MNRAS **000**, 1–17 (2016) Preprint 25 October 2016 Compiled using MNRAS LATEX style file v3.0

Evaluation of data compression techniques for the inference of stellar atmospheric parameters from high resolution spectra

A. González-Marcos ,[1★] L. M. Sarro,[2] J. Ordieres-Meré,[3] and A. Bello-García[4]
[1] *Department of Mechanical Engineering, University of La Rioja, c/ San José de Calasanz, 31, 26004 Logroño, Spain*
[2] *Dpto. de Inteligencia Artificial, ETSI Informática, UNED, c/ Juan del Rosal, 16, 28040 Madrid, Spain*
[3] *PMQ Research Team; ETSII; Universidad Politécnica de Madrid, José Gutiérrez Abascal 2, 28016 Madrid, Spain*
[4] *Dept. of Construction and Industrial Manufacturing, University of Oviedo, 33203 Gijón, Spain*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
The determination of stellar atmospheric parameters from spectra suffers the so-called curse of dimensionality problem, which is related to the higher number of input variables (measured fluxes) compared to the number of spectra available to fit a regression model. This work evaluates the utility of several techniques for alleviating this problem in regression tasks where the objective is to estimate the effective temperature, the surface gravity, the metallicity and/or the alpha to iron ratio. The goal of the techniques analyzed here is to achieve data compression by representing the spectra with a number of variables much lower than the initially measured set of fluxes. The experiments were performed with high resolution spectra in the 4000-8000 K range for different signal-to-noise ratio (SNR) regimes. We conclude that Independent Component Analysis (ICA) performs better than the rest of techniques evaluated for all SNR regimes. We also assess the necessity to adapt the SNR of the spectra used to fit a regression model (training set) to the SNR of the spectra for which the atmospheric parameters are needed (evaluation set). Within the conditions of our experiments, we conclude that at most only two such models are needed (in the case of regression models for effective temperatures, those corresponding to SNR=50 and 10) to cover the entire SNR range. Finally, we also compare the prediction accuracy of effective temperature regression models for increasing values of the training grid density and the same compression techniques.

Key words: methods: statistical – methods: data analysis – stars: fundamental parameters – methods: data compression

1 INTRODUCTION

The rapid evolution of astronomical instrumentation and the implementation of extensive surveys have permitted the acquisition of vast amounts of spectral data. The reduction and management of large spectral databases collected by large-area or all-sky surveys like Gaia/Gaia-ESO (Jordi et al. 2006; Gilmore et al. 2012), RAVE (Steinmetz et al. 2006), or APOGEE (Eisenstein et al. 2011) require the use of automatic techniques for the consistent, homogeneous, and efficient extraction of physical parameters from spectra. The availability of these huge databases opens new possibilities to better understand the stellar, Galactic, and extra-galactic astrophysics. Of special importance is the determination of intrinsic stellar physical parameters, such as effective temperature ($T_{\rm eff}$), surface gravity (log g), metallicity ([M/H]) and alpha to iron ratio ([α/Fe]). However, the difficulty that atmospheric parameter estimation poses comes from the inherent size and dimensionality of the data.

In the following, we will refer to the spectra used to infer the aforementioned physical parameters as data. Each spectrum is a high-dimensional array that containes the fluxes measured by a spectrograph or simulated by a spectra synthesis code.

★ Contact e-mail: ana.gonzalez@unirioja.es

We will assume that all spectra in a given regression application are measured at the same wavelengths. In general, the number of fluxes measured in a spectrum is very large, from several hundreds to thousands of measurements. Since these are the values we will use to infer the physical parameters, we will refer to them as predictive variables or simply variables. Hence, the input data used to infer the physical parameter of a given star is very high-dimensional. We can think of the regression process as a module that takes as input the observed or simulated spectrum and produces as output an estimate of the stellar physical parameters. The input space is the space of potential spectra (a space of high dimensionality) where each spectrum is specified by the input variables (the measured fluxes). We will sometimes refer to individual spectra as instances, cases or examples in this input space.

In order to construct regression models that infer physical parameters from input spectra, we will need a collection of examples with well known physical parameters. This collection of examples if known as the training set. The training set, in our work, will be a collection of stellar spectra with attached physical parameters.

In general, the number of measured fluxes in a single spectrum (the dimensionality of the input space) is larger than or comparable to the number of example spectra in the training data set. Thus, regression from stellar spectra suffers the so-called *curse of dimensionality*.

The *curse of dimensionality* (Bellman 1961) relates to the problem caused by the exponential increase in volume associated with adding extra dimensions to the input space of predictive variables. **When the number of examples in the training set is finite and fixed, the density of data instances (examples) decreases exponentially as the dimensionality of the input space increases. A classical example often described to illustrate this problem would consist (if translated to the domain of this work) in predicting the physical parameters of a given star by averaging the physical parameters of the most similar spectra (nearest neighbour) in the set of training examples. Let us assume, just for the sake of clarity, that our predictive variables are rescaled between 0 and 1. If we only used two fluxes, we would only need 121 spectra distributed uniformly in the 2D plane, to ensure that the nearest neighbour is at an expected Euclidean distance of** $\sqrt{0.05^2 + 0.05^2} = 0.07$**. If our spectra consist of 3 flux measurements, then the same 121 example spectra would only ensure an average minimum distance of 0.17 if (again) distributed uniformly in the unit cube. In 10 dimensions, the average minimum distance would be 1.76 (recall that we have assumed the predictive variables, that is to say, the measured fluxes) to be scaled between 0 and 1. This distance is more than half the maximum distance in the unit ten-dimensional cube. And we would need 33761 million examples to recover the minimum distance of 0.14. In other words, the nearest neighbour is further and further away as the dimensionality of the input space increases:** the available data become sparse. Because this sparsity is problematic for any method that requires statistical significance, the amount of data instances needed to support the result often grows exponentially with the dimensionality in order to obtain a statistically sound and reliable outcome.

Furthermore, typical spectra obtained in many surveys do not regularly reach the high signal-to-noise ratios (SNR) –about 100 or greater – needed to obtain robust estimates, which increases the difficulty to accurately estimate the physical parameters of spectra. In summary, stellar spectra are high dimensional noisy vectors of real numbers and thus, regression models must be both computationally efficient and robust to noise.

There are several ways to alleviate this so-called *curse of dimensionality*. It is evident that not all wavelength bins in an observed spectrum carry the same amount of information about the physical parameters of the stellar atmosphere. One way to reduce the dimensionality of the space of predictive variables is to concentrate on certain wavelength ranges that contain spectral lines that are sensitive to changes in the physical parameters. Before **large-scale spectroscopic surveys and the fast computers needed to analyse them became available**, astronomers derived physical parameters by interactively synthesizing spectra until a subjective best fit of the observed spectrum in certain spectral lines was found. But the number of spectra made available to the community in the past decades have made this manual and subjective (thus irreproducible) fitting procedure impractical. Automatic regression techniques have therefore become a necessity.

The next step consisted in using derived features of the spectrum such as fluxes, flux ratios or equivalent widths to infer the parameters via multivariate regression techniques (see Allende Prieto et al. 2006; Rojas-Ayala et al. 2010, 2012; Bruntt et al. 2010; or Mishenina et al. 2006). That way, we significantly reduce the full spectrum to a much smaller number of predictive variables, at the expense of introducing a feature extraction process: defining a continuum level and normalizing the observed spectrum in the wavelength region that contains the sensitive spectral feature. **The normalization process effectively consists in dividing two random variables: the observed flux and the estimated continuum level. The simplest hypothesis consists in assuming that both quantities are Gaussian distributed. Under these conditions, the normalized spectrum will be a collection of random variables (one per wavelength) each one distributed according to the ratio distribution (see e.g. Geary 1930; Marsaglia 1965). E**ven in the best case that the continuum flux is Gaussian distributed around a value significantly different from zero, the ratio distribution is asymmetric **(thus systematically biasing the result)** and has a heavy right tail (**meaning that values significantly larger than the mode of the distribution can occur with non-negligible probabilities**). In the cases of low signal-to-noise spectra, the situation can be catastrophic.

The potential dangers associated with the feature extraction in restricted wavelength ranges via continuum normalisation can be mitigated by projecting the observed spectra onto bases of functions spaces such as in the wavelet or

Fourier decompositions (see Manteiga et al. 2010; Lu & Li 2015; or Li et al. 2015, for examples of the two approaches). **In essence, the goal is to change the representation of the spectra, originally involving a large number of variables (measured fluxes), into a low-dimensional description using only a small number of variables (dimensions). The new representation should preserve essentially all of the useful information within the high dimensional space. Thus, by retaining only the most significant variables (dimensions) to represent the spectra, we achieve a data compression that can be of great benefit for estimating atmospheric parameters as it reduces the dimensionality of the space required to describe the data.**

The most popular data compression technique applied to stellar spectra is Principal Component Analysis (PCA). It has been widely applied in spectral classification combined with artificial neural networks (ANN) (Singh et al. 1998) or support vector machines (SVM) (Re Fiorentin et al. 2008a). For continuum emission, PCA has a proven record in representing the variation in the spectral properties of galaxies. However, it does not perform well when reconstructing high-frequency structure within a spectrum (Vanderplas & Connolly 2009). To overcome this difficulty, other methods have been used in the spectral feature extraction procedure. Locally linear embedding (LLE) (Roweis & Saul 2000) and Isometric feature map (Isomap) (Tenenbaum et al. 2000) are two widely used nonlinear data compression techniques. Some studies found that LLE is efficient in classifying galaxy spectra (Vanderplas & Connolly 2009) and stellar spectra (Daniel et al. 2011). Other authors concluded that Isomap performs better than PCA, except on spectra with low SNR (between 5 and 10) (Bu et al. 2014).

A detailed study of data compression techniques has to include the analysis of their stability properties against noise. In order to improve the overall generalisation performance of the atmospheric parameters estimators, experience shows that it is advantageous to match the noise properties of the synthetic training **example** to that of the real **observation** because it acts as a regulariser in the training phase (Re Fiorentin et al. 2008b). The impact of the SNR on the parameter estimation ($T_{\rm eff}$, log g and [Fe/H]) with ANN is explored in Snider et al. (2001). They found that reasonably accurate estimates can be obtained when networks are trained with spectra –not derived parameters– with similar SNR as those of the unlabelled data, for **SNR** as low as 13.

Recio-Blanco et al. (2006) determined three atmospheric parameters ($T_{\rm eff}$, log g and [M/H]) and individual chemical abundances from stellar spectra using the MATISSE (MATrix Inversion for Spectral SynthEsis) algorithm. They introduced Gaussian white noise to yield five values of SNR between 25 and 200 and found that errors increased considerably for SNR lower than ∼ 25. In Navarro et al. (2012) authors present a system based on ANN trained with a set of line-strength indices selected among the spectral lines more sensitive to temperature and the best luminosity tracers. They generated spectra with a range of SNR between 6 and 200 by adding Poissonian noise to each spectrum. Their scheme **allows classification of** spectra of SNR as low as 20 with an accuracy better than two spectral subtypes. For SNR ∼ 10, classification is still possible but at a lower precision.

In recent years, there seems to be a tendency to use **the spectrum rather than fluxes or equivalent widths derived from it** (see e.g. Torres et al. 2012; Recio-Blanco et al. 2014 and references therein; Ness et al. 2015; Walker et al. 2015; or Recio-Blanco et al. 2015). In this work we focus in this latter approach, and attempt to assess the relative merits of various techniques to serve as a guide for future applications of machine learning techniques for regression of stellar atmospheric physical parameters.

This paper presents a comparative study of the most popular data compression technique applied to stellar spectra (PCA) and five alternatives (two linear and three nonlinear techniques). The aims of the paper are (1) to investigate to what extent novel data compression techniques outperform the traditional PCA on stellar spectra datasets, (2) to test the robustness of these techniques and their performance in atmospheric parameters estimation for different SNRs, (3) to investigate the number of regression models of different SNRs needed to obtain the best generalisation performance for any reasonable SNR of the evaluation dataset, and (4) to analyse the effect of the grid density over the regression performance in atmospheric parameters estimation. The investigation is performed by an empirical evaluation of the selected techniques on specifically designed synthetic datasets. In Sect. 2 we describe the datasets used in our experiments. In Sect. 3 we review the data compression techniques evaluated in this work and their properties. Sect. 4 presents our results when comparing the compression techniques and compression rates in terms of the atmospheric parameter estimation errors. In Sect. 5 we evaluate the optimal match between the SNR of the training set examples to the SNR of the evaluation set, and in Sect. 6 we present the main results from the analysis of the effect of the training set grid density over the regression performance. Finally, in Sect. 7 we summarize the most relevant findings from the experiments and discuss their validity and limitations.

2 THE DATASET

The full set of spectra used in our experiments was divided into two groups:

- **The training set, which refers to the subset of spectra used to fit the regression models (the so-called training phase).**
- **The evaluation set, which refers to the subset of spectra not used for training, and used only to assess the performance of a given model when applied to previously unseen spectra.**

The synthetic spectra that form the basis of our study have been computed from MARCS model atmospheres (Gustafsson et al. 2008) and the turbospectrum code (Alvarez & Plez 1998; Plez 2012) together with atomic & molecular line lists. These spectra were kindly provided by the Gaia-ESO team in charge of producing the physical parameters for the survey (see de Laverny et al. 2012, for further details). **More specifically, our analyses were performed using spectra simulated with two different setups from the high-resolution (HR) mode of the GIRAFFE spectrograph, which was used to carry**

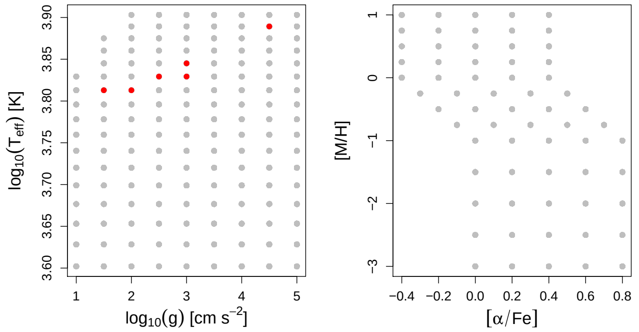

Figure 1. Coverage in parameter space of the dataset. Gray circles represent spectra available in the original collection provided by the Gaia-ESO collaboration. Red circles correspond to missing spectra that were linearly interpolated as described in the text.

out the observations of the survey: the HR10 setup (534-562 nm) and the HR21 setup (848-900 nm).

The **complete dataset (including training and evaluation data)** contains a grid of 8780 synthetic high-resolution spectra (R = 19800) between 5339 and 5619 Å (the nominal GIRAFFE HR10 setup) with effective temperatures between 4000 and 8000 K (step 250 K), logarithmic surface gravities between 1.0 and 5.0 (step 0.5), mean metallicities between -3.0 and 1.0 (with a variable step of 0.5 or 0.25 dex) and $[\alpha/Fe]$ values varying between -0.4 and +0.4 dex (step 0.2 dex) around the standard relation with the following α enhancements: $[\alpha/Fe]$ = +0.0 dex for [M/H] $\geqslant$ 0, $[\alpha/Fe]$ = +0.4 dex for [M/H] = $\leqslant$ -1.0 and $[\alpha/Fe]$ = -0.4[M/H] for [M/H] between -1.0 and +0.0 (Fig. 1). Elements considered to be α-elements are O, Ne, Mg, Si, S, Ar, Ca and Ti. The adopted solar abundances are those used by (Gustafsson et al. 2008). Fig. 2 (left) shows some example spectra from this dataset.

The sample size of our dataset (8780 spectra) is relatively small compared to the input dimension (2798 flux measurements per spectrum). With the rule of thumb of a minimum of 10 samples per dimension (Jain et al. 2000), our dataset should contain at least 10^{2798} spectra. In most real case applications in astronomy, the ratio of sample size to input dimensions is much lower and thus, the *curse of dimensionality* problem is expected to affect even more severely the inference process.

With a view to analyze the dependence of the validity of the results obtained with the selected dataset, we used a second dataset which is based on the same grid of atmospheric parameters but covering a different wavelength range. This new dataset contains high-resolution spectra (R = 16200) between 8484 and 9001 Å (the nominal GIRAFFE HR21 setup). Fig. 2 (right) shows some example spectra from this dataset. In this validity analysis, efforts were focused on the effective temperature.

3 DATA COMPRESSION

In a dynamic environment, a complete rerun of a data compression algorithm becomes prohibitively time and memory consuming. For the sake of computational efficiency, the selection of the data compression techniques tested in our experiments was done amongst those capable of projecting new data onto the reduced dimensional space defined by the training set without having to re-apply the algorithm (process also known as out-of-sample extension). Thus, in this work, we investigated three linear data compression techniques such as PCA, independent component analysis (ICA) and discriminative locality alignment (DLA), as well as three nonlinear reduction techniques that **can be generalised** to new data: wavelets, Kernel PCA and diffusion maps (DM). We aimed at minimizing the regression error in estimating stellar atmospheric parameters with no consideration of the physicality of the compression coefficients. Physicality of the coefficients is sometimes required, for example, when trying to interpret galactic spectra as a combination of non-negative components**, which closely resembles the physical process of emission in the mid-infrared**.

Other linear and nonlinear techniques could be used for data compression, such as linear discriminant analysis (LDA), locally linear embedding (LLE), Isomap, etc. When the number of variables is much higher than that of training samples, classical LDA cannot be directly applied because all scatter matrices are singular and this method requires the non-singularity of the scatter matrices involved. Isomap's performance exceeds the performance of LLE, specially when the data is sparse. However, in presence of noise or when the data is sparsely sampled, short-circuit edges pose a threat to both Isomaps and LLE algorithms (Saxena et al. 2004). Short-circuit edges can lead to low-dimensional embeddings that do not preserve a manifold's true topology (Balasubramanian et al. 2002). Furthermore, Isomap and LLE cannot be extended out-of-sample.

3.1 Principal Component Analysis (PCA)

PCA (Hotelling 1933; Pearson 1901) is by far the most popular linear technique for data compression. The aim of the method is to reduce the dimensionality of multivariate data whilst preserving as much of the relevant information (assumed to be related to the variance in the data) as possible. This is done by finding a linear basis of reduced dimensionality for the data, in which the amount of variance in the data is maximal. It is important to remark that PCA is based on the assumption that variance is tantamount to relevance for the regression task.

PCA transforms the original set of variables into a new set of uncorrelated variables, the principal components, which are linear combinations of the original variables. The new uncorrelated variables are sorted in decreasing order of variance explained. The first new variable shows the maximum amount of variance; the second new variable contains the maximum amount of variation unexplained by the first one, and is orthogonal to it, and so on. This is achieved by computing the covariance matrix for the full data set. Next, the eigenvectors and eigenvalues of the covariance matrix are computed, and sorted according to decreasing eigenvalue.

3.2 Independent Component Analysis (ICA)

ICA (Comon 1994) is very closely related to the method called blind source separation (BSS) or blind signal separa-

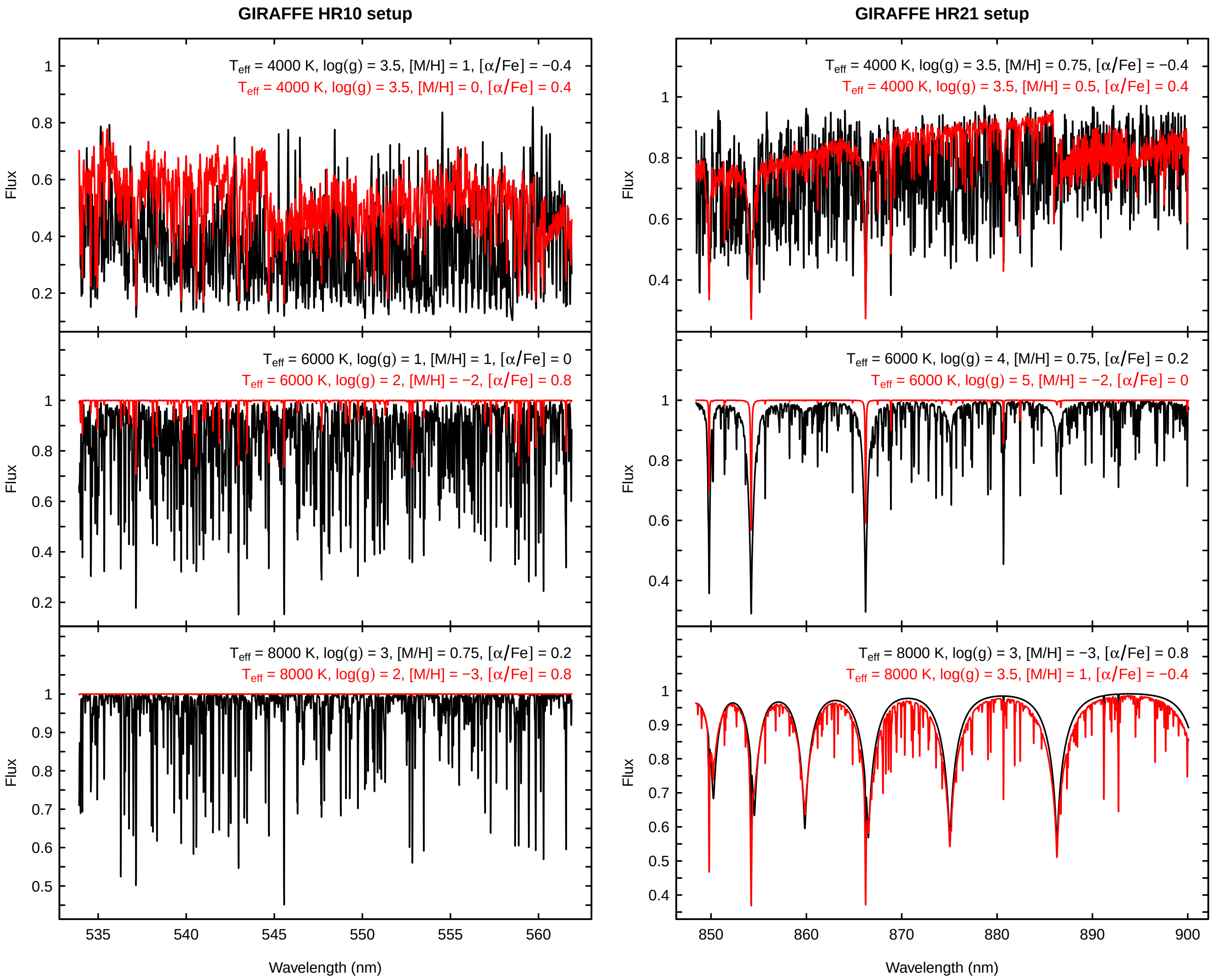

Figure 2. Example spectra from the nominal GIRAFFE HR10 setup (left) and the nominal GIRAFFE HR21 setup (right).

tion (Jutten & Hérault 1991). It is the identification and separation of mixtures of sources with little prior information. The goal of the method is to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible (Hyvärinen & Oja 2000).

Several algorithms have been developed for performing ICA (Bell & Sejnowski 1995; Belouchrani et al. 1997; Ollila & Koivunen 2006; Li & Adali 2008). A large widely used one is the FastICA algorithm (Hyvärinen & Oja 2000) which has a number of desirable properties, including fast convergence, global convergence for kurtosis-based contrasts, and the lack of any step size parameter. RobustICA (Zarzoso & Comon 2010) represents a simple modification of FastICA, and is based on the normalised kurtosis contrast function, which is optimised by a computationally efficient iterative technique. It is more robust than FastICA and has a very high convergence speed. Another widely used ICA algorithm is the Joint Approximation Diagonalisation of Eigen-matrices (JADE) (Cardoso & Souloumiac 1993). This approach exploits the fourth-order moments in order to separate the source signals from mixed signals. In this work we selected the JADE algorithm for projecting the original spectra in the space of independent components.

3.3 Discriminative Locality Alignment (DLA)

DLA (Zhang et al. 2008) is a supervised manifold learning algorithm **that performs data compression by utilizing the class label information of the data instances. In our case, we are not faced with a classification task and therefore, our examples in the training set do not have classes attached to them. In order to test the potential of this technique, we use the value of the atmospheric parameters ($T_{\rm eff}$, log g, [M/H], or $[\alpha/Fe]$) as class labels. The training set examples are spectra synthesized for a limited set of values of the physical parameters (see Sect. 2 and Fig. 1 for an illustration of the set of values used in training). It is this set of allowed values that we use as**

class label. We are aware of the gross simplification of discretizing the full range of allowed physical parameters (an interval of real numbers) into a limited subset of values.

The learning algorithm can be divided into three stages: part optimisation, sample weighting and whole alignment. In the first stage, **we define a patch $\mathcal{P}_i$ for each spectrum $\mathcal{S}_i$ as the set that includes $\mathcal{S}_i$ and its k-nearest neighbours. The set of k-nearest neighbours, in turn, is defined as the set of k spectra with minimum distances from $\mathcal{S}_i$. In our case, we use Euclidean distances.** On each patch $\mathcal{P}_i$, DLA preserves the local discriminative information through integrating the two criteria that *i*) the distances between intra-class spectra are as small as possible and *ii*) the distance between the inter-class spectra is as large as possible. In the second stage, each part optimisation is weighted by the *margin degree*, a measure of the importance of a given spectrum for classification. Finally, DLA integrates all the weighted part optimisations to form a global subspace structure through an alignment operation (Zhang & Zha 2002). The projection matrix can be obtained by solving a standard eigendecomposition problem.

DLA requires the selection of the following two parameters:

- **Cardinality of the neighbourhood in the same class (k_1): the number of nearest neighbour spectra in the same class as $\mathcal{S}_i$**
- **Cardinality of the neighbourhood in different classes (k_2): the number of nearest neighbour spectra in classes other than the class of $\mathcal{S}_i$**

This method obtains robust classification performance under the condition of small sample size. Furthermore, it does not need to compute the inverse of a matrix, and thus it does not face the matrix singularity problem that makes LDA and quadratic discriminant analysis (QDA) not directly applicable to stellar spectral data.

3.4 Diffusion Maps (DM)

DM (Coifman & Lafon 2006; Nadler et al. 2006) are a non linear data compression technique that assumes that the data (the spectra in our case) are contained in a manifold of much lower dimensionality than the embedding input space. The objective then is to find a representation in the manifold intrinsic coordinates. This is so even if the observed data (spectra) are non-uniformly distributed along the manifold, i.e., if the density of spectra is not uniform. A non-uniform data distribution may lead to reduced performance of regression algorithms.

DM are based on the assumption that there exist a low-dimensional manifold or topological space embedded in the high-dimension space of predictive variables. Thus, this technique aims to uncover the manifold structure in the data. We conjecture that a smooth variation of the stellar atmospheric parameters yield spectra that lie on a manifold. Therefore, we apply DM to attempt to discover the low-dimensional space that adequately represents such manifolds without loss of information.

DM start from a graph representation of the data, whereby each data point (spectrum) is a node. Nodes in the graph are connected by arcs with weights, and each weight measures the similarity between the nodes (spectra) it connects. Given the graph representation, we can define the quest for new coordinates in the manifold as a minimization process that involves the graph Laplacian matrix, the eigenvectors of which encode the new manifold intrinsic coordinates. Similarity can be approximated in a number of ways, including distances and kernels. The hope is that this new representation will capture the main structures of the data in few dimensions. In the low-dimensional representation of the data, DM attempt to retain the relationship between pairs of data points (spectra) as faithfully as possible.

In this work, the results were optimised by controlling the degree of locality in the diffusion weight matrix (referred to below with the parameter name *eps.val*).

Finally, the classical Nyström formula (Williams & Seeger 2001) was used to extend the diffusion coordinates computed on a subsample (the training set) to other spectra (the evaluation set).

3.5 Wavelets

Wavelets (Mallat 1998) are a set of mathematical functions used to approximate data and more complex functions by decomposing the signal in a hybrid space that incorporates both the original space where the data lie (which we will refer to as original space), and the transformed frequency domain. In our case, the original space will be the wavelength space, but in representing time series with wavelets the original space would be the time axis. The wavelet transform is a popular feature definition technique that has been developed to improve the shortcomings of the Fourier transform. Wavelets are considered better than Fourier analysis for modelling because they maintain the original space information while including information from the frequency domain.

Wavelets can be constructed from a function (named *mother wavelet*), which is confined to a finite interval in the original space. This function is used to generate a set of functions through the operation of scaling and dilation applied to the mother wavelet. The orthogonal or biorthogonal bases formed by this set allows the decomposition of any given signal using inner products, like in Fourier analysis. **That is, wavelet and Fourier analyses are similar in the sense that both of them break a signal down into its constituent parts for analysis. However, whereas the Fourier transform decomposes a signal into a series of sine waves of different frequencies, the wavelet transform decomposes the signal into its *wavelet* components (scaled and shifted versions of the *mother wavelet).***

At high frequency (short wavelength scales), the wavelets can capture discontinuities, ruptures and singularities or noise in the original spectrum. At low frequency (longer wavelength scales), the wavelet characterizes the coarse structure of the spectrum to identify the long-term trends and/or absorption bands, for example. Thus, wavelet analysis offers multi-resolution analysis in the original space and its frequency transformed domain, and it can be useful to reveal trends, breakdown points or discontinuities.

Data compression with wavelets consists of keeping a reduced number of wavelet coefficients. There are two common ways of coefficient selection: (i) to eliminate the **high frequency** coefficients that are assumed to reflect only random noise, and (ii) to keep the k most statistically significant wavelet coefficients (which yields a representation of the signal with less variance) (Li et al. 2010). There are more sophisticated ways to further reduce the number of wavelet coefficients using standard machine learning techniques for feature selection, such as the LASSO (Least Absolute Shrinkage and Selection Operator) used in Lu & Li (2015), wrapper approaches, information theory measures, etc. A full analysis of all these alternatives is out of the scope of this paper and we will only apply the first reduction mentioned above.

3.6 Kernel PCA

Kernel PCA is the reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function (Schölkopf et al. 1998). This method computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix. The reformulation of PCA in kernel space is straightforward, since a kernel matrix is similar to the inner product of the datapoints in the high-dimensional space that is constructed using the kernel function (the so-called *k*ernel trick). The application of PCA in the kernel space allows for the construction of nonlinear mappings of the input space.

Since Kernel PCA is a kernel-based method, the mapping performed relies on the choice of the kernel function. Possible choices for the kernel function include the linear kernel (i.e., traditional PCA), the polynomial kernel, and the Gaussian kernel. An important weakness of Kernel PCA is that the size of the kernel matrix is proportional to the square of the number of instances in the dataset.

In this work we used the Gaussian kernel and optimized the predictive performance by fine tuning the inverse kernel width (σ).

4 COMPARISON OF SPECTRUM COMPRESSION TECHNIQUES AND OPTIMAL RATES

We investigate the utility of six data compression techniques for feature extraction with a view to improving the performance of atmospheric parameters regression models. The robustness of these techniques against increasing SNR is evaluated, and the generalisation performance of training sets of varying SNRs is analysed.

Our set of experiments proceeds in three stages. In the first stage we aim at comparing the various compression techniques and compression rates **for different SNR regimes** in terms of the atmospheric parameter estimation errors. As a result of these experiments, we select an optimal compression approach and rate (dimensionality of the reduced space).

Different machine learning models have been used for the automatic estimation of atmospheric parameters from stellar spectra. Two of the most widely used techniques in practice are ANN and support vector machines (SVM). Unlike ANN, SVM does not need a choice of architecture before training, but there are some parameters to adjust in the kernel functions of the SVM. We use SVM with radial basis kernel functions and adjust the SVM parameters by maximizing the quality of the atmospheric parameter ($T_{\rm eff}$, log g, [M/H] or [α/Fe]) prediction as measured by the root mean squared error (RMSE, Eq. 1) in out-of-sample validation experiments.

$$RMSE_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{k;i} - \theta_{k;i}\right)^2} \tag{1}$$

where k indexes the atmospheric parameter (θ_k is one of $T_{\rm eff}$, log g, [M/H] or [α/Fe]), $\hat{\theta}_{k;i}$ and $\theta_{k;i}$ are the predicted and target values of θ_k for the i-th sample spectrum, and n represents the total number of spectra in our evaluation set.

In order to study the dependency of the estimation performance on the noise level of the input spectra, Gaussian white noise of different variances (SNRs equal to 100, 50, 25 and 10) was added to the original synthetic spectra. Then, the datasets were randomly split into two subsets, one for training (66% of the available spectra) and one for evaluation (the remaining 34%). Since the goal of these first experiments is to compare the compression techniques rather than obtaining the best predictor, splitting the dataset into training and evaluation sets is considered a good scheme. In essence, the experimental procedure consists of the following steps illustrated in Fig. 3:

(i) Compute the low-dimensional representation of the data using the training set. Because some of the techniques used to reduce the dimensionality depend on the setting of one or more parameters, a tuning process was performed in order to determine the optimal parameter values (in the sense that minimize the RMSE; see below). Table 1 presents the ranges of values that were searched, as well as the best parameter value obtained in each case.

(ii) Construct SVM models using the training set, and a varying number of dimensions (2, 5, 10, 15, 20, 25, 30 and 40) of the reduced space. The SVM parameters (kernel size and soft-margin width) and the compression parameters (where applicable; see Table 1) are fine-tuned to minimize the prediction error of the atmospheric parameter ($T_{\rm eff}$, log g, [M/H] or [α/Fe]).

(iii) Project the evaluation spectra onto the low-dimensional space computed in step (i).

(iv) Obtain atmospheric parameter predictions by applying the SVM models trained in step (ii) to the evaluation set obtained in step (iii).

(v) Calculate the performance of the predictor based on the RMSE obtained on the evaluation set.

4.1 Results

First, we compare the performance of the data compression techniques described in section 3 using noise-free synthetic spectra as well as degraded spectra with SNR levels of 100,

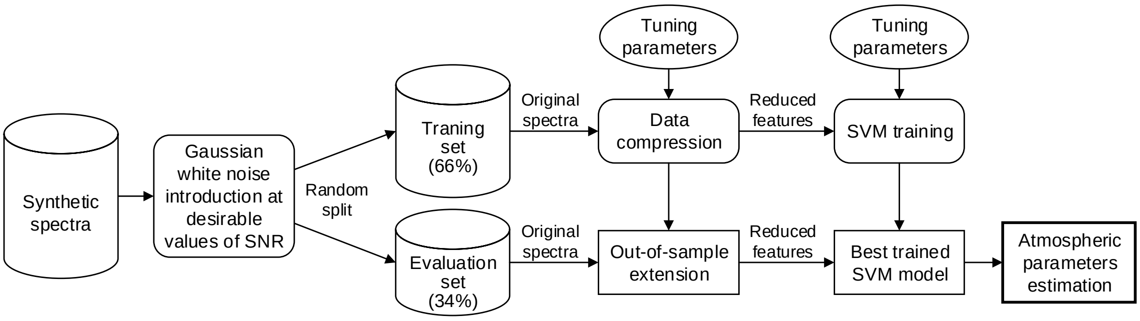

Figure 3. Process flow chart for investigating the performance of the selected data compression techniques.

Table 1. Summary of the parameters analysed for the data compression techniques.

Technique	Parameter	Analysed range	Best value
DLA	k_1	[2 - 8]	2
	k_2	[2 - 8]	3
DM	eps.val	[0.01 - 700]	600
Kernel PCA	σ	[0.0001 - 0.01]	0.001

50, 25 and 10. Figures 4 to 7 show the RMSE obtained with the evaluation set of the HR10 spectra (the 33% of the full set of spectra that was not not used to define the compression transformation or to train SVM models) grouped by SNR. Equivalent figures grouped by compression technique are included in Appendix A to facilitate comparisons.

Inspection of the figures reveals that the best strategies to compress the spectra are kernel PCA and ICA, with ICA performing marginally better than kernel PCA in most of the parameter space, except sometimes for the lowest compression rate. RMSE errors increase only moderately down to a SNR of 10, which seems to indicate that most of the examined compression techniques serve well as noise filters.

The performance comparison of the analysed data compression techniques shows that although traditional PCA is not the most efficient method, it outperforms some of the nonlinear techniques used in this study, such as DM or wavelets. The lower performance of DM compared to that of PCA could be partially explained by the Nyström extension. Although this method results in diffusion coordinates very similar to those that would be obtained by including the new spectra in the diffusion map, it may lead to small losses of prediction accuracy. As an illustration, the RMSE obtained for the $T_{\rm eff}$ in the high SNR regime (SNR=100) is between 0.5–1.5% better if the diffusion coordinates were computed from the whole dataset, instead of applying the out-of-sample extension. In the case of wavelets, it seems clear that even at the lowest compression rates of 40 components we are eliminating spectral information that is relevant for the subsequent regression task.

Overall, wavelets combined with SVM models have the highest errors regardless of the number of retained dimensions, with the exception of the [M/H] estimation where DLA performed worse for noisy synthetic spectra. Then, DLA was outperformed by most other techniques (except wavelet compression) for almost any compression rate and SNR. However, it achieved the lowest prediction errors for the hardly useful scenarios of noise-free data (not shown here for the sake of conciseness) or the highest compression rates (two or five dimensions) when estimating $T_{\rm eff}$ and log g. PCA and DM yield similar $T_{\rm eff}$ prediction errors in the high SNR regime, but DM are more robust against noise specially for the lowest compression rates examined.

It is interesting to note that compression techniques can be grouped into two categories: DLA, DM and Wavelets show a flat RMSE for target dimensions greater than ten, even for the lowest SNR explored in this Section (SNR=10); PCA, Kernel PCA and ICA show positive slopes in the RMSE curves for SNRs below 25 and target dimensionalities greater than 25, indicating that components beyond this limit are increasingly sensitive to noise.

The relative difference of DM with respect to the best performing compression techniques (ICA and kernel PCA) improves as the SNR diminishes until it becomes almost comparable for SNR=10, while at the same time rendering the SVM regression module insensitive to the introduction of irrelevant features (as shown by the flat RMSE curves for increasing numbers of dimensions used).

Table 2 quantifies the prediction errors of the best models for each SNR. It is interesting that ICA compression with 20 independent components remains as the best option for any SNR, except for the unrealistic noise-free data. These results evidence that for a given sample size (the number of spectra in the training set) there is an optimal number of features beyond which the performance of the predictor will degrade rather than improve. On the other hand, as expected, the quality of atmospheric parameter predictions degrades for lower SNR. However, RMSE errors were relatively low even for low SNR ($\sim$ 10).

4.1.1 Applicability of the HR10 results to the HR21 setup

The same analysis was carried out on the HR21 dataset characterized by a much wider wavelength range (almost twice as wide as the HR10 setup). Figure 8 and Table 3 show the results obtained for the $T_{\rm eff}$ with the evaluation set.

Some of our previous conclusions are confirmed by these results: *i*) Kernel PCA and ICA remain as the best compression techniques consistently for all SNRs, but at the lowest SNR (10), PCA and DM have comparable performances; *ii*) the SVM models trained with wavelet coefficients have the highest errors and are outperformed by PCA in most of the parameter space; and *iii*) DLA performed best for both noise-free data and the highest compression rates (two to five dimensions). However, there are also some differences.

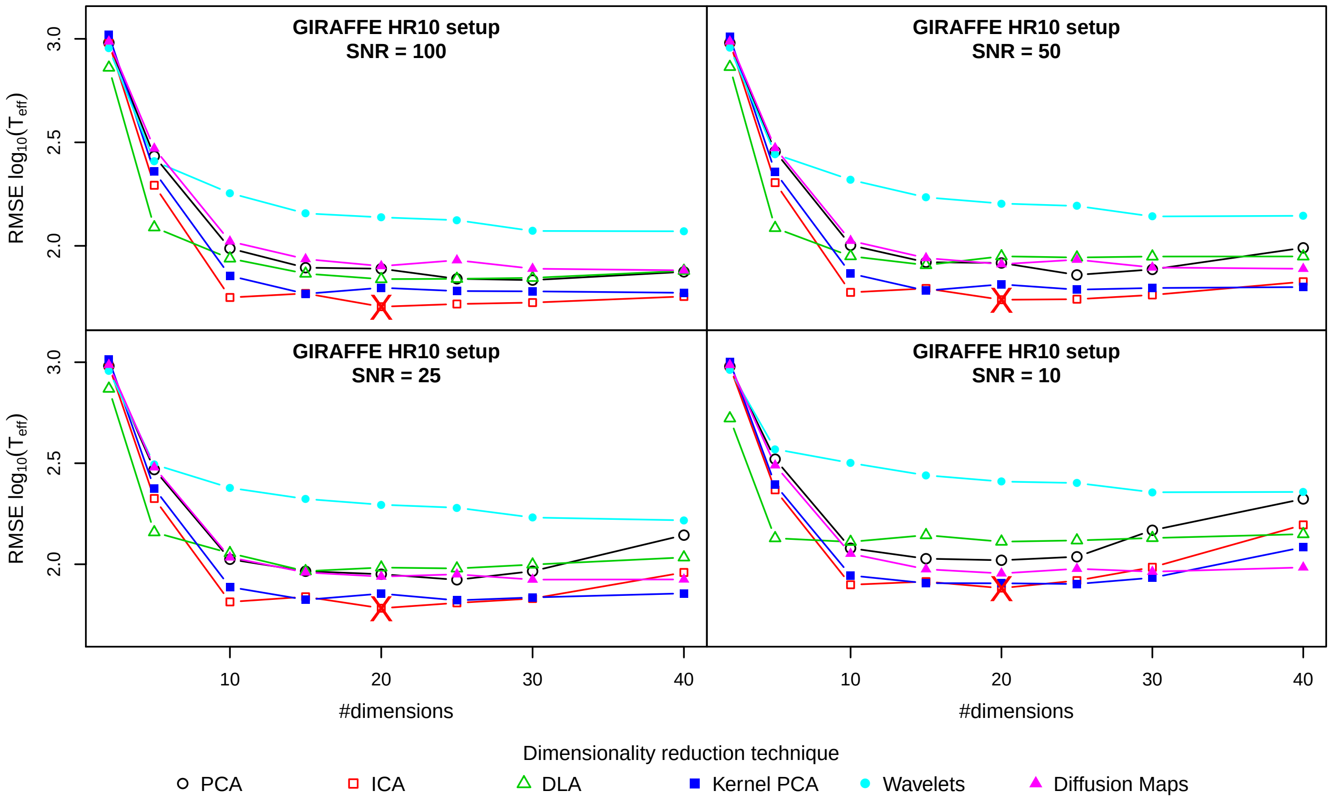

Figure 4. Temperature estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra from the nominal GIRAFFE HR10 setup.

Table 2. RMSE on the evaluation set of 2986 spectra for the best SVM trained models (HR10).

SNR	Method	Nr. Dim.	RMSE T_{eff} (K)	RMSE log g	RMSE [M/H] (dex)	RMSE $[\alpha/Fe]$ (dex)
∞	DLA / ICA[1]	40 / 30 / 20[2]	27.16	0.13	0.017	0.025
100	ICA	20	50.81	0.15	0.033	0.028
50	ICA	20	54.91	0.17	0.038	0.032
25	ICA	20	60.59	0.18	0.043	0.036
10	ICA	20	76.21	0.21	0.057	0.044

[1]The best performance for T_{eff}, log g and [M/H] was obtained with DLA, while best performance for $[\alpha/Fe]$ was obtained with ICA.
[2]The best performance for T_{eff} and log g was obtained with 40 dimensions, while for [M/H] and $[\alpha/Fe]$, 30 and 20 dimensions were needed respectively.

For low SNR data, the optimality criterion translates into retaining fewer components. This fact was first identified by Bailer-Jones et al. (1998) in the context of PCA compression of relatively low resolution spectra. We confirm this conclusion for other compression techniques in the HR21 setup where the wavelength coverage is greater than 300 Å, but not for the smaller coverage characteristic of the HR10 setup. Also, in the high SNR regimes the RMSE errors are lower with the HR21 setup than those obtained with the HR10 setup. However, the performance is considerably worsened for the lowest SNR explored in this work (SNR=10). This clearly indicates that the spectral information relevant for the prediction of effective temperatures is less robust to noise than in the case of the HR10 setup.

Table 3. RMSE on the evaluation set of 2986 spectra for the best SVM trained models (HR21).

SNR	Method	Nr. Dim.	RMSE T_{eff} (K)
∞	DLA	15	12.58
100	ICA	20	32.69
50	ICA	20	49.18
25	ICA	15	82.36
10	ICA	10	202.39

5 OPTIMAL TRAINING SET SNR

In this Section we analyse the optimal match between the SNR of the training set and that of the spectra for which

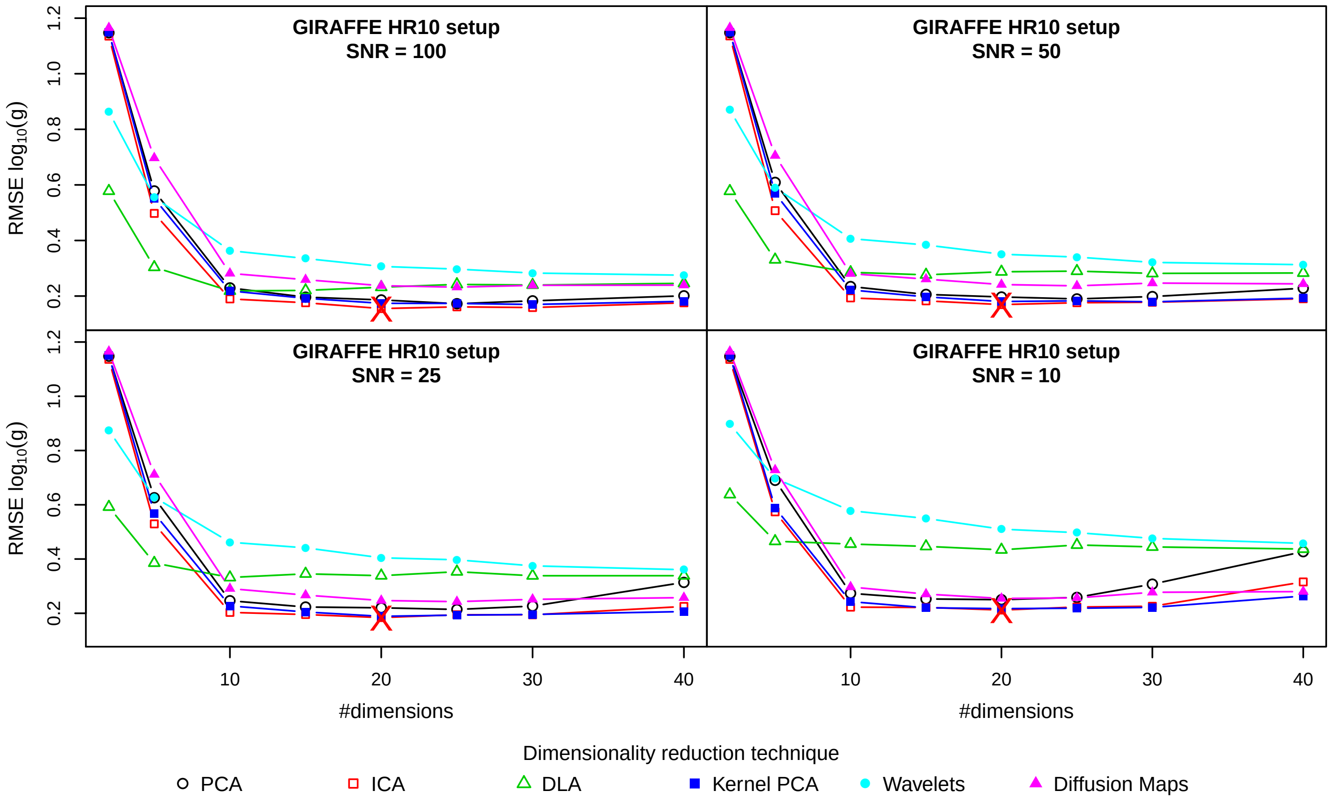

Figure 5. Surface gravity estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra from the nominal GIRAFFE HR10 setup.

the atmospheric parameter predictions are needed (in the following, the evaluation set).

In order to analyse the dependence of the predicition accuracy with the training set SNR, we generate 25 realisations of the noise for each of the following 8 finite SNR levels: 150, 125, 100, 75, 50, 25, 10 and 5. We create the 25 noise realisation in order to estimate the variance of the results and the significance of the differences. This amounts to $25 \times 8 = 200$ datasets, plus the noiseless dataset, all of which are compressed using ICA. The 20 first independent components are retained for the subsequent regression stage. The choice of compression technique and target dimensionality was dictated by the results presented in the previous Section. For each of these datasets we trained an SVM model to estimate each of the atmospheric parameters ($T_{\rm eff}$, log g, [M/H] or $[\alpha/Fe]$), and to assess the consistency of the results as the evaluation set SNR degrades. The model performances were evaluated using 10-fold cross validation as follows:

(i) The noiseless dataset is replicated 25×8 times: 25 realisations of Gaussian white noise for each of the following SNRs: 150, 125, 100, 75, 50, 25, 10, and 5. These 200 replicates together with the original noiseless dataset forms the basis for the next steps.

(ii) Each spectrum in each dataset is projected onto 20 independent components (as suggested by the experiments described in Section 4).

(iii) Each of the 201 compressed datasets is then split into 10 subsets or *folds*. The splitting is unique for the 201 datasets, which means that each spectrum belongs to the same fold across all 201 datasets.

(iv) An SVM model is trained using 9 folds of each dataset (all characterized by the same SNR). This amounts to 201 models.

(v) The model constructed in step (iv) is used to predict physical parameters for the tenth fold in all its 201 versions. The RMSE is calculated independently for each value of the SNR and noise realisation.

(vi) Steps (iv) to (v) are repeated 10 times (using each time a different fold for evaluation) and the performance measure is calculated by averaging the values obtained in the loop.

5.1 Results

Fig. 9 shows the mean (averaged over the 25 noise realisations) RMSE results and the 95% confidence interval for the mean as a function of the SNR of the evaluation set. The nine different lines correspond to the SNR of the training set used to generate both the projector and the atmospheric parameters predictor. The main conclusions of the analysis can be summarised as follows.

The analysis yields the very important (albeit somehow predictable) consequence that models trained with noise-free spectra are the worst choice for spectra with SNRs up 50/75, and are unnecessary for $T_{\rm eff}$, log g and $[\alpha/Fe]$ in contexts

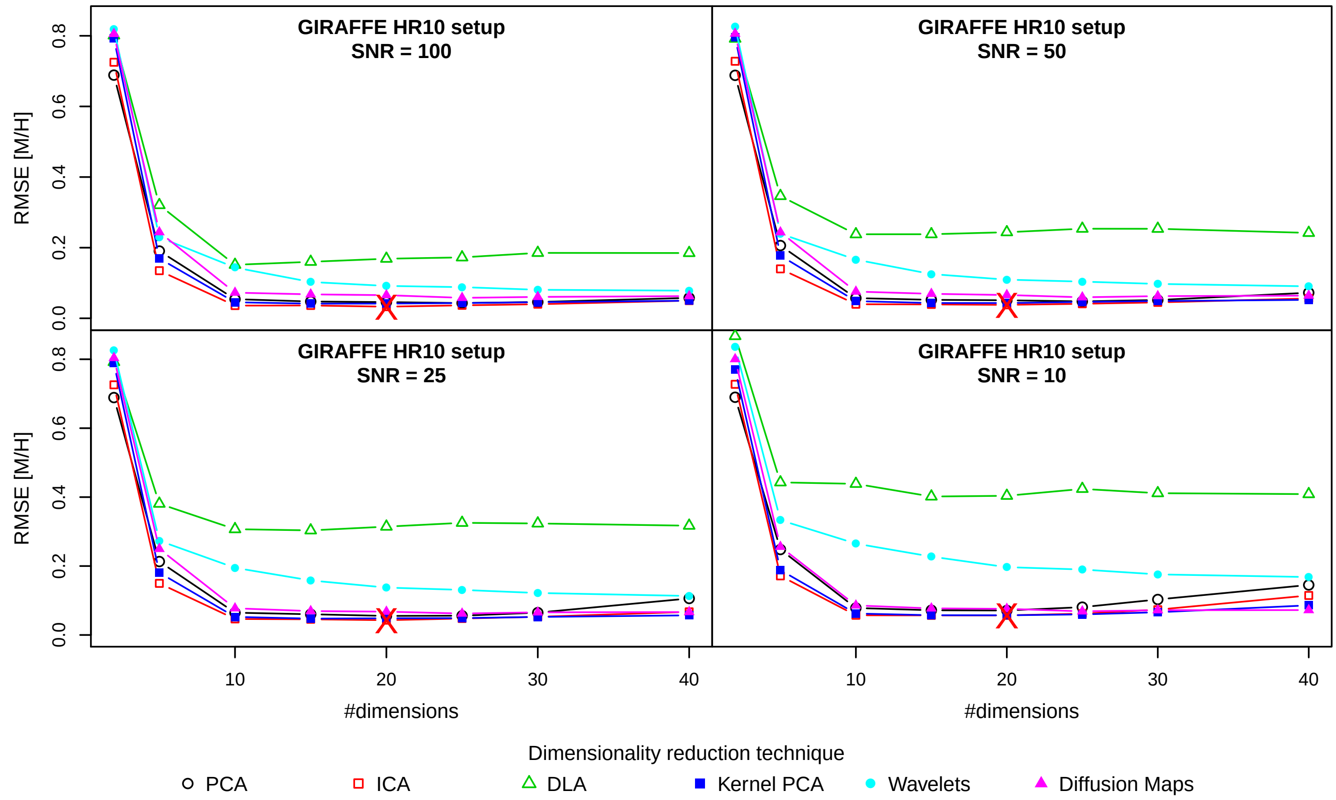

Figure 6. Metallicity estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra from the nominal GIRAFFE HR10 setup.

of higher SNRs. Only the [M/H] regression models slightly benefit from training with noiseless spectra if the evaluation spectra are in the SNR$\geq$ 50 regime. The accuracy of the model trained with noise-free spectra degrades exponentially for SNR<50.

There are no large discrepancies amongst the estimations obtained by applying the 25 models trained with a given SNR to different noise realisations, which translates into small confidence intervals and error bars in the plot. This is so even for the lowest SNR tested (SNR=5).

For the effective temperature and metallicity estimation from evaluation spectra with SNR > 50, there are minimal differences in the precision achieved by models trained with spectra of SNR$\geq$ 50, while for evaluation sets with 50 $\geq$ SNR > 10, the best accuracy is obtained with the model constructed from spectra with SNR of 50. For SNR lower than 10, the model with best generalisation performance is that trained with SNR=10. Hence, two models suffice to obtain the best performance across the entire SNR range explored in this set of experiments: one trained with SNR=50 examples for evaluation spectra of any SNR above 10, and one trained with SNR = 10 examples for lower SNRs.

Finally, only one ICA+SVM model trained with SNR=25 examples would be enough to estimate the surface gravity for spectra of all SNRs with the best performance (the SNR=50 model yielding similar although lower performances), and only one ICA+SVM model trained with SNR of 50 would be enough to estimate the alpha to iron ratio for spectra of all SNRs.

As a summary, models trained with noiseless spectra are either catastrophic choices or just equivalent to other models. Moreover, there is no need to match the SNR of the training set to that of the real spectra because only two ICA+SVM models would be enough to estimate $T_{\rm eff}$ and [M/H] in all SNR regimes, and a single model is needed for the optimal prediction of surface gravities and alpha to iron ratios.

5.1.1 *Application to the HR21 setup spectra*

The same evaluation procedure described above was applied to the HR21 setup spectra in order to check for the applicability of our conclusions in different wavelength ranges and coverages. Figure 10 shows the results obtained for the prediction of $T_{\rm eff}$.

Again, we observe that there is no need to match the SNR of the training set to that of the real spectra. Models trained with noise-free spectra are only adequate to estimate $T_{\rm eff}$ of noise-free spectra, and completely useless in any other SNR regime. This effect is much more evident here than in the case of the HR10 setup.

It is also clear that again, if the evaluation spectra are in the SNR>25 regime, the $T_{\rm eff}$ regression models have to be trained with SNR$\geq$50 examples. For evaluation spectra with SNR$\geq$100, the differences in the precision achieved by mod-

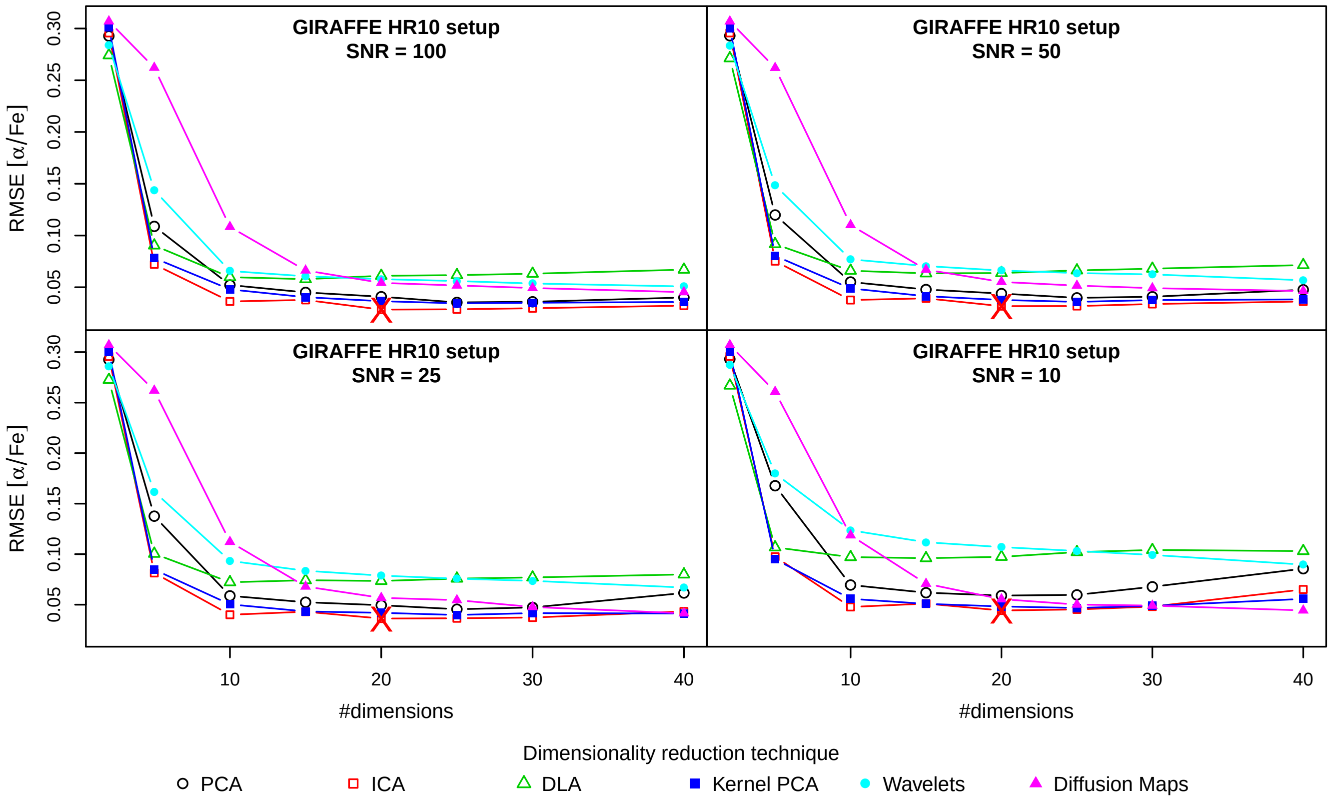

Figure 7. $[\alpha/Fe]$ estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra from the nominal GIRAFFE HR10 setup.

els trained with spectra of SNR≤50 are easier to notice than in the HR10 setup. There, the best option is one ICA+SVM model trained with SNR of 125.

In summary, our conclusions for the HR10 setup remain valid, except that a third model a model trained with SNR=125 examples would be marginally better than the SNR=50 one in the highest SNR regime (that is, above 100).

6 TRAINING SET DENSITY

In this Section we evaluate the dependence of the regression RMSE with the training set density. In order to simplify the interpretation of the results, we restrict the problem to solar metallicities and alpha abundace ratios. This simplification reduces the set of available spectra from 8780 to only 137 spectra in HR10 setup dataset with solar [M/H] and [alpha/Fe]. These 137 spectra are situated at the nodes of a regular grid except for a few gaps (see Fig. 1) that were interpolated as a weighted bilinear combination four nearest neighbours in the space of physical parameters. Thereafter, succesive grid refinements were obtained by recursively interpolating spectra at intermediate points between grid nodes. These interpolated spectra were obtained again as weighted linear combinations of the bracketing spectra, with weights given by the inverse square of the normalized euclidean distance to the nearest neighbours.

A total of six grids of synthetic spectra with different grid densities were used to train SVM models. The $T_{\rm eff}$ values varied between 4000 and 8000 K with step-sizes equal to 50, 62.5, 100, 125, 200, and 250 K. The other grid parameters were established as follows: the log g were regularly sampled from 1 to 5 dex in steps of 0.5 dex and both [M/H] and $[\alpha/Fe]$ were set equal to zero. Table 4 presents the step-sizes used in this study as well as the number of synthetic spectra available in each grid. In addition to this, noisy replicates of these grids were generated for four different SNR levels (100, 50, 25, 10).

Table 4. Size of the new datasets computed with different grid densities.

$T_{\rm eff}$ step-size (K)	Number of spectra
50	679
62.5	545
100	343
125	277
200	175
250	143

We evaluated the performance of the SVM regression models using 10-fold cross validation. Figures 11 and 12 present the $T_{\rm eff}$ estimation errors obtained with the different grid densities and the two optimal training set SNRs (50 and 10) found in the previous Section. Similar figures for SNR=25 and 100 are shown in Appendix B.

As expected, the estimation errors increase when the grid density decreases. We see how ICA remains as a winning alternative in this second scenario (a simplified train-

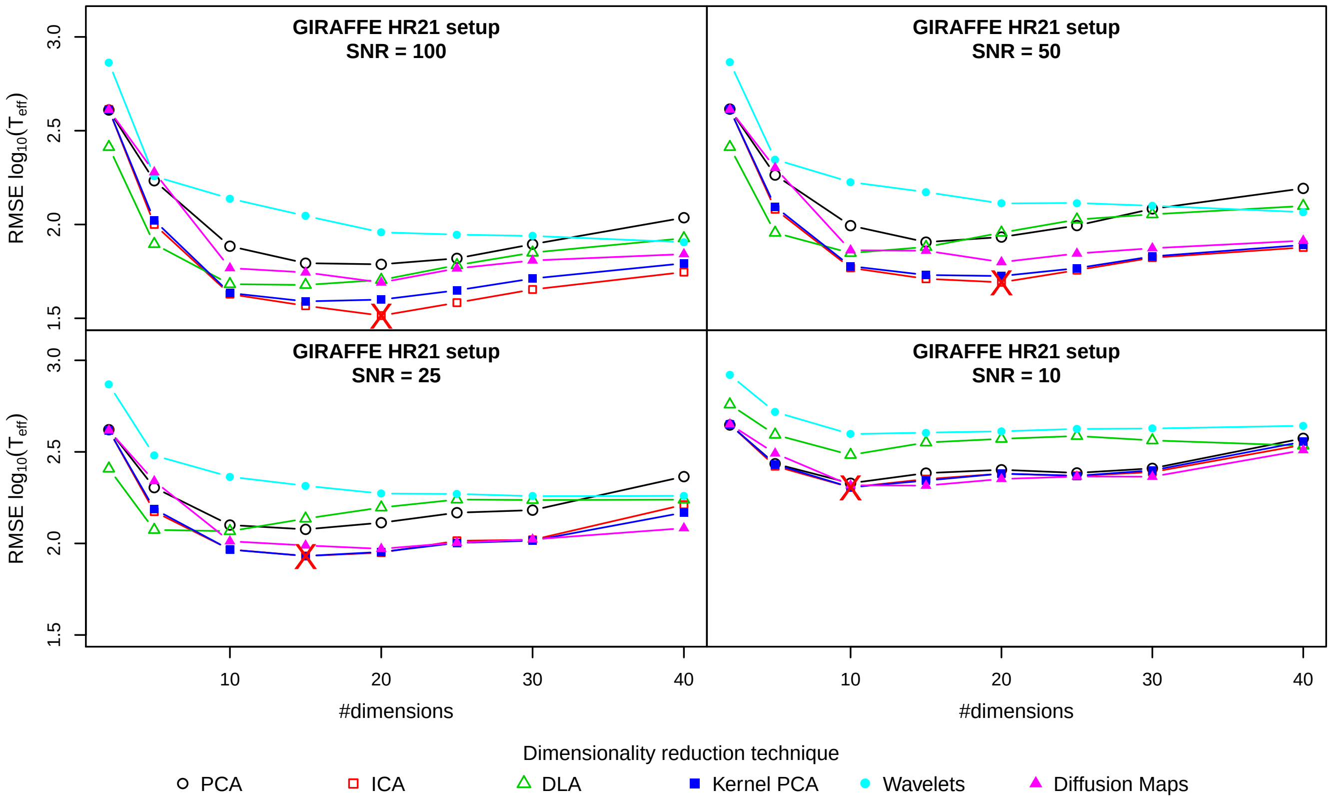

Figure 8. Temperature estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra from the nominal GIRAFFE HR21 setup.

ing set with no variation in metallicity or $[\alpha/Fe]$), where kernel PCA becomes non optimal and another linear technique (PCA) takes its place amongst the best performing techniques.

The prevalence of our conclusion for ICA as a winning alternative regardless of the grid spacing is reassuring. However, the non linear version of PCA lost its place amongst the best performing compression techniques. It is evident from the comparison of Figs. A1 and 11 that it is only at the largest grid spacings (250 K) that the non linear version of PCA performs better than the linear version (consistent with the results declared in Section 4), because the latter improves faster due to the grid refinement. It remains to be tested whether this faster decrease in the RMSE is due to the reduction in the training set complexity brought about by the removal of the non solar metallicities and $[\alpha/Fe]$ ratios, or it is still present in the full space of four physical parameters. It is plausible that the simplification to solar abundances brings the distribution of examples in feature space closer to a gaussian distribution where indeed the first principal components are effectively more correlated with the effective temperature.

It is interesting to note that the (non linear) compression with Diffusion Maps benefits much less from the grid refinement than the linear compression techniques PCA and ICA. Given the high dimensionality of the input space, it may be the case that much finer grid spacings are needed for the benefits of Diffusion Maps to become apparent. More experiments are needed to confirm this hypothesis, but insofar as the grid spacings are constrained to the values tested here, Diffusion Maps remain suboptimal choices.

7 CONCLUSIONS

In this work we have carried out a complete set of experiments to guide users of spectral archives to overcome the problems associated with the curse-of-dimensionality, when inferring astrophysical parameters from stellar spectra using standard machine learning algorithms.

In Section 4 we demostrate that, taken globally (that is, including the four stellar atmospheric parameters, a range of SNRs, and a range of compression ratios), Independent Component Analysis outperforms all other techniques, followed by Kernel Principal Component Analysis. The comparative advantage of using ICA is clearer for the $T_{\rm eff}$ and $[\alpha/Fe]$ regression models, and less evident for $\log g$ and [M/H]. Furthermore, we prove that this advantage holds too for a completely different wavelength range and a wavelength coverage twice as large. This is not enough to recommend ICA compression of high resolution spectra for any spectrograph observations, but it is a good indication that our results are not strongly dependent on the actual characteristics of the spectra.

The conclusions drawn from the set of experiments described in Section 4 are tied to the restricted range of physical parameters, wavelengths, and the spectral resolution of

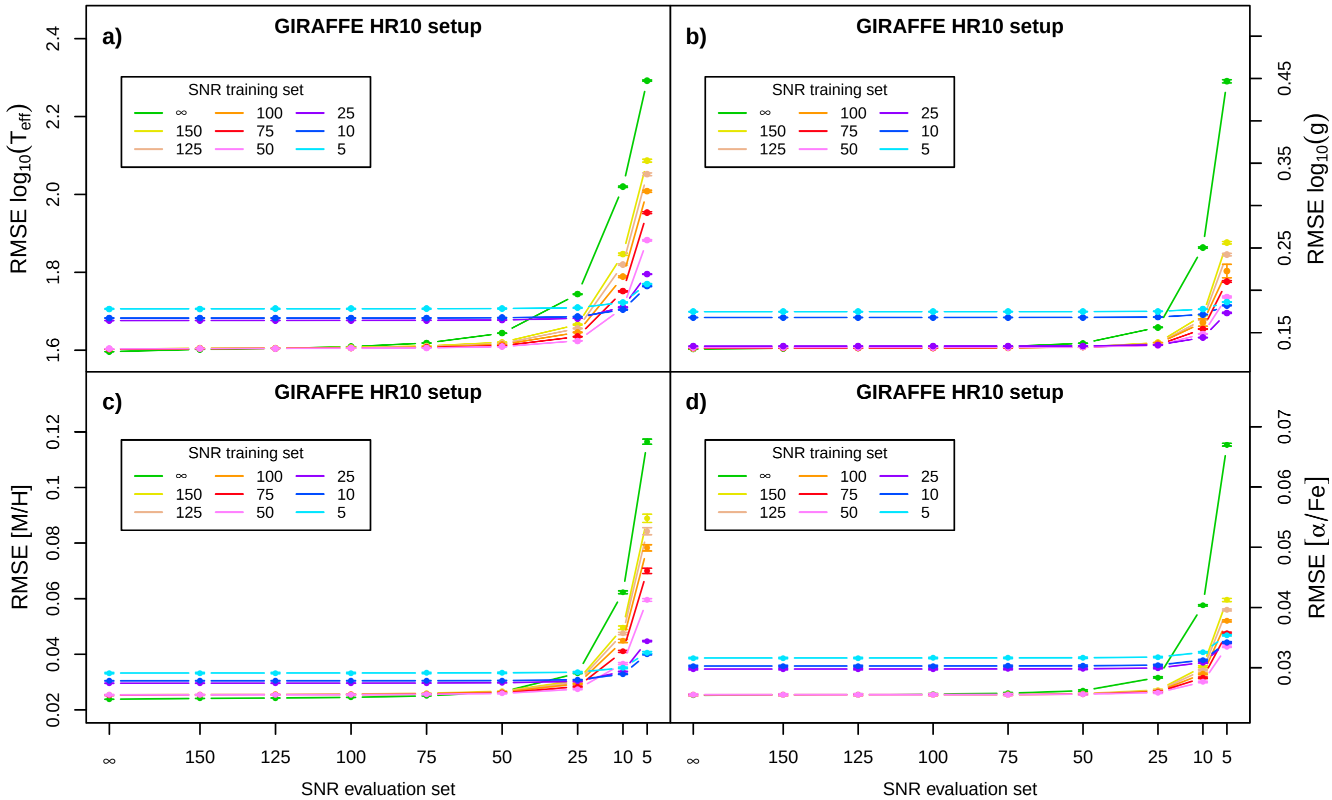

Figure 9. Estimation errors as a function of the SNR of the evaluation set for $T_{\rm eff}$ (a), $\log(g)$ (b) and [M/H] (c) and $[\alpha/Fe]$ (d). Each line corresponds to a model trained with a specific SNR (nominal GIRAFFE HR10 setup).

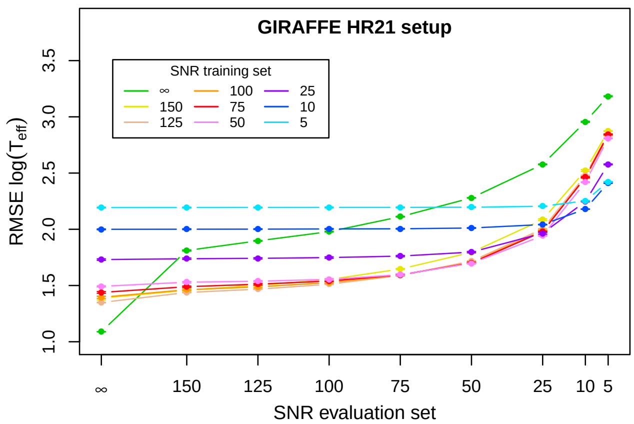

Figure 10. Estimation errors as a function of the SNR of the HR21 evaluation set for $T_{\rm eff}$. Each line corresponds to a model trained with a specific SNR.

the datasets adopted (the HR 10 and 21 setups), but we hope that they still hold for datasets of similar characteristics (different wavelength ranges but similar resolutions and parameter subspaces). In completely different scenarios such as the coolest regions of the Hertzsprung-Russell diagram, where spectra are dominated by molecular absoption bands, the validity of our results still remains to be proved.

In Section 5 we show that there is no need to match the SNR of unlabelled spectra (the spectra for which we want to predict astrophysical parameters) with a regression model trained with the same SNR. On the contrary, only two models are needed to achieve optimal performance in $T_{\rm eff}$ and $[\alpha/Fe]$ regression models (one trained with SNR=50 examples for SNR $>$ 10 spectra and one trained with SNR=10 examples for the lowest SNR regime), and only one model is needed for the prediction of $\log g$ and [M/H] (trained with SNR=25 and 50 examples respectively). The $T_{\rm eff}$ result holds also for the HR21 setup regression, although the model trained with SNR=125 is marginally better than the SNR=50 one in the highest SNR regime above 100.

In Section 6 we demostrate in a very simplified setup with no metallicity or alpha-enhancement effects incorporated in the training set, the importance of dense training sets in reducing the cross-validation errors, even in the context of compressed data spaces. We emphasize that this is only applicable to cross-validation errors (that is, errors estimated from spectra entirely equivalent to those used for training). These cross-validation errors are often known as internal errors as they do not take into account systematic differences between the training and evaluation sets. In our case, we have used MARCS model atmospheres, not observed spectra of real stars. In practical applications of the results presented above, the mismatch between the training set and the observed spectra inevitably leads to additional errors. It seems a reasonable working hypothesis to assume

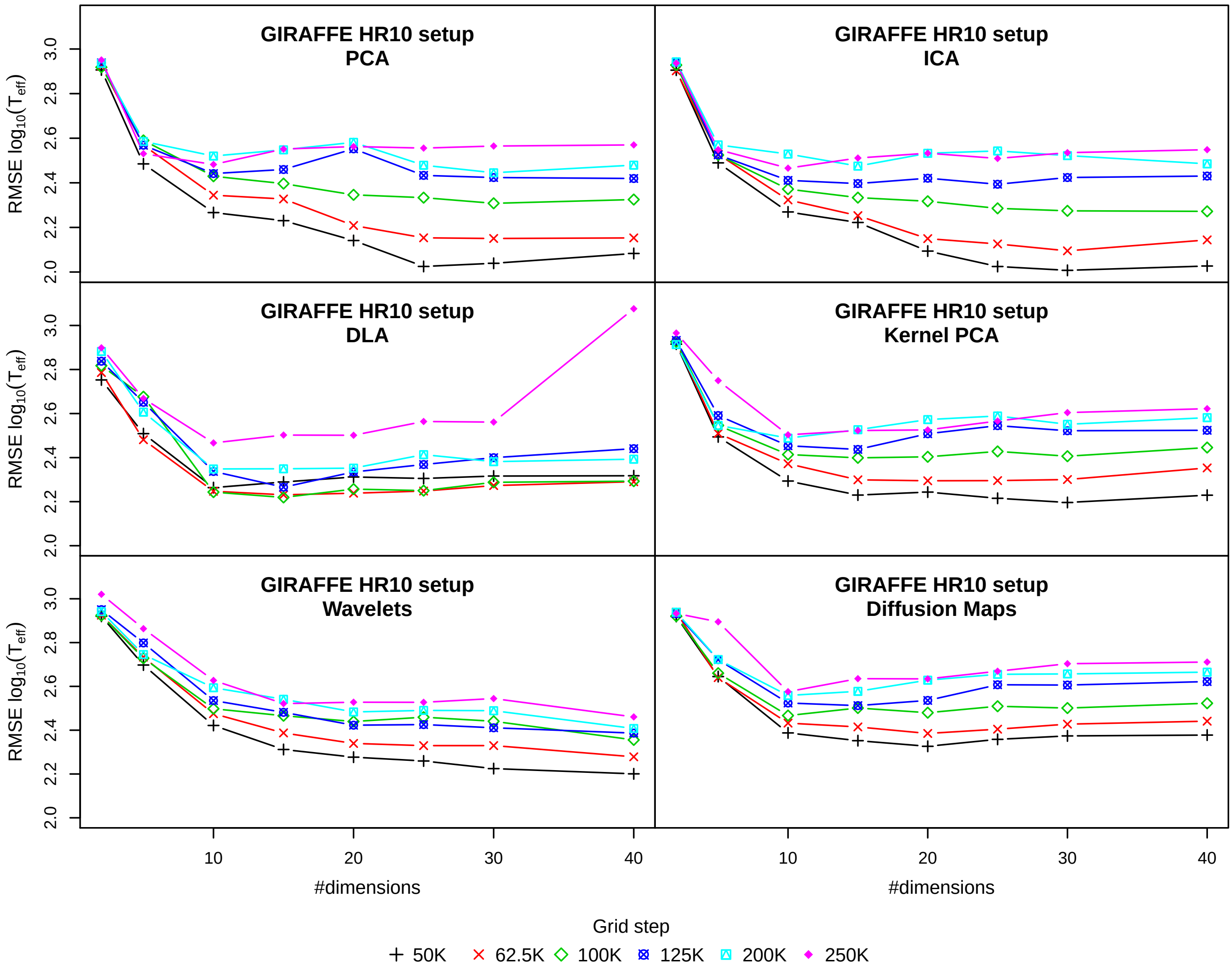

Figure 11. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 50)

that there is a limit beyond which the total errors are dominated by this mismatch and further increasing the training grid density will not significantly decrease the total errors.

Again, ICA turns out to be the best performing compression technique in the simplified experiments described in Section 6. The underlying assumptions of ICA may not be fulfilled to their full extent in the context of stellar spectra, but certainly we have reasonable hints that they apply, even if approximately. Our working hypothesis is that the independent components reflect groupings of spectral lines of various atomic elements with similar behaviour, such that the strengths and shapes of the lines pertaining to a given component respond in the same way to changes in the atmospheric parameters. Any such component would certainly have a non gaussian distribution across our training set (assumption one), albeit the fulfillment of the statistical independence assumption is, however, less clear under our interpretation of the ICA components. JADE maximizes non-gaussianity (rather than minimizing mutual information as in other flavours of ICA) via the fourth order moment of the distribution, and this turns out to result in the best projection amongst those tested in our regression models. This is certainly a result that holds for the synthetic spectra that constitute our working data set, but we have hints that this holds too for observed spectra (Sarro et al. 2013).

There are other reasons that may limit the applicability of the results presented in this work. Extending the applicability analysis to prove our conclusions universally valid is beyond the scope of this article.

In the first place, we have used the most standard or general versions of the techniques evaluated here. In the case of wavelet compression, for example, there are approaches to coefficient shrinkage other than simply removing the smallest spatial scales. The bibliography is endless and it would be impossible to test each and every proposed variation of the techniques presented here. In any case, it is important to note that the validity of our conclusions is limited to the standard versions tested here.

Another source of limitation is due to the use of a single regression model to assess the prediction errors. Again,

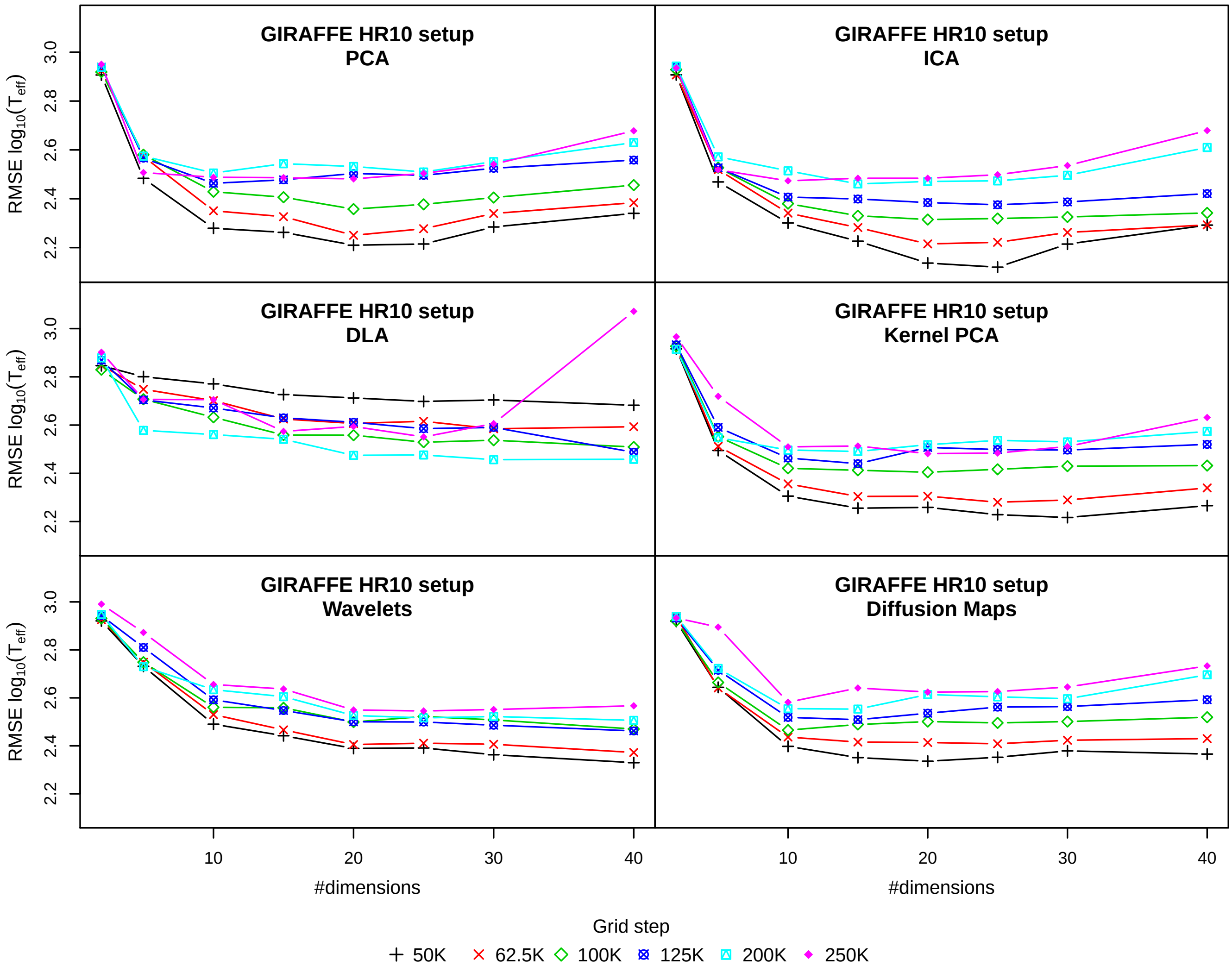

Figure 12. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 10)

Support Vector Machines and empirical risk minimization are very standard and robust statistical learning techniques amongst the top performing models for a very wide range of real life problems (van Gestel et al. 2004). Of course, the no-free-lunch theorem (see Igel & Toussaint 2005, and references therein for a formal statement of the theorem) always allows for the existence of algorithms that perform better than SVMs for this particular problem, but in the absence of free lunches, SVMs are a very reasonable choice and a good standard to measure the compression techniques.

Finally, we have focused our research in a battery of discriminative regression models where the *curse-of-dimensionality* may lead to severe problems. Forward models such as the *Cannon* (Ness et al. 2015) are not affected by this *curse-of-dimensionality* but would certainly benefit from a reduction of the more than 80000 parameters needed to model each and every flux in the spectrum.

ACKNOWLEDGEMENTS

This research was supported by the Spanish Ministry of Economy and Competitiveness through grant AyA2011-24052.

REFERENCES

Allende Prieto C., Beers T. C., Wilhelm R., Newberg H. J., Rockosi C. M., Yanny B., Lee Y. S., 2006, The Astrophysical Journal, 636, 804

Alvarez R., Plez B., 1998, Astronomy and Astrophysics, 330, 1109

Bailer-Jones C. A. L., Irwin M., von Hippel T., 1998, MNRAS, 298, 361

Balasubramanian M., Schwartz E. L., Tenenbaum J. B., de Silva V., Langford J. C., 2002, Science, 295(5552), 7

Bell A., Sejnowski T. J., 1995, Neural Computation, 7(6), 1129

Bellman R., 1961, Adaptive Control Processes: A Guided Tour. Princeton University Press

Belouchrani A., Meraim K. A., Cardoso J. F., Moulines E., 1997, IEEE Transaction on Signal Processing, 45(2), 434

Bruntt H., et al., 2010, MNRAS, 405, 1907
Bu Y., Chen F., Pan J., 2014, New Astronomy, 28, 35
Cardoso J. F., Souloumiac A., 1993, IEEE Transactions on Signal Processing, 140(6), 362
Coifman R. R., Lafon S., 2006, Applied and Computational Harmonic Analysis, 21(1), 5
Comon P., 1994, Signal Processing, 36, 287
Daniel S. F., Connolly A., Schneider J., Vanderplas J., Xiong L., 2011, The Astronomical Journal, 142, 203
Eisenstein D. J., et al., 2011, AJ, 142, 72
Geary R. C., 1930, Journal of the Royal Statistical Society, 93, 442
Gilmore G., et al., 2012, The Messenger, 147, 25
Gustafsson B., Edvardsson B., Eriksson K., JÃÿrgensen U. G., Nordlund A., Plez B., 2008, Astronomy and Astrophysics, 486(3), 951
Hotelling H., 1933, Journal of Educational Psychology, 24(6&7), 447
Hyvärinen A., Oja E., 2000, Neural Networks, 13(4-5), 411
Igel C., Toussaint M., 2005, Journal of Mathematical Modelling and Algorithms, 3, 313
Jain A. K., Duin R. P., Mao J., 2000, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 4
Jordi C., et al., 2006, MNRAS, 367, 290
Jutten C., Hérault J., 1991, Signal Processing, 24, 1
Li H., Adali T., 2008, IEEE Transactions on Neural Networks, 19(3), 408
Li T., Ma S., Ogihara M., 2010, in Maimon O., Rokach L., eds, , Data Mining and Knowledge Discovery Handbook. Springer, pp 553–571
Li X., Lu Y., Comte G., Luo A., Zhao Y., Wang Y., 2015, ApJS, 218, 3
Lu Y., Li X., 2015, MNRAS, 452, 1394
Mallat S., 1998, A Wavelet Tour of Signal Processing. Academic Press
Manteiga M., Ordóñez D., Dafonte C., Arcay B., 2010, PASP, 122, 608
Marsaglia G., 1965, Journal of the American Statistical Association, 60, 193
Mishenina T. V., Bienaymé O., Gorbaneva T. I., Charbonnel C., Soubiran C., Korotin S. A., Kovtyukh V. V., 2006, A&A, 456, 1109
Nadler B., Lafon S., Coifman R. R., Kevrekidis I. G., 2006, Applied and Computational Harmonic Analysis: Special Issue on Diffusion Maps and Wavelets, 21, 113
Navarro S. G., Corradi R. L. M., Mampaso A., 2012, Astronomy and Astrophysics, 538, A76, 1
Ness M., Hogg D. W., Rix H.-W., Ho A. Y. Q., Zasowski G., 2015, ApJ, 808, 16
Ollila E., Koivunen V., 2006, IEEE Transactions on Signal Processing, 89(4), 365
Pearson K., 1901, Philosophical Magazine, 2(11), 559
Plez B., 2012, Turbospectrum: Code for spectral synthesis, record ascl:1205.004, http://adsabs.harvard.edu/abs/2012ascl.soft05004P
Re Fiorentin P., Bailer-Jones C., Beers T., Zwitter T., 2008a, in Proceedings of the International Conference: "Classification and Discovery in Large Astronomical Surveys". pp 76–82
Re Fiorentin P., Bailer-Jones C. A. L., Lee Y. S., Beers T. C., Sivarani T., Wilhelm R., Allende Prieto C., Norris J. E., 2008b, Astronomy and Astrophysics, 467(3), 1373
Recio-Blanco A., Bijaoui A., de Laverny P., 2006, Monthly Notices of the Royal Astronomical Society, 370, 141
Recio-Blanco A., et al., 2014, A&A, 567, A5
Recio-Blanco A., et al., 2015, preprint, (arXiv:1510.00111)
Rojas-Ayala B., Covey K. R., Muirhead P. S., Lloyd J. P., 2010, ApJ, 720, L113
Rojas-Ayala B., Covey K. R., Muirhead P. S., Lloyd J. P., 2012, ApJ, 748, 93
Roweis S., Saul L., 2000, Science, 290(5500), 2323
Sarro L. M., et al., 2013, A&A, 550, A120
Saxena A., Gupta A., Mukerjee A., 2004, in Pal N., Kasabov N., Mudi R., Pal S., Parui S., eds, Lecture Notes in Computer Science, Vol. 3316, Neural Information Processing. Springer Berlin Heidelberg, pp 1038–1043
Schölkopf B., Smola A., K.-R.MÃijller 1998, Neural Computation, 10(5), 1299
Singh H., Gulati R., Gupta R., 1998, Monthly Notices of the Royal Astronomical Society, 295(2), 312
Snider S., Allende Prieto C., von Hippel T., Beers T., Sneden C., Qu Y., Rossi S., 2001, The Astrophysical Journal, 562, 528
Steinmetz M., et al., 2006, AJ, 132, 1645
Tenenbaum J. B., de Silva V., Langford J. C., 2000, Science, 290(5500), 2319
Torres G., Fischer D. A., Sozzetti A., Buchhave L. A., Winn J. N., Holman M. J., Carter J. A., 2012, ApJ, 757, 161
Vanderplas J., Connolly A., 2009, The Astronomical Journal, 138, 1365
Walker M. G., Olszewski E. W., Mateo M., 2015, MNRAS, 448, 2717
Williams C. K. I., Seeger M., 2001, in Leen T. K., Dietterich T. G., Tresp V., eds, , Advances in Neural Information Processing Systems 13. MIT Press, pp 682–688
Zarzoso V., Comon P., 2010, IEEE Transactions on Neural Networks, 21(2), 248
Zhang Z., Zha H., 2002, eprint arXiv:cs/0212008,
Zhang T., Tao D., Yang J., 2008, in Forsyth D., Torr P., Zisserman A., eds, Lecture Notes in Computer Science, Vol. 5302, Computer Vision - ECCV 2008. Springer Berlin Heidelberg, pp 725–738
de Laverny P., Recio-Blanco A., Worley C. C., Plez B., 2012, A&A, 544, A126
van Gestel T., Suykens J. A., Baesens B., Viaene S., Vanthienen J., Dedene G., de Moor B., Vandewalle J., 2004, Machine Learning, 54, 5

APPENDIX A: REGRESSION ERRORS FOR THE GIRAFFE HR10 SETUP GROUPED BY COMPRESSION TECHNIQUE.

APPENDIX B: EFFECTIVE TEMPERATURE REGRESSION ERRORS AS A FUNCTION OF THE GRID SPACING AND GROUPED BY COMPRESSION TECHNIQUE FOR SNR=25 AND 100

This paper has been typeset from a TEX/LATEX file prepared by the author.

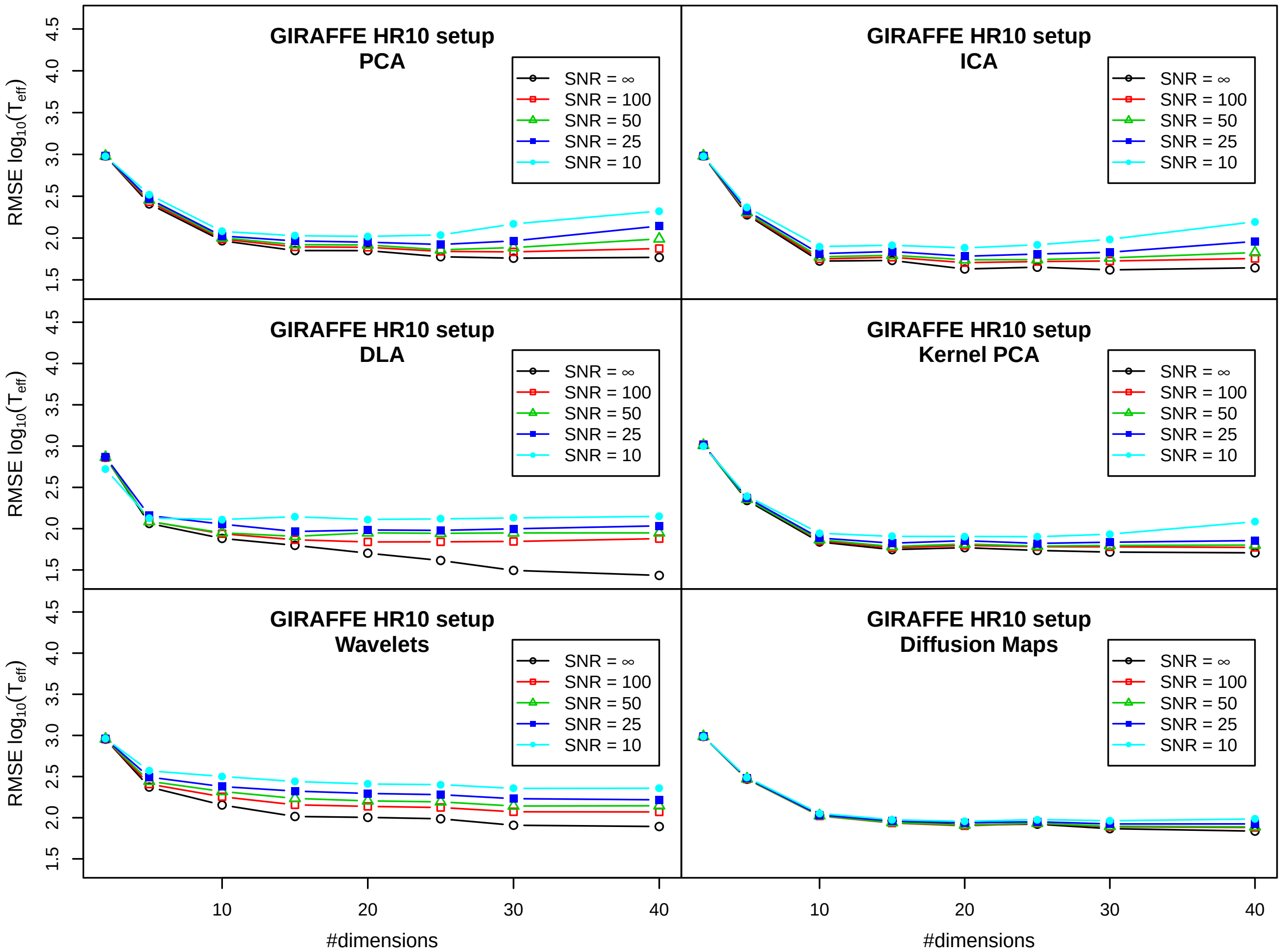

Figure A1. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

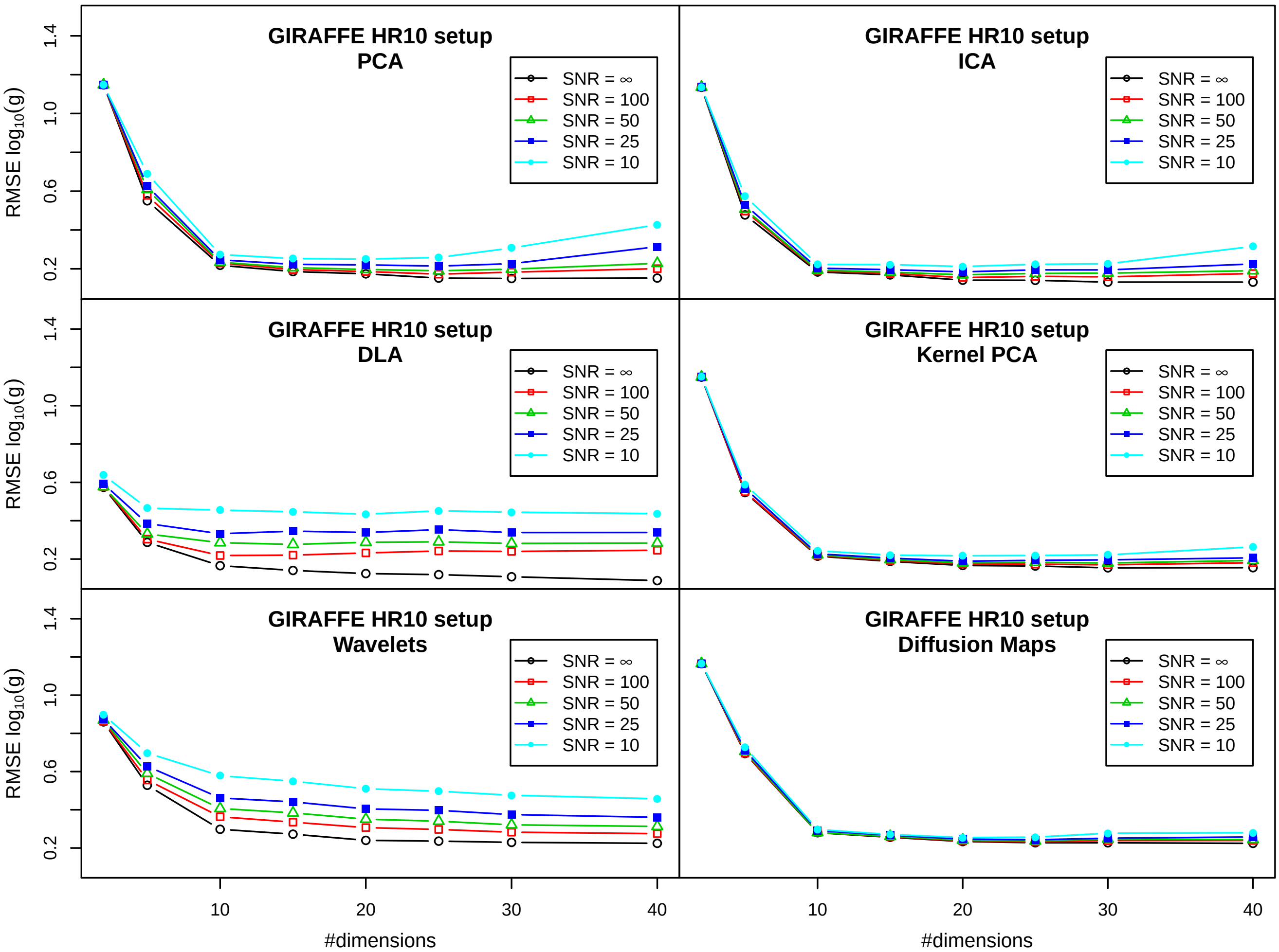

Figure A2. Surface gravity estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

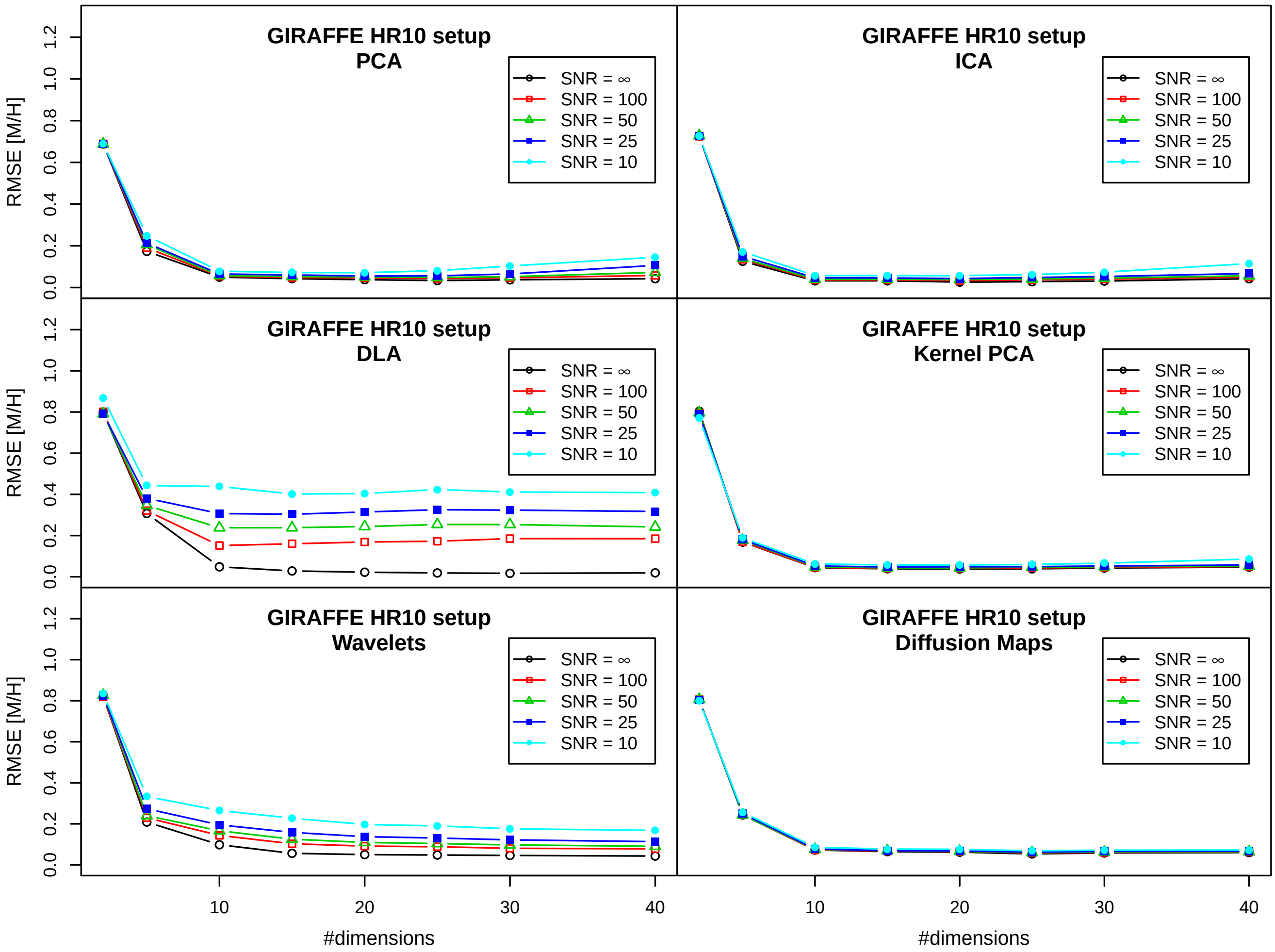

Figure A3. Metallicity estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

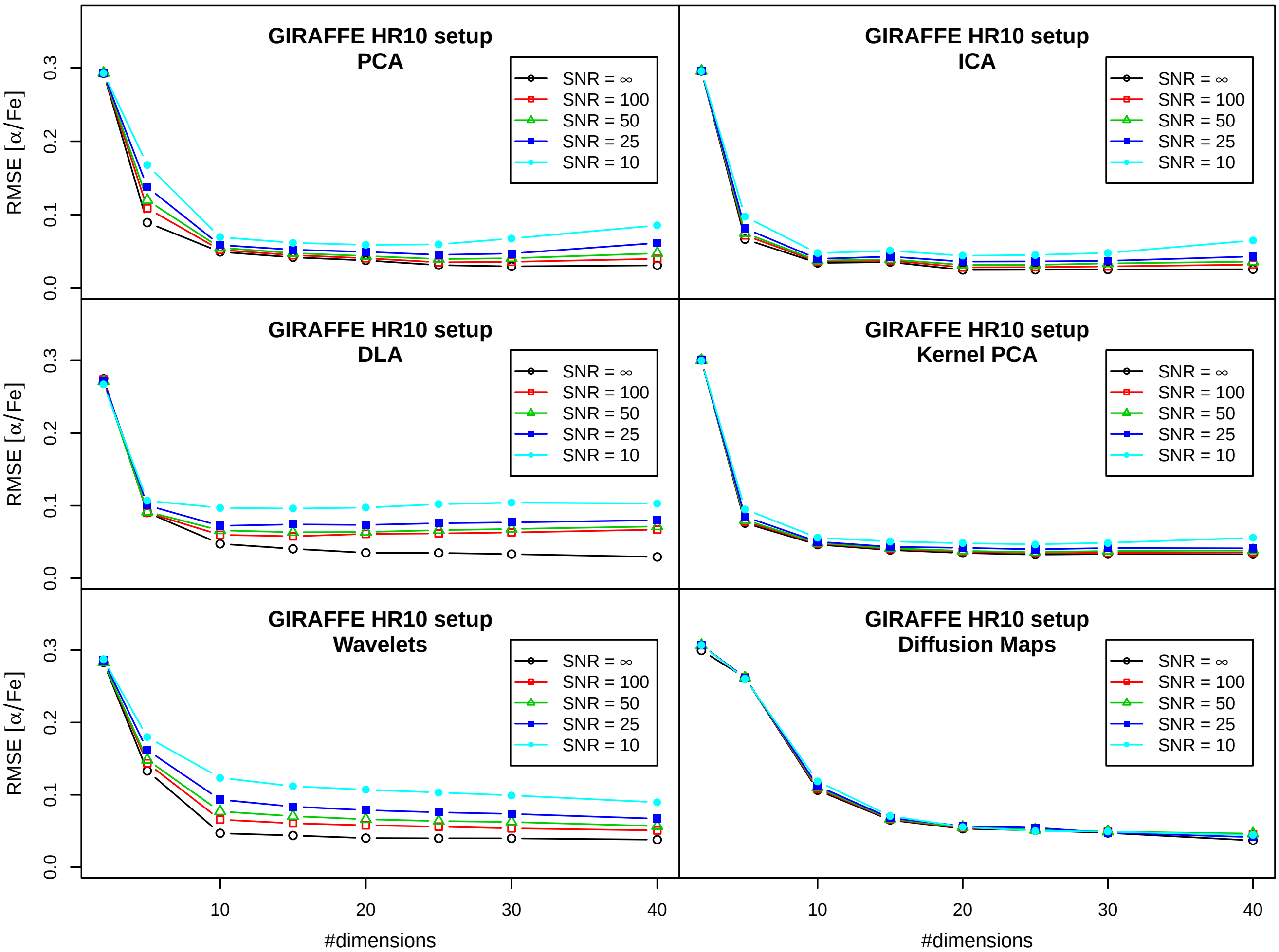

Figure A4. $[\alpha/Fe]$ estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

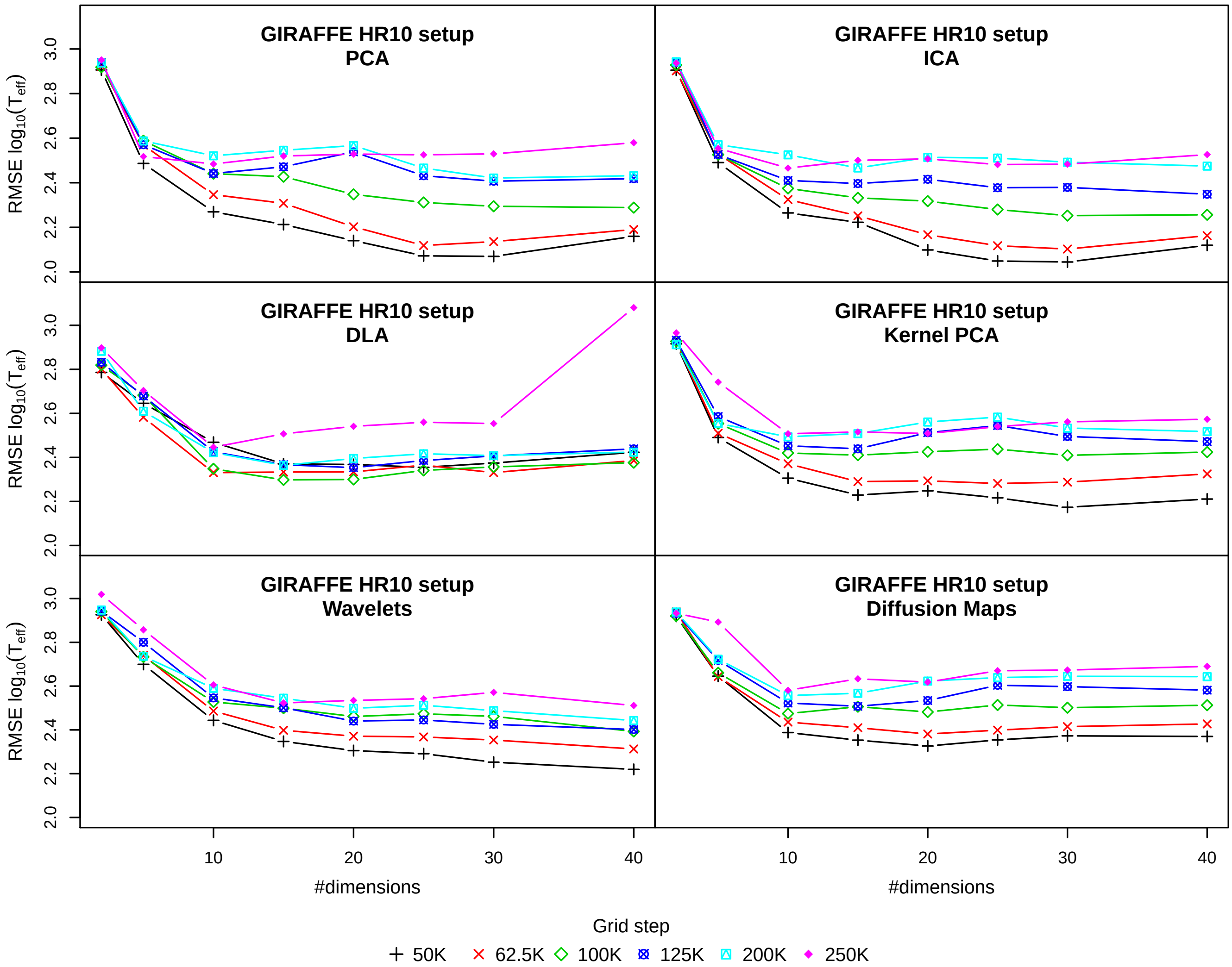

Figure B1. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 25)

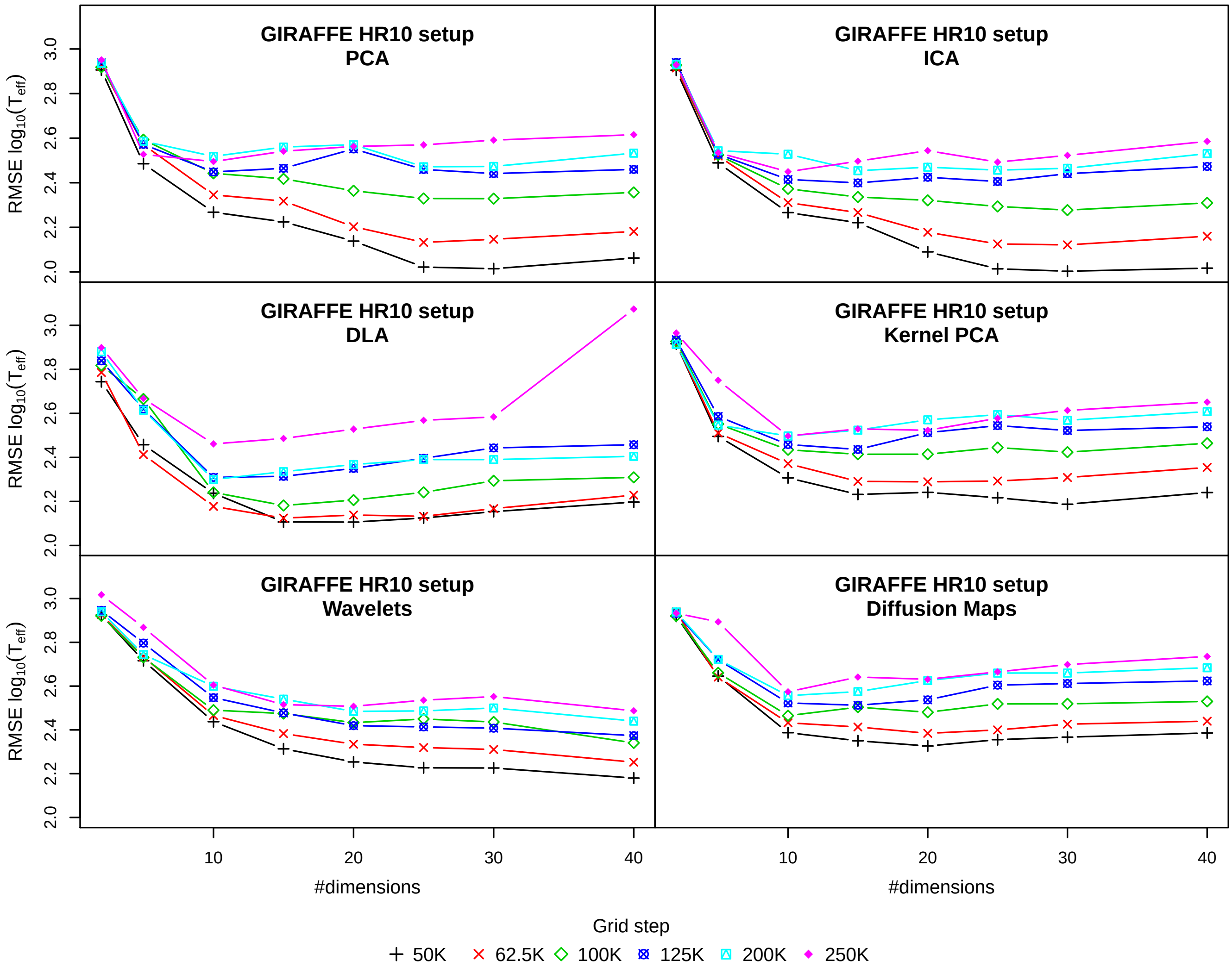

Figure B2. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 100)