	0.18	-	-	-
SDSS + UKIDSS					
MLPQNA	0.008	0.11	0.027	0.11	0.040
SDSS + GALEX + UKIDSS					
MLPQNA	0.005	0.087	0.022	0.088	0.032
SDSS + GALEX + UKIDSS + WISE					
MLPQNA	0.004	0.069	0.020	0.069	0.029

Table 5. Comparison in terms of outliers percentages among the different references. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; Column 2-3 are fractions of outliers at different σ based on $\Delta z = (z_{spec} - z_{phot})$; Column 4-5 are the fractions of outliers at different σ based on $\Delta z_{norm} = (z_{spec} - z_{phot}) \,/\, (1 + z_{spec})$. The column 4 reports our catastrophic outliers, defined as $|\Delta z_{norm}| > 2\sigma(\Delta z_{norm})$.

Exp	Outliers ($\|\Delta z\|$)		Outliers ($\|\Delta z_{norm}\|$)	
	$> 2\sigma(\Delta z)$	$> 4\sigma(\Delta z)$	$> 2\sigma(\Delta z_{norm})$	$> 4\sigma(\Delta z_{norm})$
	SDSS			
MLPQNA	7.68	0.38	6.53	1.24
Bovy et al.		0.51		
	SDSS + GALEX			
MLPQNA	4.88	1.61	4.57	1.37
Bovy et al.		1.86		
	SDSS + UKIDSS			
MLPQNA	4.00	1.73	3.82	1.38
Bovy et al.		1.92		
	SDSS + GALEX + UKIDSS			
MLPQNA	2.86	1.47	3.05	0.23
Bovy et al.		1.13		
	SDSS + GALEX + UKIDSS + WISE			
MLPQNA	2.57	0.87	2.88	0.91

Table 6. Catastrophic outliers evaluation and comparison between the residual $\sigma_{clean}(\Delta z_{norm})$ and $NMAD(\Delta z_{norm})$. The reported number of objects, for each cross-matched catalog, is referred to the test sets only. Catastrophic outliers are defined as objects where $|\Delta z_{norm}| > 2\sigma\,(\Delta z_{norm})$. The standard deviation $\sigma_{clean}(\Delta z_{norm})$ is calculated after having removed the catastrophic outliers, i.e. on the data sample for which $|\Delta z_{norm}| \leq 2\sigma\,(\Delta z_{norm})$

Exp	n. obj.	$\sigma\,(\Delta z_{norm})$	% catas. outliers	$\sigma_{clean}\,(\Delta z_{norm})$	$NMAD\,(\Delta z_{norm})$
SDSS	41431	0.15	6.53	0.062	0.058
SDSS + GALEX	17876	0.11	4.57	0.045	0.043
SDSS+UKIDSS	12438	0.11	3.82	0.041	0.040
SDSS+GALEX+UKIDSS	5836	0.087	3.05	0.040	0.032
SDSS+GALEX+UKIDSS+WISE	5716	0.069	2.88	0.035	0.029

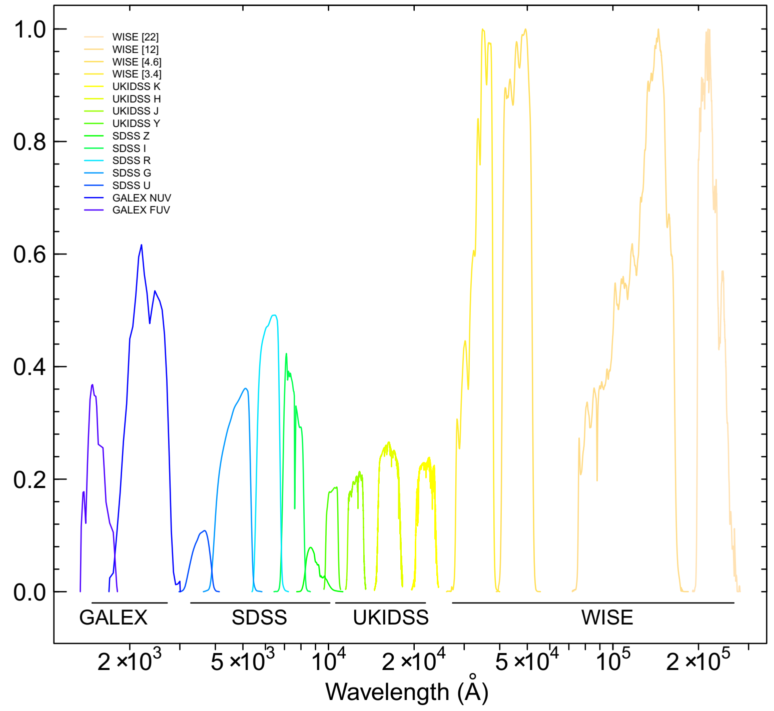

Fig. 1.— Transmission curves for all filters in the four surveys considered.

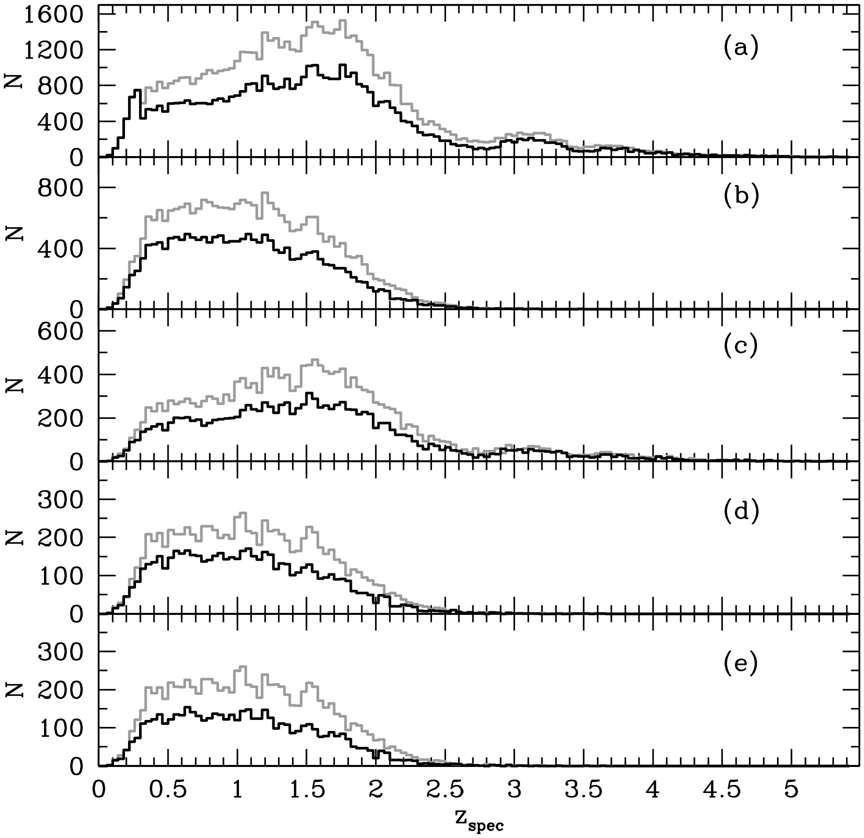

Fig. 2.— Histograms of spectroscopic redshift distribution in the five survey cross-matched samples as derived from the SDSS spectroscopic data. (a) SDSS; (b) SDSS+GALEX; (c) SDSS+UKIDSS; (d) SDSS+GALEX+UKIDSS; (e) SDSS+GALEX+UKIDSS+WISE. Gray dotted line is the training sample. Black line is the test sample.

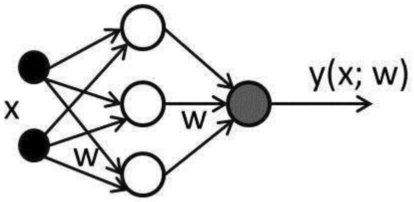

Fig. 3.— Scheme of a Multi Layer Perceptron general architecture for two input variables, one hidden layer with three neurons and one output value.

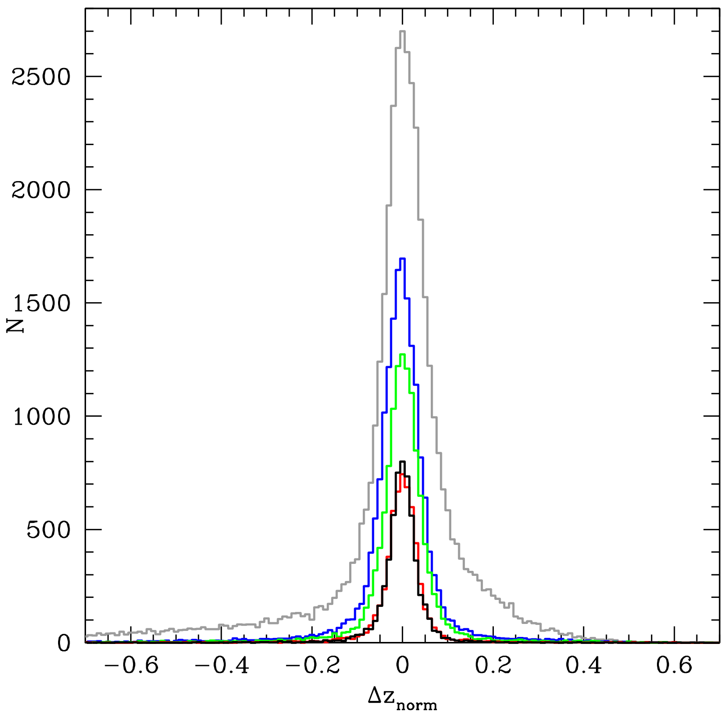

Fig. 4.— Δz_{norm} distributions for all five cross-matched test data sets. Lines are referred to, respectively, SDSS (gray), SDSS+GALEX (blue), SDSS+UKIDSS (green), SDSS+GALEX+UKIDSS (red) and SDSS+GALEX+UKIDSS+WISE (black).

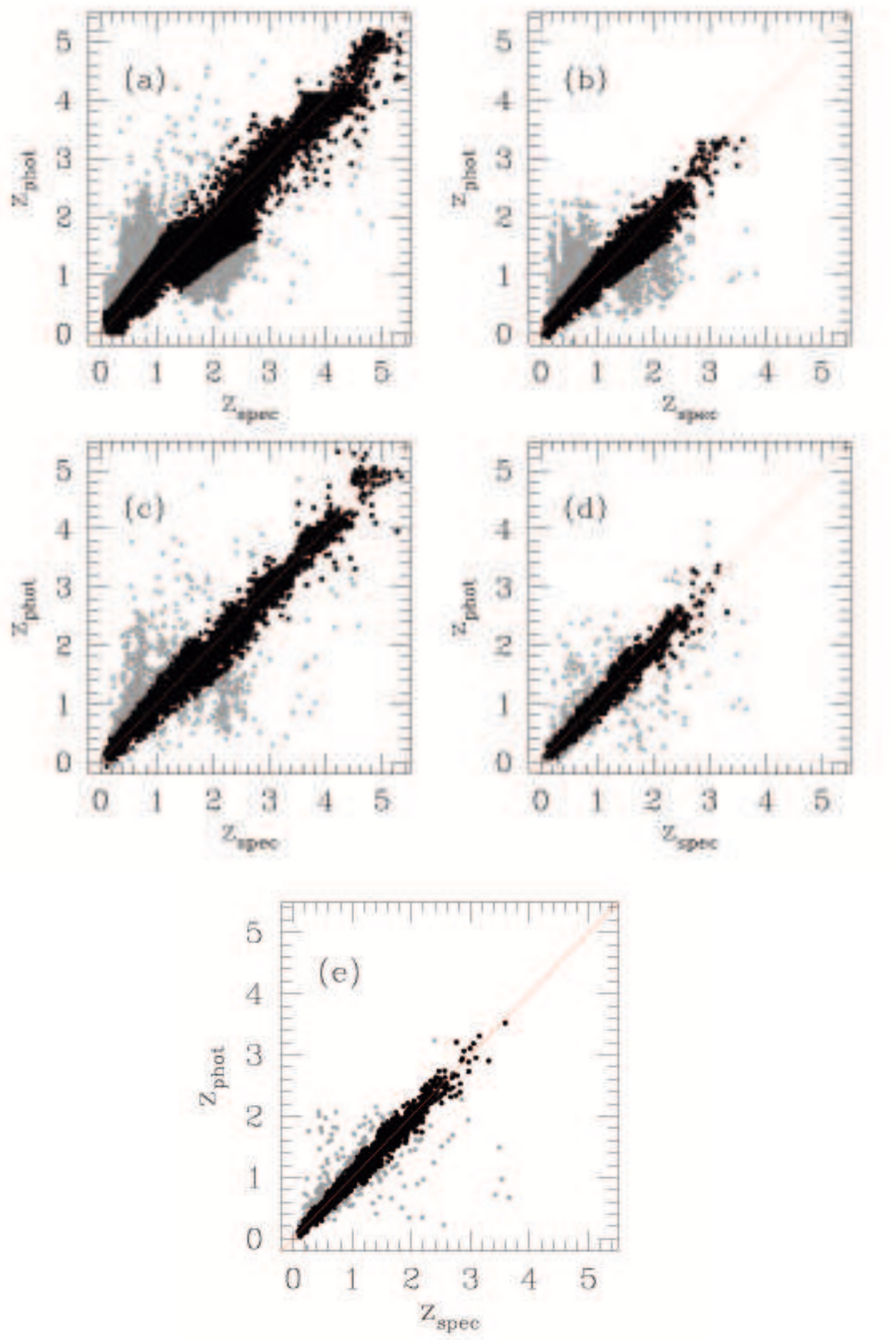

Fig. 5.— Scatter plots (z_{spec} vs z_{phot}); (a) SDSS, (b) SDSS+GALEX, (c) SDSS+UKIDSS, (d) SDSS+GALEX+UKIDSS and (e) SDSS+GALEX+UKIDSS+WISE. All diagrams refer to results on test sets. Gray points are catastrophic outliers (defined in Tab. 5). Red line is the dot-to-dot straight line passing through photometric and spectroscopic redshift limits in the available Knowledge Base.

A. Appendix: The Quasi Newton learning rule

Most Newton methods use the Hessian of the function to find the stationary point of a quadratic form. It needs to be stressed, however, that the Hessian of a function is not always available and in many cases it is far too complex to be computed in an analytical way. More often it is easier to compute the function gradient which can be used to approximate the Hessian via N consequent gradient calculations. In order to better understand why QNA are so powerful, it is convenient to start from the classical and quite common Gradient Descent Algorithm (GDA) used for Back Propagation (Bishop 2006). In GDA, the direction of each updating step for the MLP weights is derived from the error descent gradient, while the length of the step is determined from the learning rate. In case of particularly complex problems this method is inaccurate and ineffective and therefore may get stuck in local minima. A more effective approach is to move towards the negative direction of the gradient (*line search direction*) not by a fixed step, but by moving towards the minimum of the function along that direction. This can be achieved by first deriving the descent gradient and then by analyzing it with the variation of the learning rate (Brescia 2012). Let us suppose that at step t, the current weight vector is $w^{(t)}$, and let us consider a search direction $d^{(t)} = -\nabla E^{(t)}$. If we select the parameter λ in order to minimize $E(\lambda) = E(w^{(t)} + \lambda d^{(t)})$, the new weight vector can be expressed as:

$$w^{(t+1)} = w^{(t)} + \lambda d^{(t)} \tag{A1}$$

and the problem of *line search* becomes a 1-dimensional minimization problem which can be solved in many different ways. Simple variants are: i) to move $E(\lambda)$ by varying λ by small intervals, then evaluate the error function at each new position, and stop when the error begins to increase, or ii) to use the parabolic search for a minimum and compute the parabolic curve crossing pre-defined learning rate points. The minimum d of the parabolic curve is a good approximation of the minimum of $E(\lambda)$ and it can be derived by means of the parabolic curve which crosses the fixed points with the lowest error values.

Another approach makes instead use of *trust region* based strategies which minimize the error function, by iteratively growing or contracting the region of the function by adjusting a quadratic model function which best approximates the error function. In this sense this technique can be considered as a dual to line search, since it tries to find the best size of the region by fixing the step size (while the line search strategy always chooses the step direction before selecting the step size), (Celis et al. 1985). All these approaches, however, rely on the assumption that the optimal search direction is given at each step by the negative gradient: an assumption which not only is not always true, but can also lead to serious wrong convergence. In fact, if the minimization is done along the negative gradient direction, the subsequent search direction (the new gradient) will be orthogonal to the previous one: in fact, note that when the line search founds the minimum, it is:

$$\frac{\partial E}{\partial \lambda}(w^{(t)} + \lambda d^{(t)}) = 0 \tag{A2}$$

and hence,

$$g^{(t+1)T} d^{(t)} = 0 \tag{A3}$$

where $g \equiv \nabla E$. The iteration of the process therefore leads to oscillations of the error function which slow down the convergence process. The method implemented here relies on selecting other directions so that the gradient component, parallel to the previous search direction, would remain unchanged at each step. Suppose that you have already minimized with respect to the direction $d^{(t)}$ starting from the point $w^{(t)}$ and reaching the point $w^{(t+1)}$, where Eq. A3 becomes:

$$g(w^{(t+1)})^T d^{(t)} = 0 \tag{A4}$$

by choosing $d^{(t+1)}$ so to preserve the gradient component parallel to $d^{(t)}$ equal to zero, it is possible to build a sequence of directions d in such a way that each direction is conjugated to the previous one on the dimension $|w|$ of the search space (this is known as conjugate gradients method; Golub & Ye (1999)). In presence of a squared error function, the update weights algorithm is:

$$w^{(t+1)} = w^{(t)} + \alpha^{(t)} d^{(t)} \tag{A5}$$

with:

$$\alpha^{(t)} = -\frac{d^{(t)T} g^{(t)}}{d^{(t)T} H d^{(t)}} \tag{A6}$$

Furthermore, d can be obtained for the first time via the negative gradient and in the subsequent iterations, as a linear combination of the current gradient and of the previous search directions:

$$d^{(t+1)} = -g^{(t+1)} + \beta^{(t)} d^{(t)} \tag{A7}$$

with:

$$\beta^{(t)} = \frac{g^{(t+1)T} H d^{(t)}}{d^{(t)T} H d^{(t)}} \tag{A8}$$

This algorithm finds the minimum of a square error function at most in $|w|$ steps but at the price of a high computational cost, since in order to determine the values of α and β, it makes use of that *hessian matrix* H which, as we already mentioned, is very demanding in terms of computing time. A fact which puts serious constraints on the application of this family of methods to medium/large data sets. Excellent approximations for the coefficients α and β can, however, be obtained from analytical expressions that do not use the Hessian matrix explicitly. For instance, β can be calculated through any one of the following expressions (respectively Hestenes & Stiefel (1952); Fletcher & Reeves (1964); Polak & Ribiere (1969)):

$$Hestenes - Sitefel : \beta^{(t)} = \frac{g^{(t+1)T}(g^{(t+1)} - g^{(t)})}{d^{(t)T}(g^{(t+1)} - g^{(t)})} \tag{A9}$$

$$Fletcher - Reeves : \beta^{(t)} = \frac{g^{(t+1)T}g^{(t+1)}}{g^{(t)T}g^{(t)}} \tag{A10}$$

$$Polak - Ribiere : \beta^{(t)} = \frac{g^{(t+1)T}(g^{(t+1)} - g^{(t)})}{g^{(t)T}g^{(t)}} \tag{A11}$$

These expressions are all equivalent if the error function is square-typed, otherwise they assume different values. Typically the Polak-Ribiere equation obtains better results because, if the algorithm is slow and subsequent gradients are quite alike between them, its equation produces values of β such that the search direction tends to assume the negative gradient direction (Vetterling & Flannery 1992).

Concerning the parameter α, its value can be obtained by using the line search method directly. The method of conjugate gradients reduces the number of steps to minimize the error up to a maximum of $|w|$ because there could be almost $|w|$ conjugate directions in a $|w|$-dimensional space. In practice however, the algorithm is slower because, during the learning process, the property *conjugate* of the search directions tends to deteriorate. It is useful, to avoid the deterioration, to restart the algorithm after $|w|$ steps, by resetting the search direction with the negative gradient direction.

By using a local square approximation of the error function, we can obtain an expression for the minimum position. The gradient in every point w is in fact given by:

$$\nabla E = H \times (w - w^*) \tag{A12}$$

where w^* corresponds to the minimum of the error function, which satisfies the condition:

$$w^* = w - H^{-1} \times \nabla E \tag{A13}$$

The vector $-H^{-1} \times \nabla E$ is known as Newton direction and it is the base for a variety of optimization strategies, such as for instance the QNA, which instead of calculating the H matrix and then its inverse, uses a series of intermediate steps of lower computational cost to generate a sequence of matrices which are more and more accurate approximations of H^{-1}.

From the Newton formula (Eq. A13) we note that the weight vectors on steps t and $t+1$ are correlated to the correspondent gradients by the formula:

$$w^{(t+1)} - w^{(t)} = -H^{(-1)}(g^{(t+1)} - g^{(t)}) \tag{A14}$$

which is known as *Quasi Newton Condition*. The approximation G is therefore built in order to satisfy this condition. The formula for G is:

$$G^{(t+1)} = G^{(t)} + \frac{pp^T}{p^T\nu} - \frac{(G^{(t)}\nu)\nu^T G^{(t)}}{\nu^T G^{(t)}\nu} + (\nu^T G^{(t)}\nu)uu^T \tag{A15}$$

where the vectors are:

$$p = w^{(t+1)} - w^{(t)}; \nu = g^{(t+1)} - g^{(t)}; u = \frac{p}{p^T\nu} - \frac{G^{(t)}\nu}{\nu^T G^{(t)}\nu} \tag{A16}$$

Using the identity matrix to initialize the procedure is equivalent to consider, step by step, the direction of the negative gradient while, at each next step, the direction $-Gg$ is for sure a descent direction. The above expression could carry the search out of the interval of validity for the squared approximation. The solution is hence to use the *line search* to found the minimum of function along the search direction. By using such system, the weight updating expression (Eq. A5) can be formulated as follows:

$$w^{(t+1)} = w^{(t)} + \alpha^{(t)} G^{(T)} g^{(t)} \tag{A17}$$

where α is obtained by the *line search*.

One of the main advantages of QNA, compared with conjugate gradients, is that the *line search* does not require the calculation of α with a high precision, because it is not a critical parameter. Unfortunately, however, again, it requires a large amount of memory to calculate the matrix G ($|w| \times |w|$), for large $|w|$. One way to reduce the required memory is to replace at each step the matrix G with a unitary matrix. With such replacement and after multiplying by g (the current gradient), we obtain:

$$d^{(t+1)} = -g^{(t)} + Ap + B\nu \tag{A18}$$

Note that if the line search returns exact values, then the above equation produces mutually conjugate directions. A and B are scalar values defined as:

$$\begin{aligned} A &= -(1 + \tfrac{\nu^T\nu}{p^T\nu})\tfrac{p^T g^{(t+1)}}{p^T\nu} + \tfrac{\nu^T g^{(t+1)}}{p^T\nu} \\ B &= \tfrac{p^T g^{(t+1)}}{p^T\nu} \end{aligned} \tag{A19}$$

REFERENCES

Abazajian, K.N. et al. 2009, ApJS, 182, 2, 543-558

Aihara, H., et al., 2011, ApJS, 193, 29

Ball, N. M. et al. 2008, ApJ, 683, 1, 12-21

Baum, W. A., 1962, in Problems of Extra-Galactic Research, Proceedings from IAU Symposium no. 15, edited by McVittie, G. C., 390

Baum, E., and Wilczek, F., 1988, Supervised learning of probability distributions by neural networks. Neural Information Processing Systems, Anderson, D.Z. ed., American Institute of Physics, New York, pp. 52-61

Bengio, Y., & LeCun, J., 2007, in Large-Scale Kernel Machines. MIT Press.

Bishop, C. M., 2006, Pattern Recognition and Machine Learning. Springer ISBN 0-387-31073-8.

Bovy, J., et al.; Astrophysical Journal, 749, 41.

Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., Puzia, T., 2012a, Monthly Notices of the Royal Astronomical Society, Volume 421, Issue 2, pp. 1155-1165.

Brescia, M., 2012, New Trends in E-Science: Machine Learning and Knowledge Discovery in Databases. In Horizons in Computer Science Research, Thomas S. Clary (eds.), Series Horizons in Computer Science Vol. 7, Nova Science Publishers, ISBN: 978-1-61942-774-7.

Broyden, C. G. , 1970, Journ. of the Inst. of Math. and Its Appl., 6, 76

Budavari, T., et al., 2001, AJ, 122, 1163

Byrd, R.H., Nocedal, J., Schnabel, R.B., 1994, Mathematical Programming, 63, 4, pp. 129-156

Cavuoti, S., Brescia, M., Longo, G., Mercurio, A., 2012a, A&A, Vol. 546, A13, pp. 1-8 arXiv:1206.0876.

Cavuoti, S., Brescia, M., Longo, G., Garofalo, M., Nocella, A., 2012b, Science - Image in Action, World Scientific Publishing, pp. 241-247

Celis, M., Dennis, J. E., Tapia, R. A., 1985, Numerical Optimization, P. Boggs, R. Byrd and R. Schnabel (eds.), SIAM, Philadelphia USA, pp. 71-82.

Chambers, K.C., 2011, Status and Early Science from the PS1 Science Mission, American Astronomical Society. Bulletin of the American Astronomical Society, Vol. 43.

Collister, A.A., & Lahav, O., 2004, PASP, Vol. 116, Issue 818, 345-351

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

D'Abrusco, R., et al. 2009, MNRAS, 396, 1, pp. 223-262

D'Abrusco, R., et al. 2007, ApJ, 663, 2, pp. 752-764

Davidon, W.C., 1968, Comput. J. 10, 406

The Dark Energy Survey Collaboration, 2005, The Dark Energy Survey, White Paper submitted to the Dark Energy Task Force, 42 pages, arXiv:0510346

Fernandez-Soto, A., Lanzetta, K.M., Chen, H.W., Pascarelle, S.M., Yahata, N., 2001, ApJ Supplement Series, 2001, 135, pp. 41-61 (39)

Fletcher, R., 1970, Computer Journal 13: 317

Fletcher, R., Reeves, C. M., 1964, Function minimization by conjugate gradients. Comput. J. 7, 2, 149-154. MR 0187375

Floudas, C.A., & Jongen, H. T., 2005, Journal of Global Optimization, Vol. 32, Number 3, 409-415

Fu, Limin., 1994, Neural Networks in Computer Intelligence. E.M. Munson and L. Goldberg (eds.), McGraw-Hill NY

Geisser, S., 1975, Journal of the American Statistical Association, 70 (350), 320-328.

Giannantonio, T., et al., 2006, Phys. Rev. D, 74, 063520

Giannantonio, T., Scranton, R., Crittenden, R. G., Nichol, R. C., Boughn, S. P., Myers, D., & Richards, G. T., 2008, Phys. Rev. D, 77, 123520

Goldfarb, D., 1970, Mathematics of Computation, 24, 23

Golub, G. H., & Ye, Q., 1999, SIAM Journal of Scientific Computation, Vol. 21, pp. 1305-1320

Guyon, I., Elisseeff, A., 2003, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, Vol. 3, pp. 1157-1182

Guyon, I., Elisseeff, A., 2006, In Feature Extraction, Foundations and Applications, Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. A. Editors; Series: Studies in Fuzziness and Soft Computing, Springer, Vol. 207

Haykin, Simon, 1998. Neural Networks: A Comprehensive Foundation, Vol. 2, Prentice Hall

Hennawi, J. F., et al., 2006, AJ, 131, 1

Hestenes, M. R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bur. Standards 49 (1952), 6, 409-439. MR 0060307

Hildebrandt, H., et al., 2010, PHAT: PHoto-z Accuracy Testing, Astronomy and Astrophysics, Vol. 523, 21 pp.

Kearns, M., 1996, A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for Training-Test Split, Neural Information Processing 8, D.S. Touretzky, M,C. Mozer and M.E. Hasselmo (eds.), Morgan Kaufmann, pp. 183-189

Lanzetta, K.M., Yahil, A., Fernandez-Soto, A., 1998, Astronomical Journal, 116, pp. 1066-1073 (18)

Laurino, O., D'Abrusco, R., Longo, G., Riccio, G., 2011, MNRAS, 418, 4, pp. 2165-2195

Lawrence, A., et al., 2007, MNRAS, 379, 1599

Lindeberg, T., 1998, Feature detection with automatic scale selection, International Journal of Computer Vision 30 (2), pp 77-116

Marlin, B.M., 2008, Missing data problems in machine learning, Library and Archives:Canada.

Martin, D. C., et al., 2005, ApJ, 619, L1

Myers, A. D., Brunner, R. J., Richards, G. T., Nichol, R. C., Schneider, D. P., Vanden Berk, D. E., Scranton, R., Gray, A. G., & Brinkmann, Jon, 2006, ApJ, 638, 622

Mobasher, B., Capak, P., Scoville, N. Z., Dahlen, T., Salvato, M., Aussel, H., Thompson, D. J., Feldmann, R., Tasca, L., Le Fevre, O., Lilly, S., Carollo, C. M., Kartaltepe, J. S., McCracken, H., Mould, J., Renzini, A., Sanders, D. B., Shopbell, P. L., Taniguchi, Y., Ajiki, M., Shioya, Y., Contini, T., Giavalisco, M., Ilbert, O., Iovino, A., Le Brun, V., Mainieri, V., Mignoli, M., Scodeggio, M., 2007. The Astrophysical Journal Supplement Series, Volume 172, Issue 1, pp. 117-131

Myers, A. D., Brunner, R. J., Nichol, R. C., Richards, G. T., Schneider, D. P. & Bahcall, N. A., 2007a, ApJ, 658, 85

Myers, A. D., Brunner, R. J., Nichol, R. C., Richards, G. T., Schneider, D. P. & Bahcall, N. A., 2007b, ApJ, 658, 99

Polak, E., Ribiere, G., 1969, Note sur la convergence de methodes des directions conjugees. Revue Fr. Inf. Rech. Oper. 16-R1, 35-43. MR 0255025

Refregier, A., et al., 2010, Euclid Imaging Consortium Science Book, arXiv:1001.0061

Richards, G. T., et al., 2001a, AJ, 121, 2308

Richards, G. T., et al., 2001b, AJ, 122, 1151

Richards, G. T., et al., 2002, AJ, 123, 2945

Richards, G. T. et al. 2009, ApJS, 180, 67-83

Ripley, B.D., 1996, Pattern Recognition and Neural Networks, Cambridge University Press

Rubinstein, R.Y., Kroese, D.P., 2004, The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York NY

Schneider, D. P., Richards, G. T., Hall, P. B., Strauss, M. A., Anderson, S. F., Boroson, T. A., Ross, N. P., Shen, Y., Brandt, W. N., Fan, X., Inada, N., Jester, S., Knapp, G. R., Krawczyk, C. M., Thakar, A. R., Vanden Berk, D. E., Voges, W., Yanny, B., York, D. G., Bahcall, N. A., Bizyaev, D., Blanton, M. R., Brewington, H., Brinkmann, J., Eisenstein, D., Frieman, J. A., Fukugita, M., Gray, J., Gunn, J. E., Hibon, P., Ivezic, Z., Kent, S. M., Kron, R. G., Lee, M. G., Lupton, R. H., Malanushenko, E., Malanushenko, V., Oravetz, D., Pan, K., Pier, J. R., Price, T. N., Saxe, D. H., Schlegel, D. J., Simmons, A., Snedden, S. A., SubbaRao, M. U., Szalay, A. S., Weinberg, D. H., Scranton, R., et al., 2010, AJ, Vol. 139, Issue 6, article id. 2360

Scranton, R., et al., 2005, ApJ, 633, 589

Shanno, D. F., 1990, Recent Advances in Numerical Techniques for large-scale optimization, Neural Networks for Control, MIT Press, Cambridge MA.

Shanno, D. F., 1970, Math. Comput. 24: 647

Sylvain, A., & Celisse, A., 2010, A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79. doi: 10.1214/09-SS054.

R. Tagliaferri, G. Longo, S. Andreon, S. Capozziello, C. Donalek, and G. Giordano, 2002, Neural networks and photometric redshifts, astro-ph/0203445.

Vetterling, T., Flannery, B. P., 1992, Conjugate Gradients Methods in Multidimensions. Numerical Recipes in C - The Art of Scientific Computing, W. H. Press and S. A. Teukolsky (eds.), Cambridge University Press; 2nd edition.

Veron-Cetty, M.-P., Veron, P., 2000, A Catalogue of quasars and active nucleai. ESO Scientific Report, n. 19.

Wright, E. L., et al., 2010, AJ, 140, 1868

Wolf, C., et al., 2004, A&A, 421, 913

This preprint was prepared with the AAS LaTeX macros v5.2.