MNRAS 000, 1–18 (2015) Preprint 12 July 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

Keith T. Smith,[1★] A. N. Other,[2] Third Author[2,3] and Fourth Author[3]
[1] *Royal Astronomical Society, Burlington House, Piccadilly, London W1J 0BQ, UK*
[2] *Department, Institution, Street Address, City Postal Code, Country*
[3] *Another Department, Different Institution, Street Address, City Postal Code, Country*

Accepted XXX. Received YYY; in original form ZZZ

**ABSTRACT**
We evaluate the utility of several data compression techniques for alleviating the curse-of-dimensionality problem in regression tasks where the objective is to estimate stellar atmospheric parameters from high resolution spectra in the 4000-8000 K range. We conclude that ICA and kernel-PCA perform better than the rest of the techniques evaluated for all compression ratios. We also assess the necessity to adapt the signal-to-noise ratio (SNR) of the training set examples to the SNR of each test spectrum and conclude that within the conditions of our experiments, only two such models are needed (SNR=50 and 10) to cover the entire range.

**Key words:** Dimensionality Reduction – keyword2 – keyword3

## 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 properties 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 properties, such as effective temperature ($T_{\rm eff}$), surface gravity ($\log g$), metallicity ([M/H]) and alpha to iron ratio ($[\alpha/Fe]$). However, the difficulty that atmospheric parameter estimation poses comes from the inherent size and dimensionality of the data. Regression from stellar spectra suffers the so-called *curse of dimensionality* problem because the number of variables (wavelengths) is much higher than the number of training samples.

The *curse of dimensionality* (Bellman 1961) relates to the problem caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space. When the dimensionality increases, the volume of the space increases so fast that the available data become sparse. Because this sparsity is problematic for any method that requires statistical significance, the amount of data 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 independent variables is to concentrate on certain wavelength ranges that contain spectral lines that are sensitive to changes in the physical parameters. Before the advent of the large-scale spectroscopic surveys, 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.

★ E-mail: mn@ras.org.uk (KTS)

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), Muirhead et al. (2012), or Mishenina et al. (2006)). That way, we significantly reduce the full spectrum to a much smaller number of independent 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. This is potentially dangerous because, even in the best case that the continuum flux is Gaussian distributed around a value significantly different from zero, the ratio distribution is asymmetric and has a heavy right tail. 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 recent years, there seems to be a tendency to use the full spectrum rather than selected wavelength ranges (see e.g. Recio-Blanco et al. (2014), 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 techiques to serve as a guide for future applications of machine learning techiques for regression of stellar atmospheric physical parameters.

The most popular dimensionality reduction technique applied to stellar spectra is Principal Component Analysis (PCA). It has been widely applied in spectral classification combined with artificial neural networks (ANNs) (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 dimensionality reduction 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 sample to that of the real sample 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 artificial neural networks (ANNs) 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 ratios 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 $\sim$ 25. In Navarro et al. (2012) authors present a system based on ANNs trained with a set of line-strength indexes 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 to classify spectra of SNR as low as 20 with an accuracy better than two spectral subtypes. For SNR $\sim$ 10, classification is still possible but at a lower precision.

This paper presents a comparative study of the most popular dimensionality reduction 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 dimensionality reduction 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 test data, 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 review the data compression techniques evaluated in this work and their properties. In Sect. 3 we describe the dataset used in our experiments. 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 prediction test, 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 DIMENSIONALITY REDUCTION

For the sake of computational efficiency in a dynamic environment where a complete rerun of a dimensionality reduction algorithm becomes prohibitively time consuming, the selection of the dimensionality reduction 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 dimensionality reduction techniques such as PCA, independent component analysis (ICA) and discriminative locality alignment (DLA), as well as three nonlinear reduction techniques that do not lack generalisation 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.

Other linear and nonlinear techniques could be used for dimensionality reduction, 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.

### 2.1 Principal Component Analysis (PCA)

Principal Components Analysis (PCA) (Hotelling 1933; Pearson 1901) is by far the most popular (unsupervised) linear technique for dimensionality reduction. 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.

### 2.2 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) (Comon 1994) is very closely related to the method called blind source separation (BSS) or blind signal separation (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.

### 2.3 Discriminative Locality Alignment (DLA)

Discriminative Locality Alignment (DLA) (Zhang et al. 2008) is a supervised manifold learning algorithm which can be divided into three stages: part optimisation, sample weighting and whole alignment. In the first stage, for each sample (each spectrum in our case) a patch is defined by the given sample and its neighbours. On each patch, DLA preserves the local discriminative information through integrating the two criteria that *i*) the distances between intra-class samples are as small as possible and *ii*) the distance between the inter-class samples 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 sample 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:

- Neighbour samples from an identical class ($k_1$): the number of nearest neighbours with respect to $x_i$ from samples in the same class with $x_i$
- Neighbour samples from different classes ($k_2$): the number of nearest neighbours with respect to $x_i$ from samples in different classes with $x_i$

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 linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) not directly applicable to stellar spectral data.

### 2.4 Diffusion Maps

Diffusion maps (DM) (Coifman & Lafon 2006; Nadler et al. 2006) are a non linear dimensionality reduction technique for finding the feature representation of the datasets even if observed samples are non-uniformly distributed.

DMs achieve dimensionality reduction by re-organizing data according to parameters of its underlying geometry. DM are based on defining a Markov random walk on the data. By performing the random walk for a number of time steps, a measure for the proximity of the data points is obtained (*diffusion distance*). In the low-dimensional representation of the data, DMs attempt to retain the pairwise diffusion distances as faithfully as possible (under a squared error criterion). The key idea behind the diffusion distance is that it is based on integrating over all paths through the graph. This makes the diffusion distance more robust to short-circuiting than, e.g., the geodesic distance that is employed in Isomap (Tenenbaum et al. 2000).

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

### 2.5 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. This method 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.

Dimensionality reduction 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 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 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.

### 2.6 Kernel PCA

Kernel PCA (KPCA) 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 ($\sigma$).

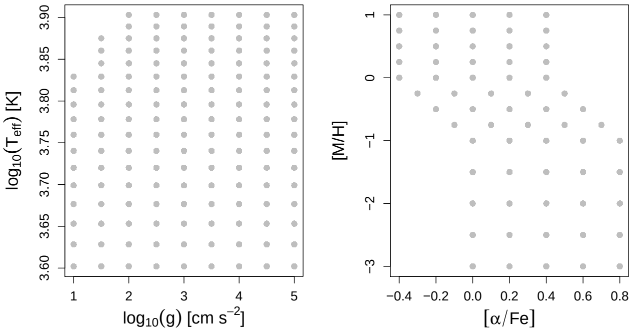


**Figure 1.** Coverage in parameter space of the dataset

## 3 THE DATASET

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.

The dataset 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 $\alpha$ 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 $\alpha$-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). For example, with an amount of information about 10 samples per dimension –a rule of thumb is to have at least 10 training samples per feature di-

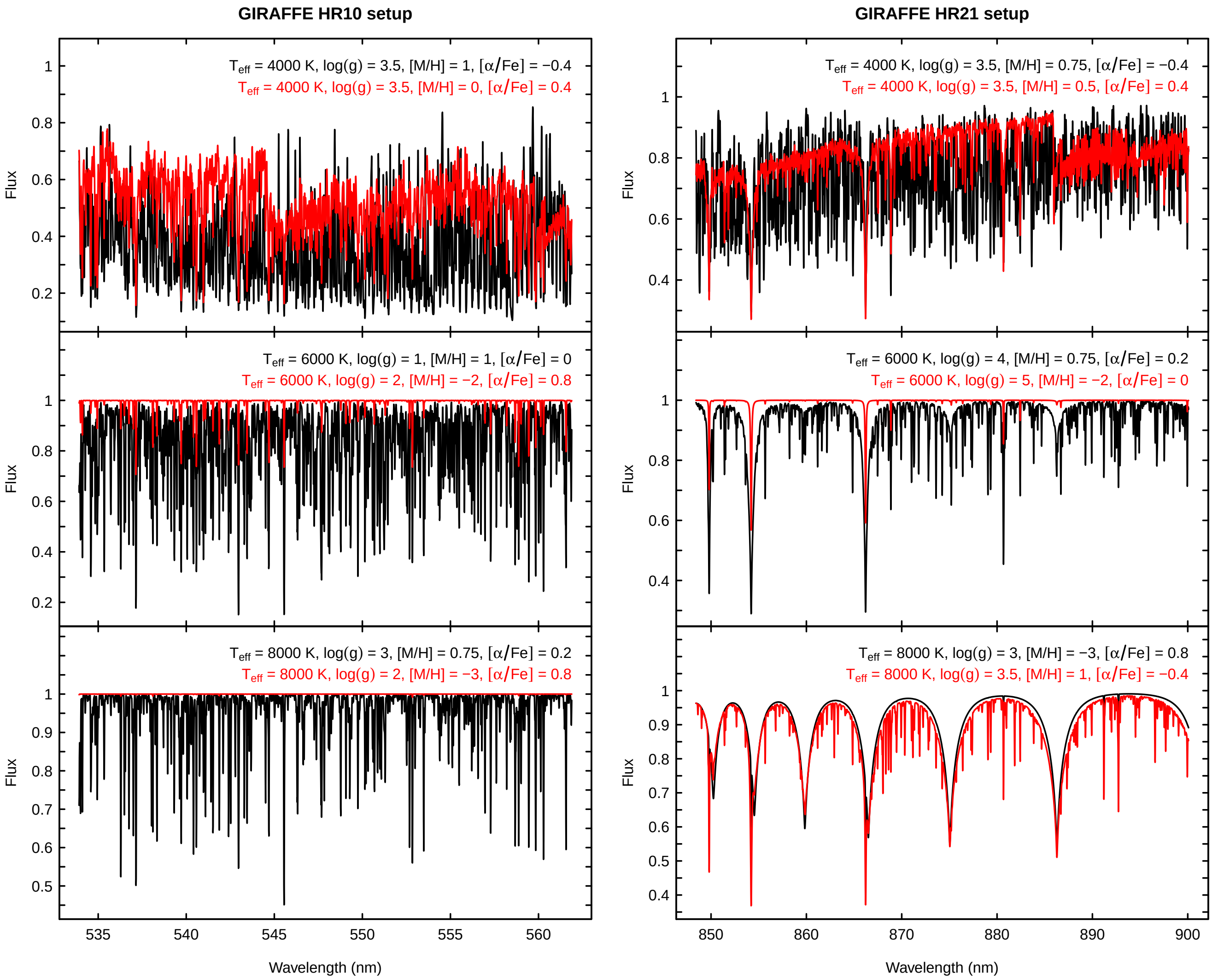


**Figure 2.** Example spectra from the nominal GIRAFFE HR10 setup (left) and the nominal GIRAFFE HR21 setup (right).

mension (Jain et al. 2000), the dataset should contain $10^{2798}$ spectra. In most real case applications, 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 validity of the results obtained with the selected dataset, we used a second dataset which is based on the same grid but covering a different wavelength. 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.

The conclusions drawn from the set of experiments described below depend on the restricted range of physical parameters, wavelengths, and spectral resolution adopted in the dataset, 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 reddest regions of the Hertzsprung-Russell diagram, where spectra are dominated by molecular absoption bands, our results cannot be expected to apply in general.

## 4 COMPARISON OF SPECTRUM COMPRESSION TECHNIQUES AND OPTIMAL RATES

We investigate the utility of six dimensionality reduction 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 in terms of the atmospheric parameter estimation errors. As a result of these

**Table 1.** Summary of the parameters analysed for the dimensionality reduction 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 |
| KPCA | $\sigma$ | [0.0001 - 0.01] | 0.001 |

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 artificial neural networks (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 SVMs 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 $[\alpha/Fe]$) prediction as measured by the root mean squared error (RMSE) (equation (1)) in out-of-sample validation experiments.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{T}_i - T_i\right)^2} \tag{1}$$

where $\hat{T}_i$ is the predicted atmospheric parameter ($T_{\rm eff}$, log $g$, [M/H] or $[\alpha/Fe]$) and $T_i$ is the target value.

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 reduction 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 (Fig. 4):

(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. Table 1 presents the values that were evaluated, 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) are fine-tuned to minimize the prediction error of the atmospheric parameter ($T_{\rm eff}$, log $g$, [M/H] or $[\alpha/Fe]$).

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

(iv) Obtain atmospheric parameter estimations by applying the SVM models trained in step (ii) to the test cases obtained in step (iii).

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

The procedure described above is repeated for different SNR regimes in order to study the dependency of the estimation performace 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.

### 4.1 Results

First, we compare the performance of the dimensionality reduction techniques described in section 2 using noise-free synthetic spectra as well as degraded spectra with SNR levels of 100, 50, 25 and 10. Figures 5 to 8 show the RMSE obtained with the evaluation set of the first dataset (the 34% of the full set of spectra that was not not used to define the compression transformation or to train SVM models).

Inspection of the figures reveals that the best strategies to compress the spectra are kernel PCA and ICA, with ICA outperforming 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 dimensionality reduction 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 –or even to other dimensionality reduction techniques– could be explained by the tuning process that was carried out to determine the optimal DM parameter values. Thus, a smaller grid size of tuning parameters would improve the DM performance. In the case of wavelets, it seems that we are eliminating valuable spectra information instead of additive noise by removing the high frequency coefficients.

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. DLA achieved the lowest prediction errors for the unrealistic noise-free data (not shown here for the sake of conciseness). Furthermore, this technique yields the best performance for the highest compression rates (two or five dimensions) when estimating $T_{\rm eff}$ and log $g$. However, it was outperformed by most other techniques for almost any other compression rate. PCA and DMs yield similar $T_{\rm eff}$ prediction errors in the high SNR regime, but DMs 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 merit 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

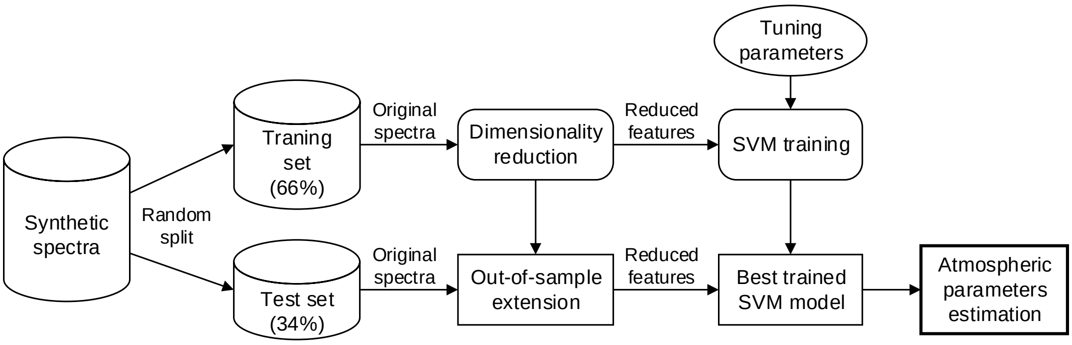


**Figure 3.** Process flow chart for investigating the performance of the selected dimensionality reduction techniques.

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 (shown in Fig. A1 in Appendix A). These results evidence that for a given sample size (the number of spectra in this particular application) 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 ($T_{\rm eff}$, log $g$, [M/H] or $[\alpha/Fe]$) predictions degrades for lower SNR. However, RMSE errors were relatively low even for low SNR ($\sim$ 10).

### *4.1.1 Second dataset analysis*

The same analysis was carried out by using a dataset that contains a wider wavelength range (almost twice as wide than the first dataset). Figure 9 and table 3 show the results obtained for $T_{\rm eff}$ with the evaluation set.

Some of our previous conclusions are confirmed by these results:

- Kernel PCA and ICA are revealed to be the best compression techniques for stellar spectra.
- Wavelets combined with SVM models have the highest errors and they are outperformed by PCA in most of the parameter space.
- DLA performed best for both noise-free data and the highest compression rates (two to five dimensions).

However, there are also some differences:

- For lower SNR data, the optimality criterion translates into retaining fewer components. This fact was identified by (Bailer-Jones et al. 1998) in the case of compression of spectra with PCA. Our results extend this conclusion to other compression techniques when the wavelength range is higher than 300 Å.
- In the high SNR regimes, RMSE errors were lower than those obtained with the first dataset. However, the performance is considerably worsened for the lowest SNR explored in this work (SNR=10).

## 5 OPTIMAL TRAINING SET SNR

In the second stage, we study the optimal match between the training set SNR and that of the spectra for which the atmospheric parameter predictions are needed (in the following, the prediction 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. This amounts to $25 \times 8 = 200$ datasets, plus the noiseless dataset. We create the 25 noise realisation to estimate the variance of the results. 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 test set SNR degrades. The model performances were evaluated using 10-fold cross validation as follows:

(i) The noiseless dataset is replicated $25 \times 8$ 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.

[1] The best performance for $T_{\rm eff}$, log $g$ and [M/H] was obtained with DLA, while best performance for $[\alpha/Fe]$ was obtained with ICA.

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

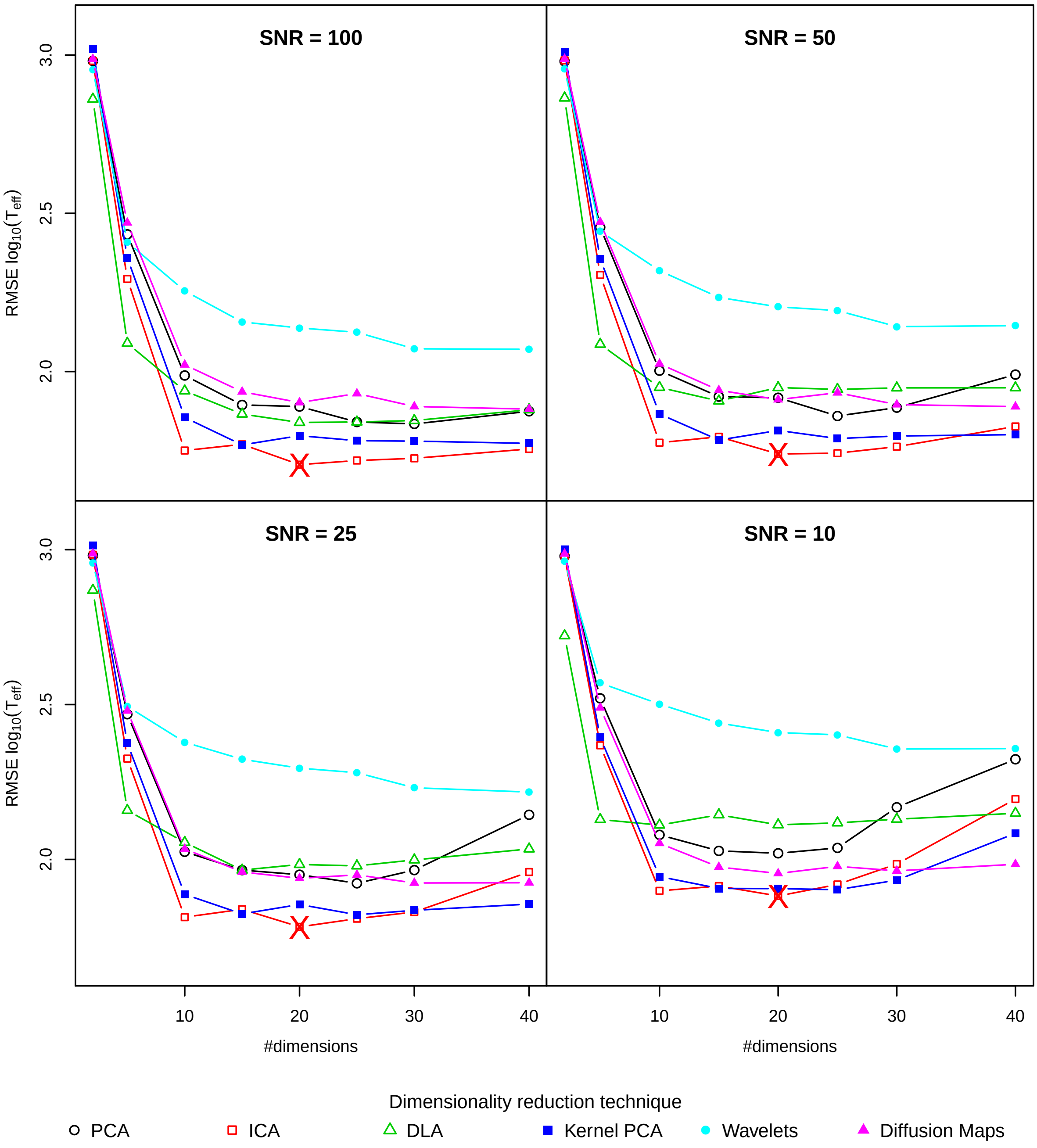


**Figure 4.** Temperature estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

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

| SNR | Method | Nr. Dim. | RMSE $T_{\rm eff}$ (K) | RMSE log $g$ | RMSE [M/H] (dex) | RMSE $[\alpha/Fe]$ (dex) |
|---|---|---|---|---|---|---|
| $\infty$ | 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 |

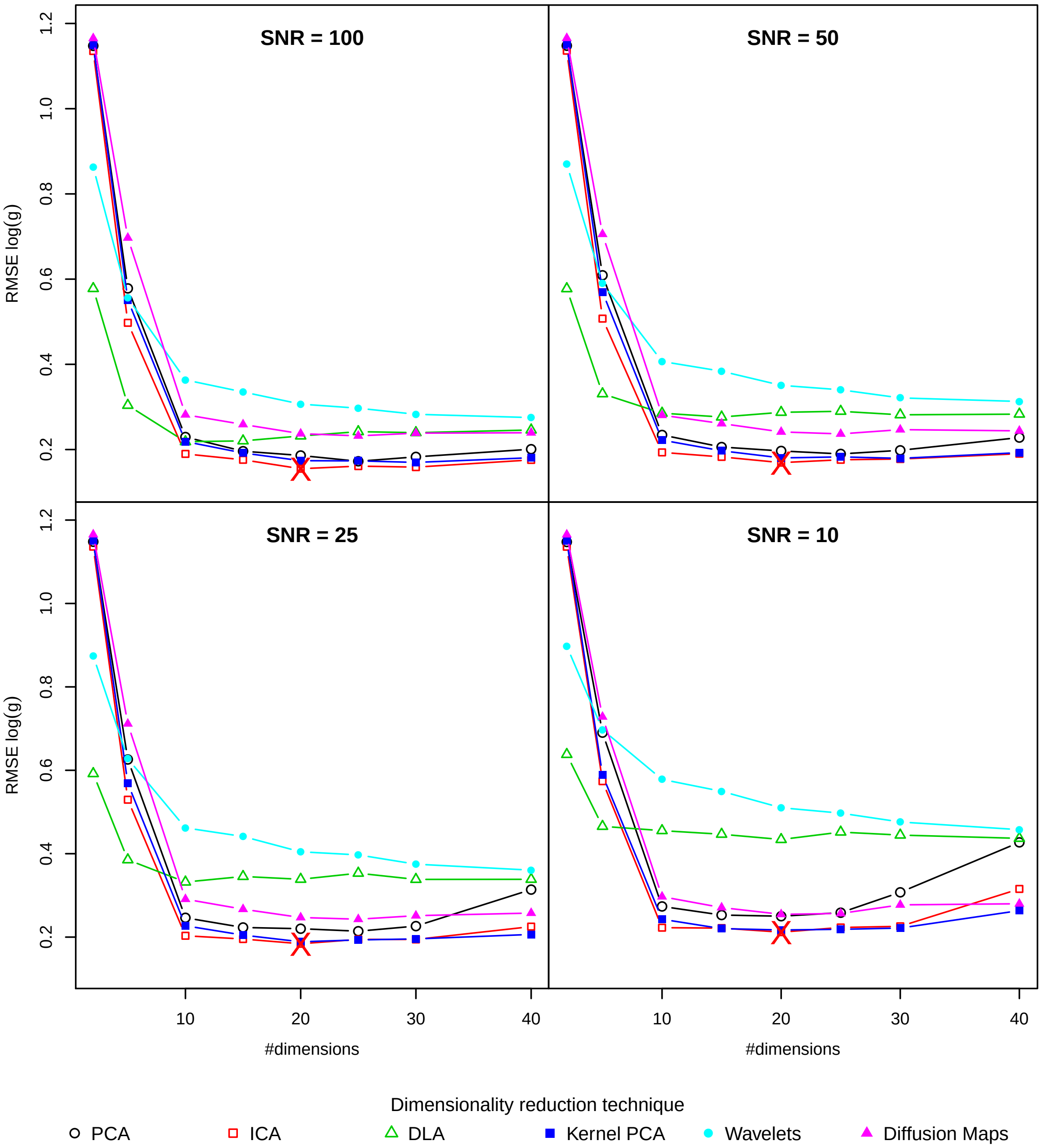


**Figure 5.** Surface gravity estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

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

| SNR | Method | Nr. Dim. | RMSE $T_{\rm eff}$ (K) |
|---|---|---|---|
| $\infty$ | DLA | 15 | 12.58 |
| 100 | ICA | 20 | 32.69 |
| 50 | ICA | 20 | 49.18 |
| 25 | ICA | 15 | 82.36 |
| 10 | ICA | 10 | 202.39 |

### 5.1 Results

Fig. 10 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:

- This analysis yields the very important consequence that models trained with noise-free spectra are not adequate

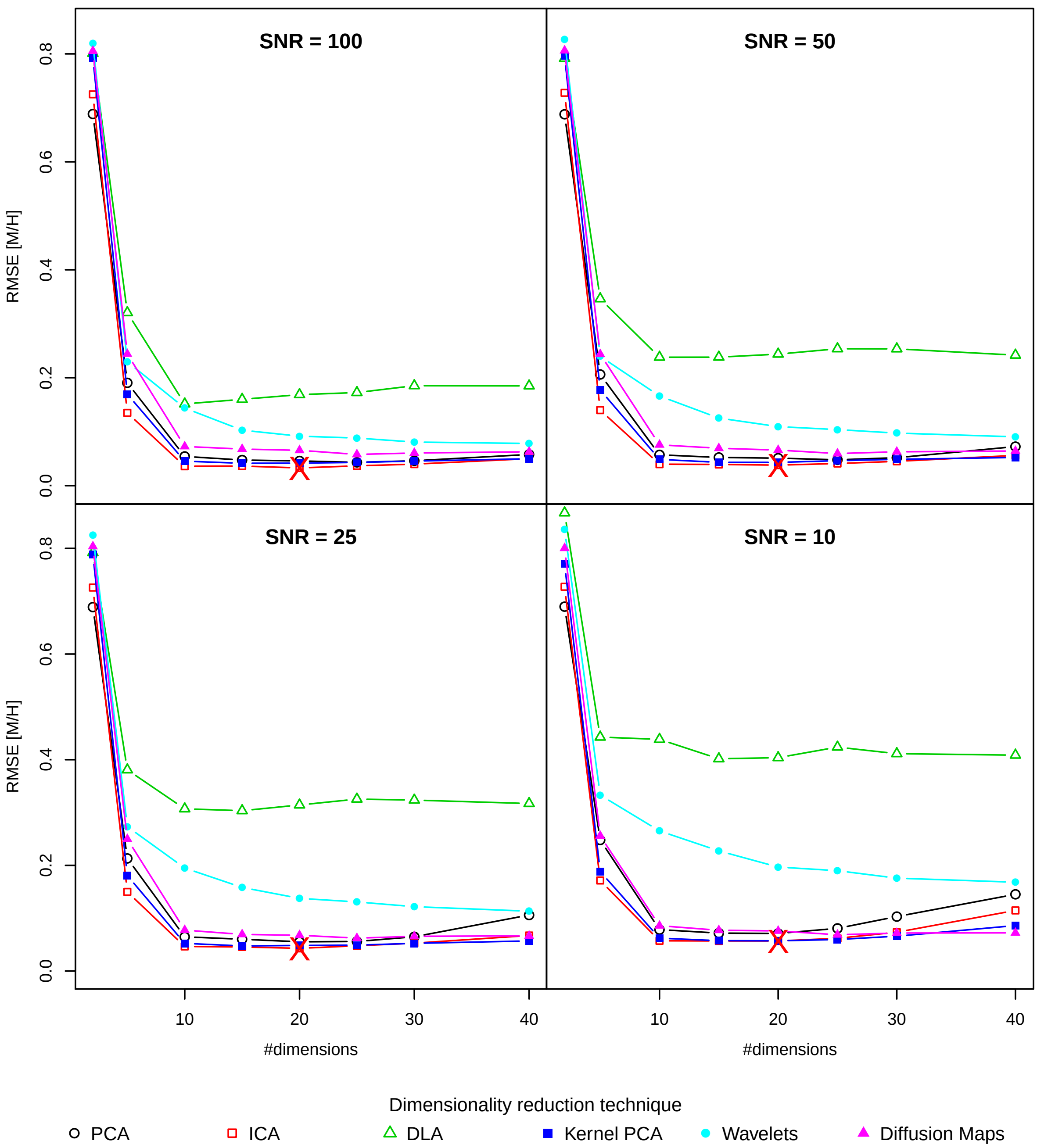


**Figure 6.** Metallicity estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

to estimate atmospheric parameters of spectra with SNRs up 50/75, and are unnecessary for $T_{\rm eff}$, log $g$ and $[\alpha/Fe]$ in contexts of even higher SNRs. Only the [M/H] regression models slightly benefit from training with noiseless spectra if the test spectra are in the SNR≥ 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).

- Only one ICA+SVM model trained with SNR of 25 would be enough to estimate the surface gravity for spectra of all SNRs with the best performance.

- 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 with the best performance.

- For evaluation spectra with SNR≥ 100, there are minimal differences in the precision achieved by models trained with spectra of SNR≥ 50.

- For evaluation sets with $100 \geq SNR > 10$, the best accuracy is obtained with the model constructed from spectra with SNR of 50 (except in the case of log $g$, where the

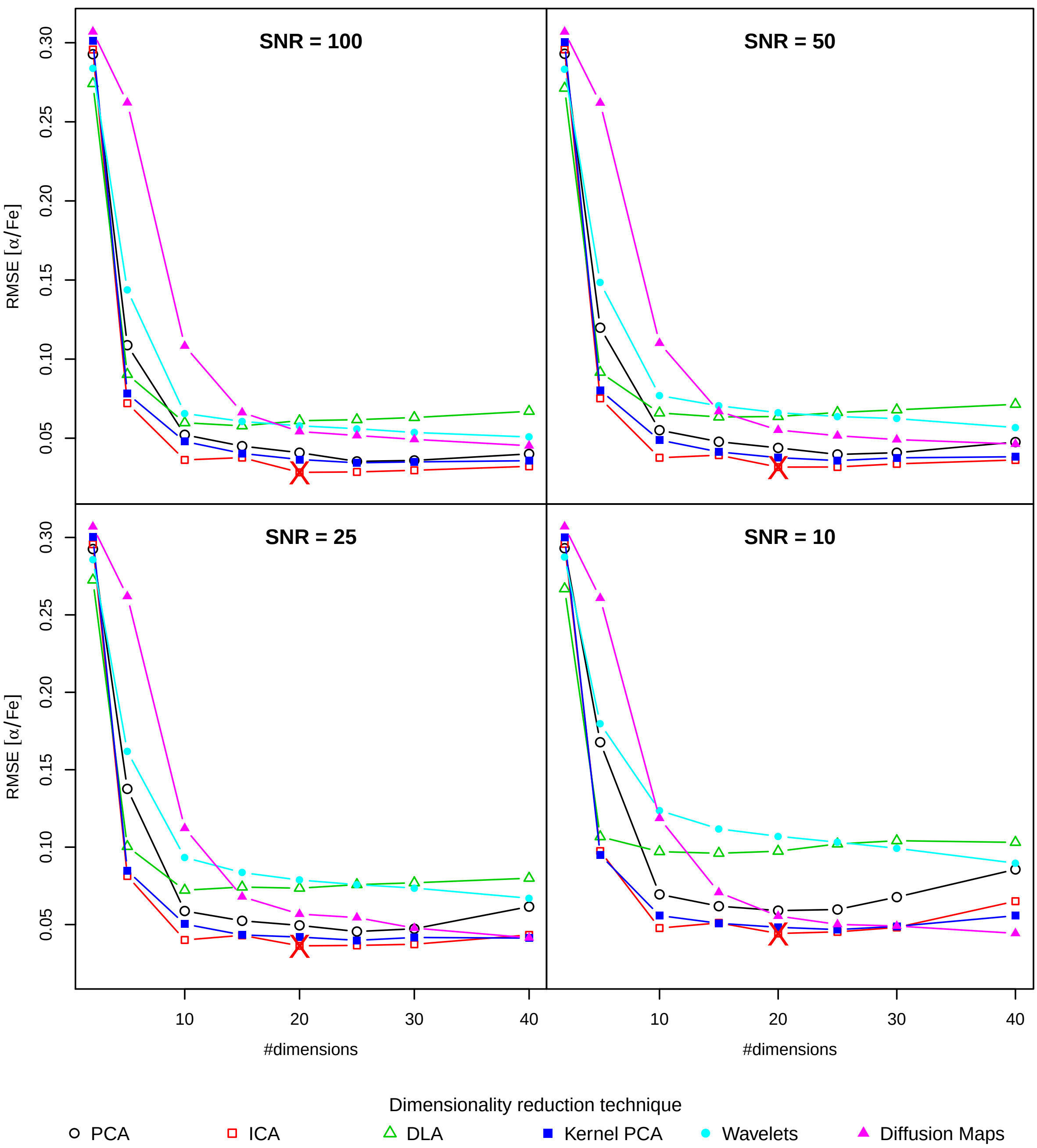


**Figure 7.** $[\alpha/Fe]$ estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

SNR=25 training set outperforms SNR=50 as noted above, but the difference is small).

- For SNR lower than 10, the model with best generalisation performance is that trained with SNR equal to 10 for $T_{\rm eff}$ and [M/H].

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: the one trained with SNR=50 for SNR≥25 and the one trained with SNR=10 for spectra with SNR≤ 10. For the prediction of surface gravities, the SNR=25 model is sufficient for any spectrum of whatever SNR. For the prediction of the ratio between the alpha-elements and iron, the SNR=50 model is sufficient for any spectrum of whatever SNR.

### *5.1.1 Second dataset analysis*

The optimal training SNR was also evaluated with the dataset that contains a wider wavelength range. In a similar

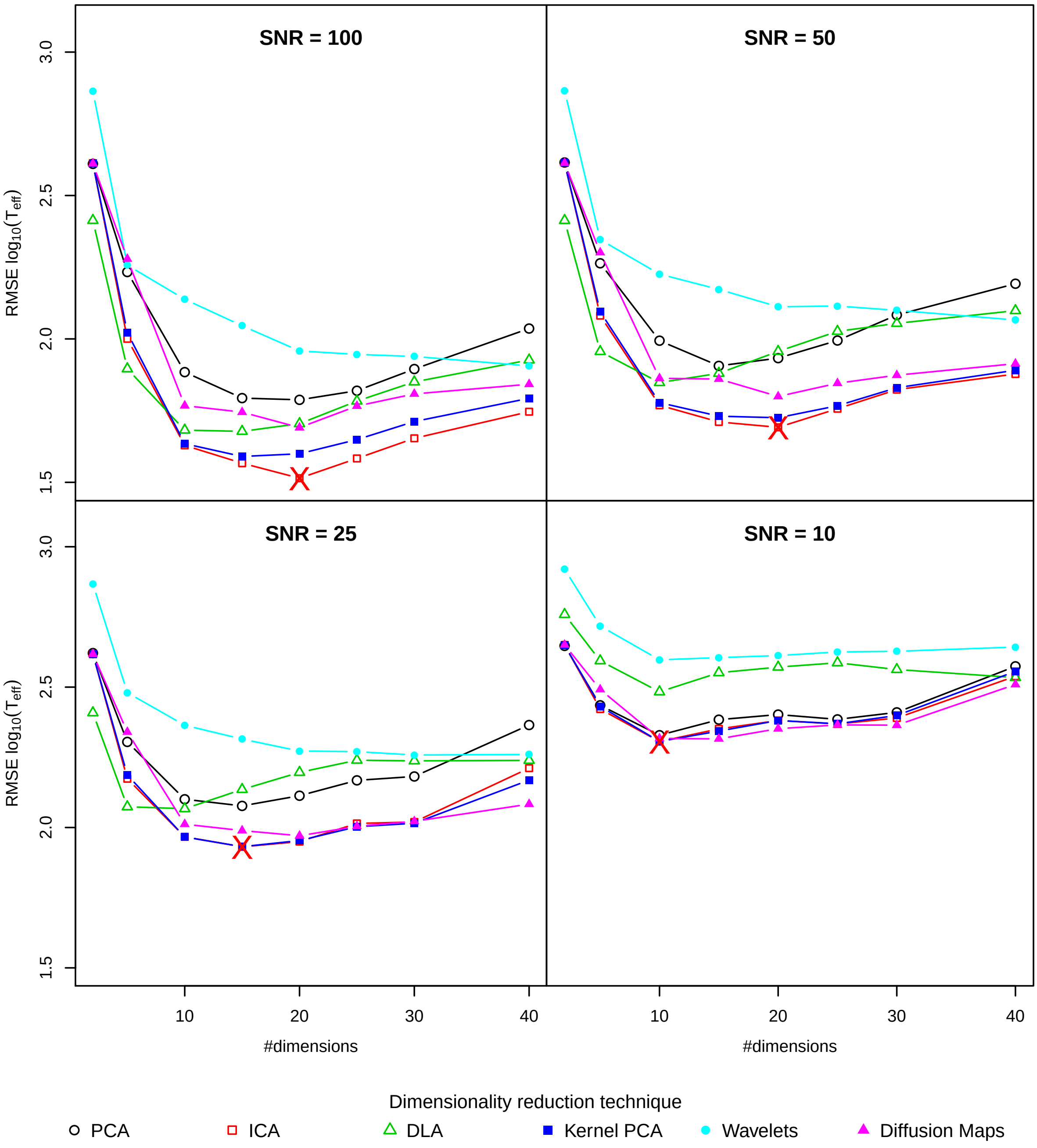


**Figure 8.** Temperature estimation errors as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

way to Fig. 10, Fig. 11 shows the results obtained for the prediction of $T_{\mathrm{eff}}$.

Again, it is possible to 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 $T_{\mathrm{eff}}$ of noise-free spectra, and completely useless in any other SNR regime. This effect is much more evident with this dataset than with the first one.
- Also, it is clear that, if the test spectra are in the SNR>25 regime, the $T_{\mathrm{eff}}$ regression models do not benefit at all from training with SNR≤25.
- In this case, for evaluation spectra with SNR≥100, differences in the precision achieved by models trained with spectra of SNR≤50 are easier to notice. The best option is one ICA+SVM model trained with SNR of 125.

In summary, a model trained with SNR=50 would be enough to estimate $T_{\mathrm{eff}}$ of spectra with 100≥SNR≥25. A second model trained with SNR=10 would be adequate for spectra with SNR≤10. And, differently to the previous results, a third model trained with SNR=125 would be necessary for SNR≥100.

Finally, it is noteworthy that RMSE errors are slightly

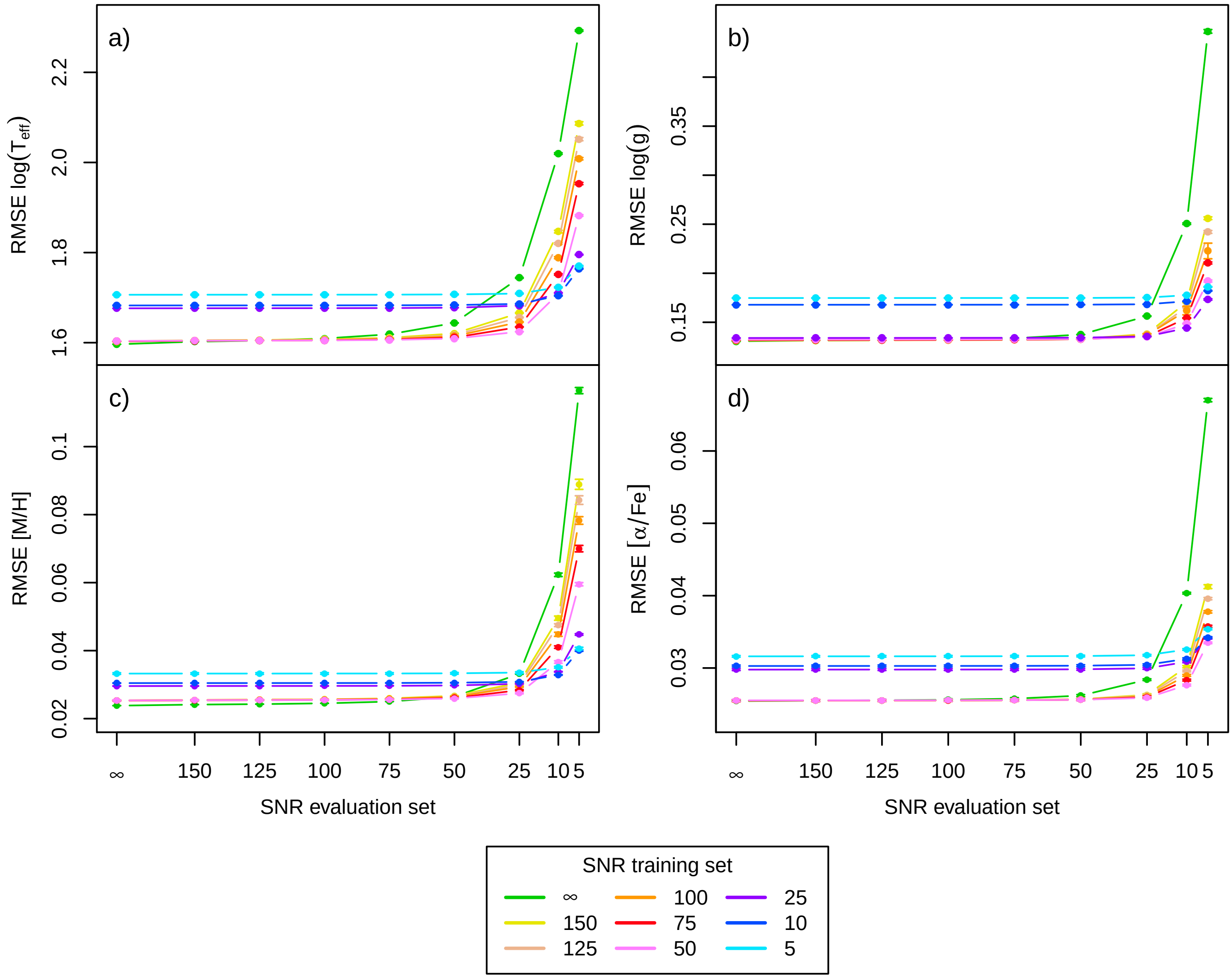


**Figure 9.** Estimation errors as a function of the SNR of the evaluation set for $T_{\rm eff}$ (a), $\log(g)$ (b) and [M/H] (c) and $[\alpha/Fe]$ (d). Each line corresponds to a model trained with a specific SNR

lower than those from the first dataset in the high SNR regime, and that RMSE errors are higher than those form the first dataset in the low SNR regime. Since the wavelength range of the second dataset is almost twice as wide than the first dataset, this results illustrate the effect of both the input dimension and the additive noise introduced in spectra on the the *compression* and *denoising* ability of the investigated dimensionality reduction techniques.

**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 |

## 6 TRAINING SET DENSITY

Finally, we carry out an analysis of the effect of the training set grid density over the regression performance. To do this, six new 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 a variable step-size between 50 K and 250 K. The other grid parameters were established as follows: the log $g$ were regularly sampled from 1 to 5 in 0.5 steps and both [M/H] and $[\alpha/Fe]$ were set equal to zero for simplicity. 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 of different SNR levels (100, 50, 25, 10).

We evaluated the performance of the SVM regression models using 10-fold cross validation. Figures 12 and 13 presents the $T_{\rm eff}$ estimation errors obtained with the different grid densities and the two optimal training set SNRs (50

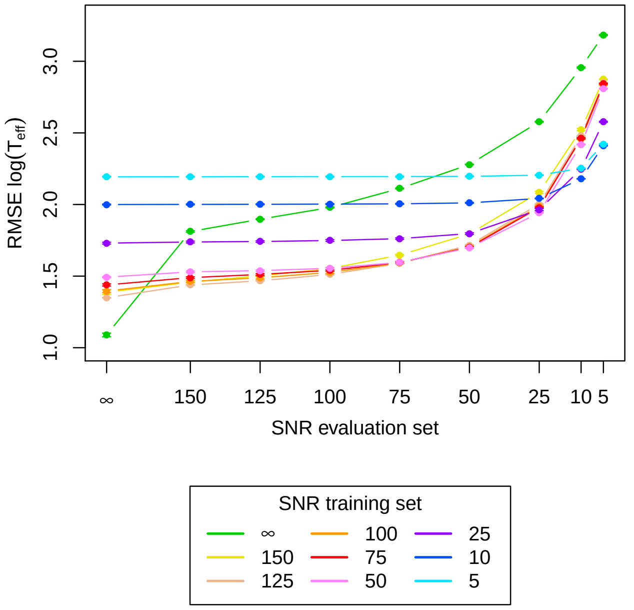


**Figure 10.** Estimation errors as a function of the SNR of the second dataset evaluation set for $T_{\rm eff}$. Each line corresponds to a model trained with a specific SNR

and 10) found in the previous Section. The main conclusions of the analysis of these figures are the following:

• As expected, estimation errors increase when the grid density decreases.

• Overall, the accuracy obtained against the grid density is more variable when the number of dimensions retained increases.

• PCA and ICA show a similar behaviour with the grid density variation.

• The grid density appears to have less effect over the performance of Wavelets and diffusion maps.

## 7 CONCLUSIONS

**Here we should discuss the validity of our conclusions. The validity depends on our assumptions and the experiments carried out. For example, they are based on SVM models with radial kernel functions and the implications should be stressed. Also the spectra were trimmed in a wavelength range: how is this range? Also compare our RMSE with those in the bibliography, for example, those of MATISSE, the Gaia-ESO results...**

**Discuss the impact of including other metallicities and alpha abundances in the training set. To be analysed in a subsequent paper.**

**Discuss the relationship between the methods tested here and those in the bibliography.**

## 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
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
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
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, 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
Mishenina T. V., Bienaymé O., Gorbaneva T. I., Charbonnel C., Soubiran C., Korotin S. A., Kovtyukh V. V., 2006, A&A, 456, 1109
Muirhead P. S., Hamren K., Schlawin E., Rojas-Ayala B., Covey K. R., Lloyd J. P., 2012, ApJ, 750, L37
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

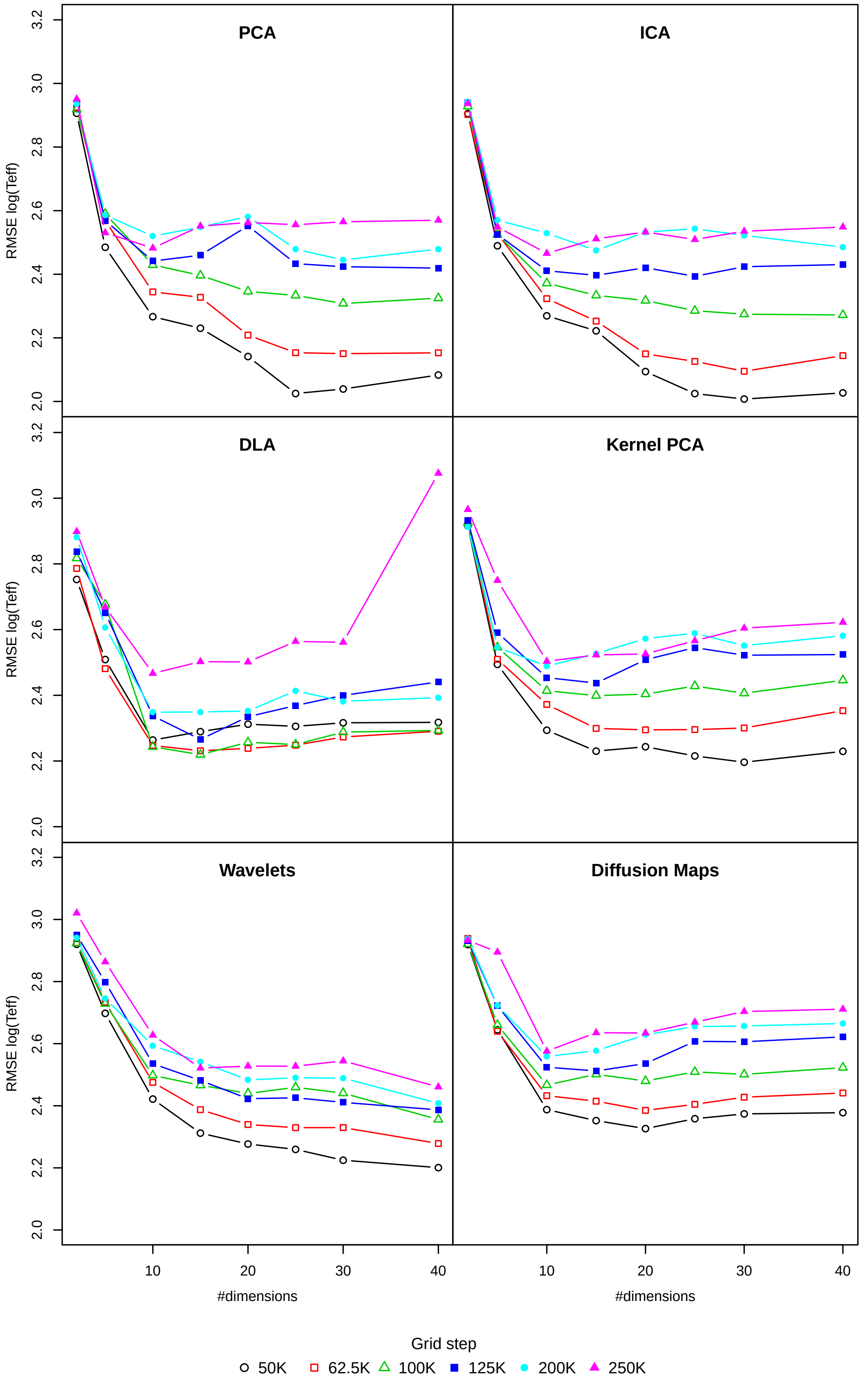


**Figure 11.** Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 50)

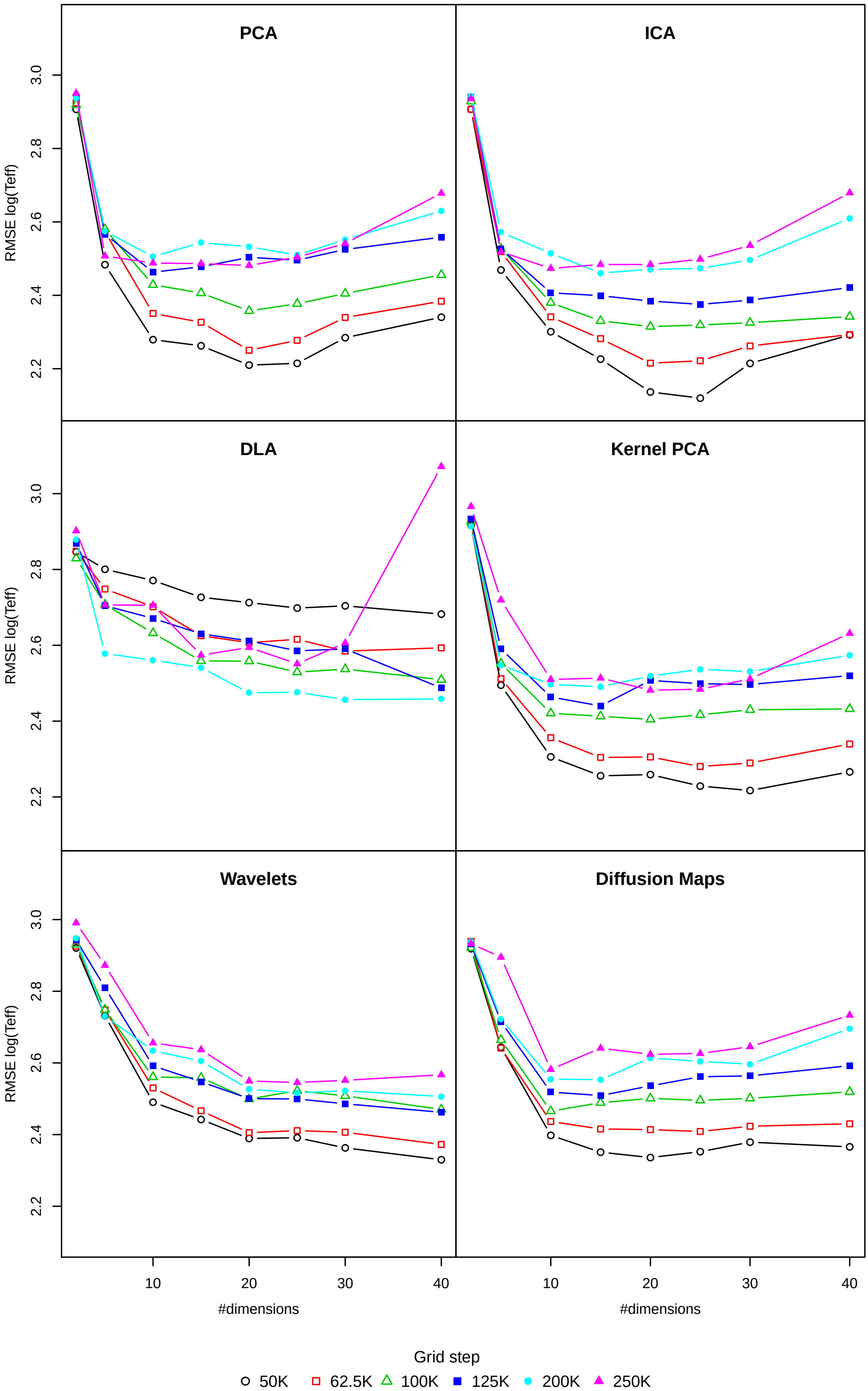


**Figure 12.** Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 10)

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)

Roweis S., Saul L., 2000, Science, 290(5500), 2323

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

Vanderplas J., Connolly A., 2009, The Astronomical Journal, 138, 1365

Walker M. G., Olszewski E. W., Mateo M., 2015, MNRAS, 448, 2717

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

## APPENDIX A: ADDITIONAL FIGURES

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

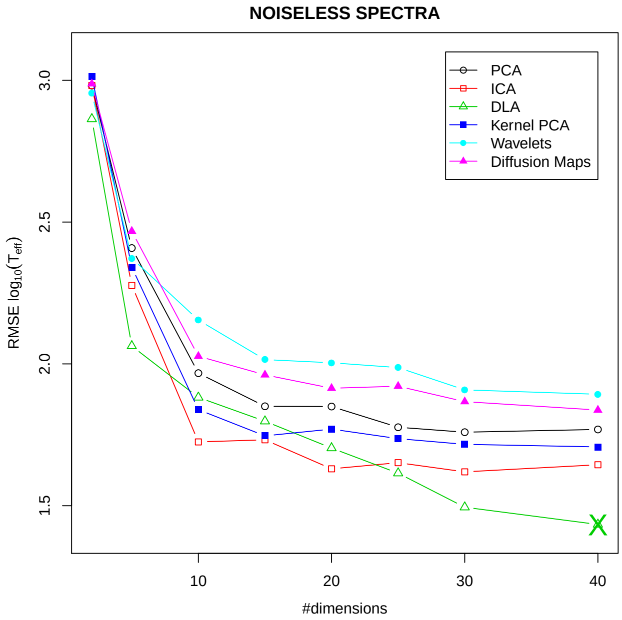


**Figure A1.** Temperature estimation error as a function of the number of dimensions used for data compression, for synthetic spectra

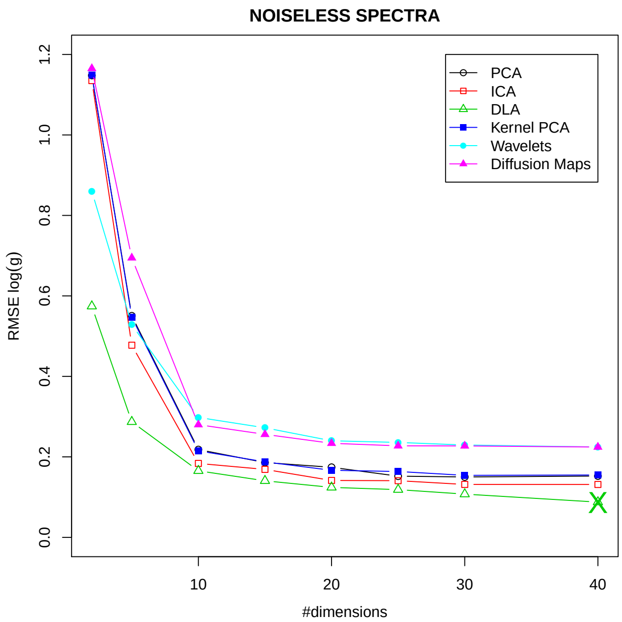


**Figure A2.** Surface gravity estimation error as a function of the number of dimensions used for data compression, for synthetic spectra

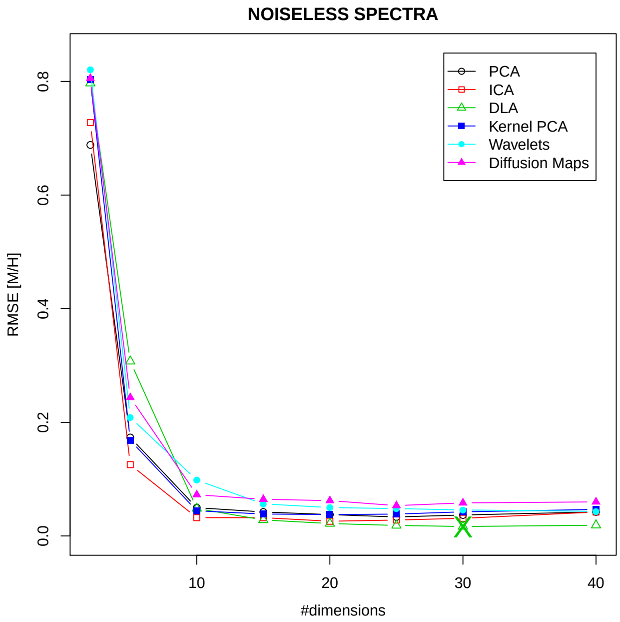


**Figure A3.** Metallicity estimation error as a function of the number of dimensions used for data compression, for synthetic spectra

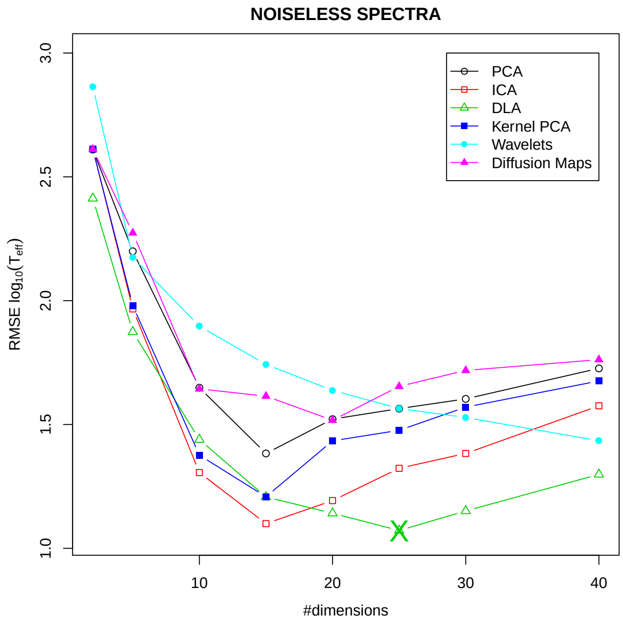


**Figure A5.** Temperature estimation error as a function of the number of dimensions used for data compression, for synthetic spectra from the second dataset

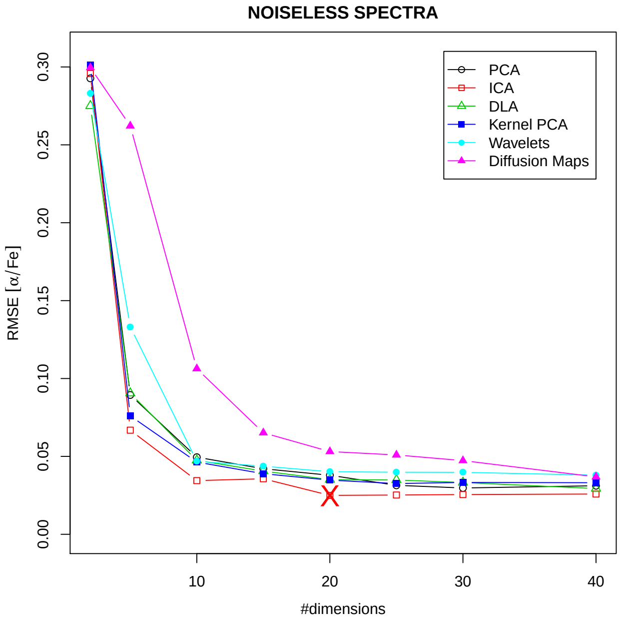


**Figure A4.** $[\alpha/Fe]$ estimation error as a function of the number of dimensions used for data compression, for synthetic spectra

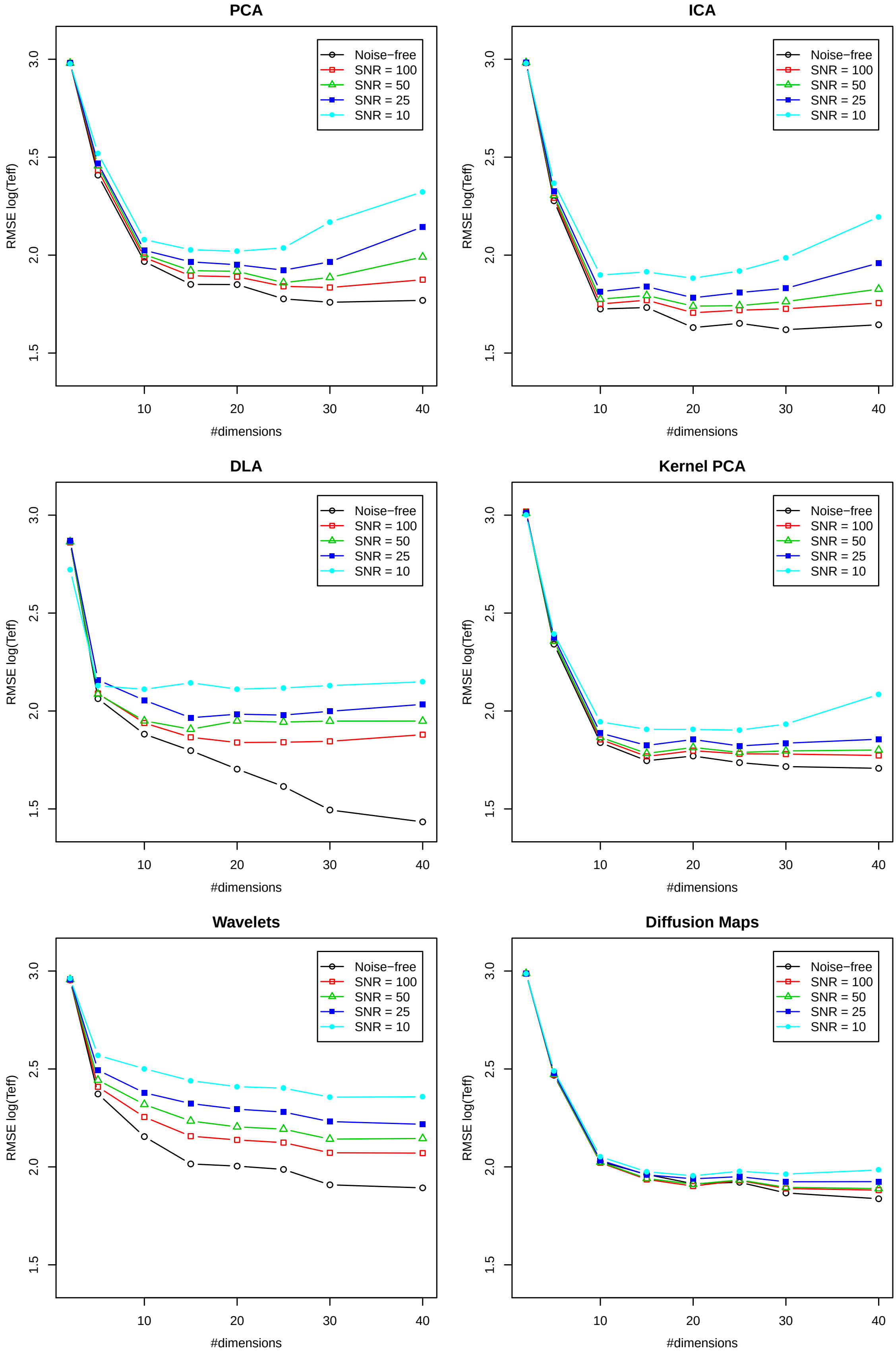


**Figure A6.** Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

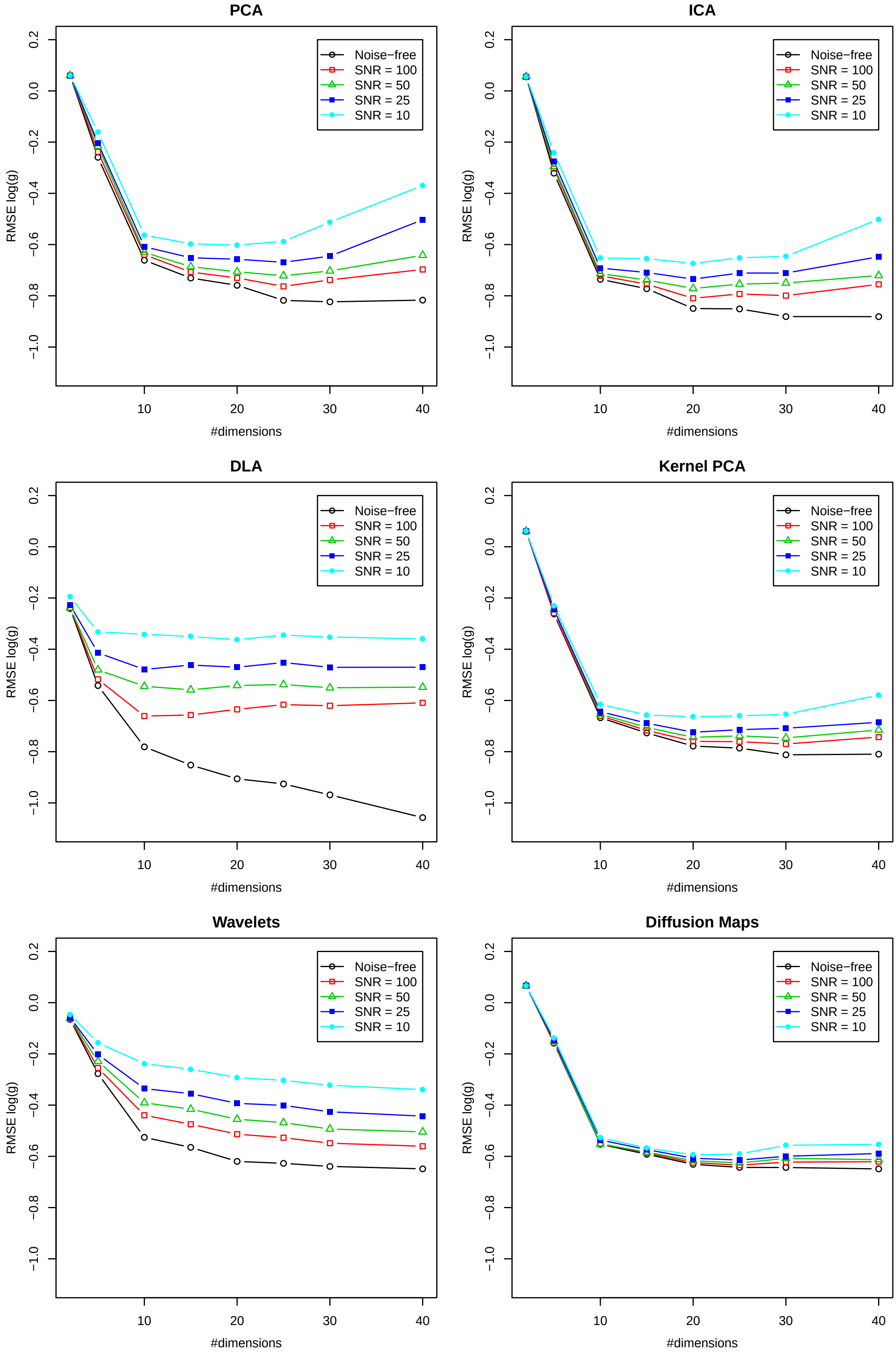


**Figure A7.** Surface gravity estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

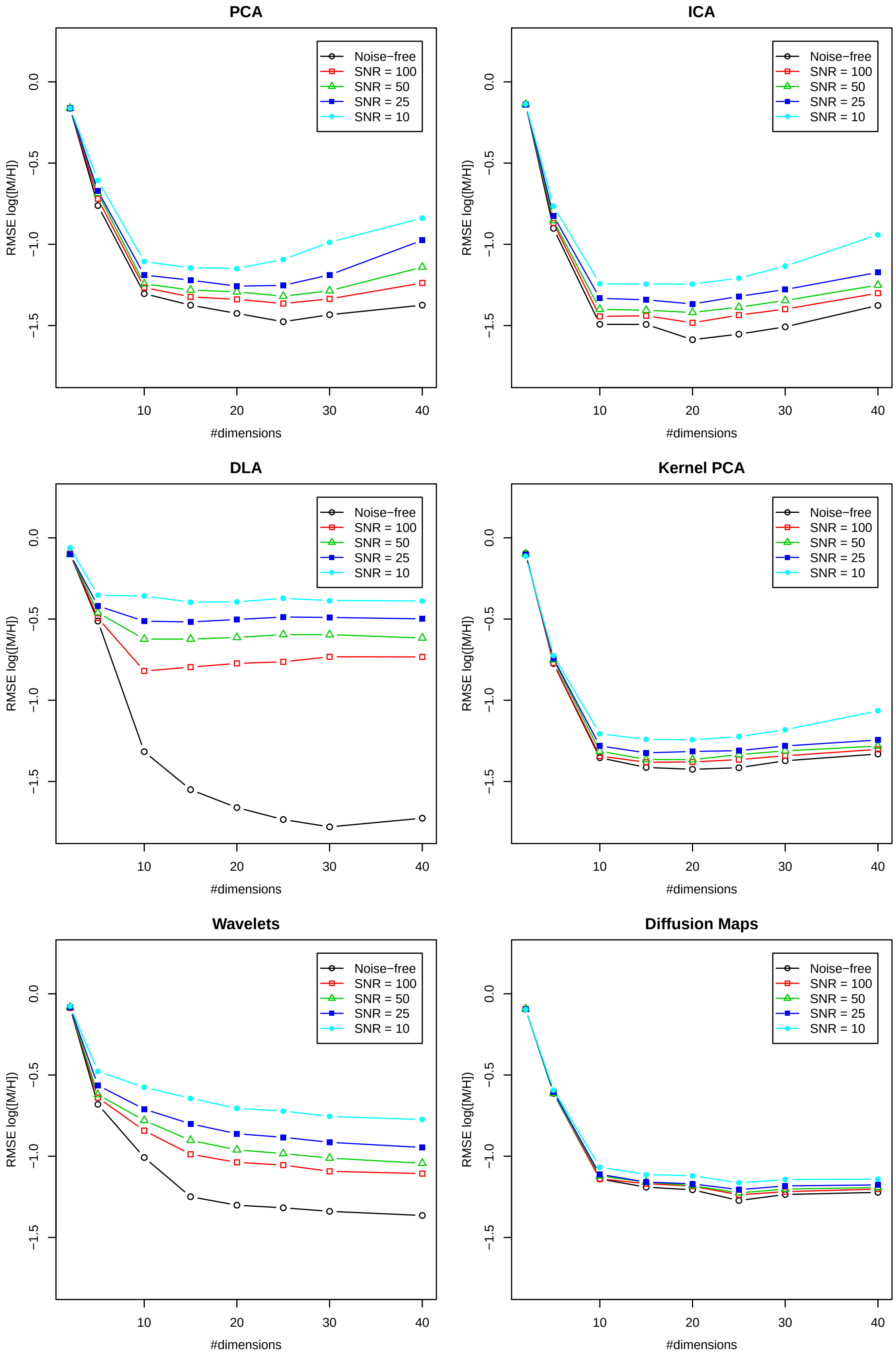


**Figure A8.** Metallicity estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

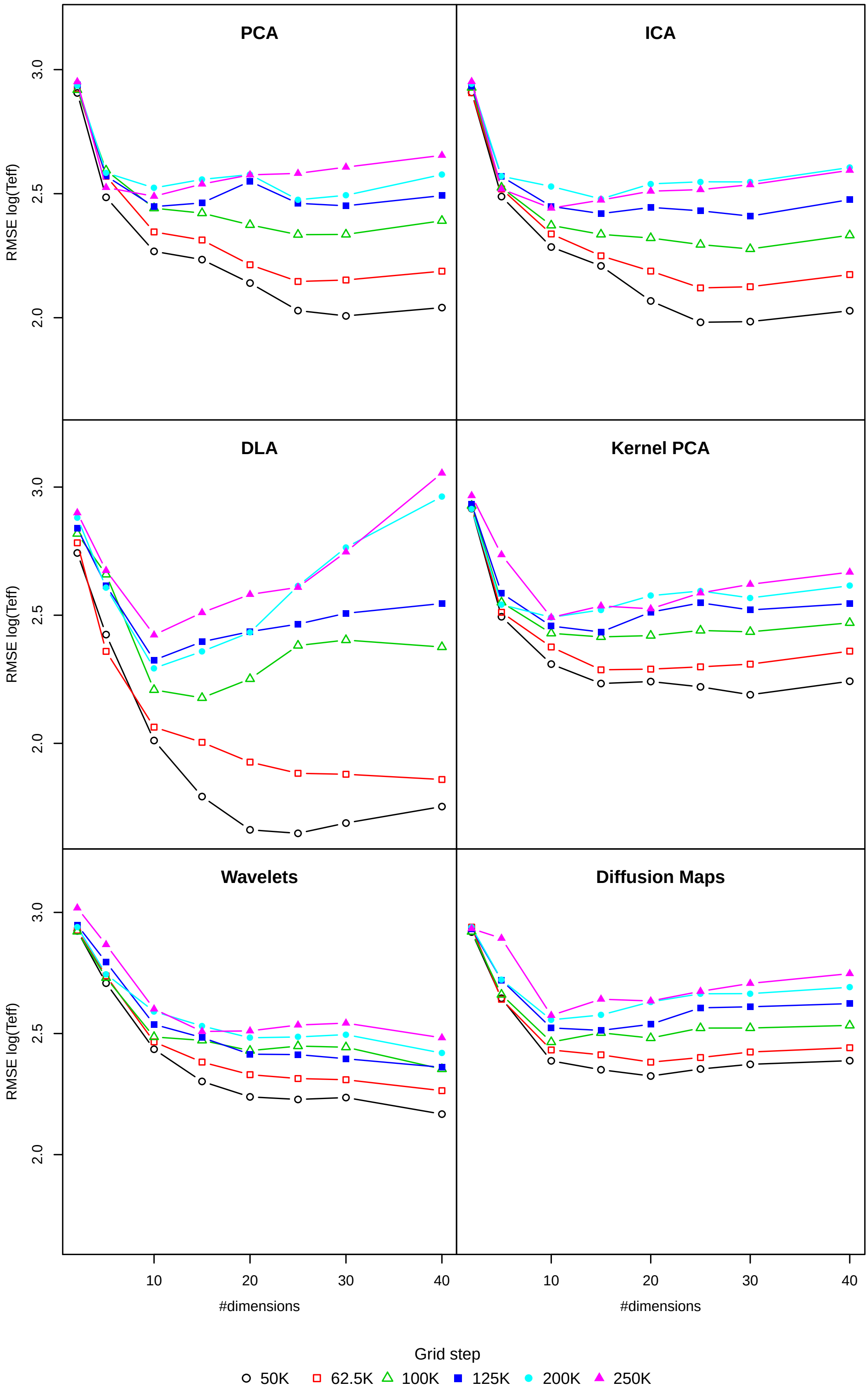


**Figure A9.** Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (Noise-free spectra)

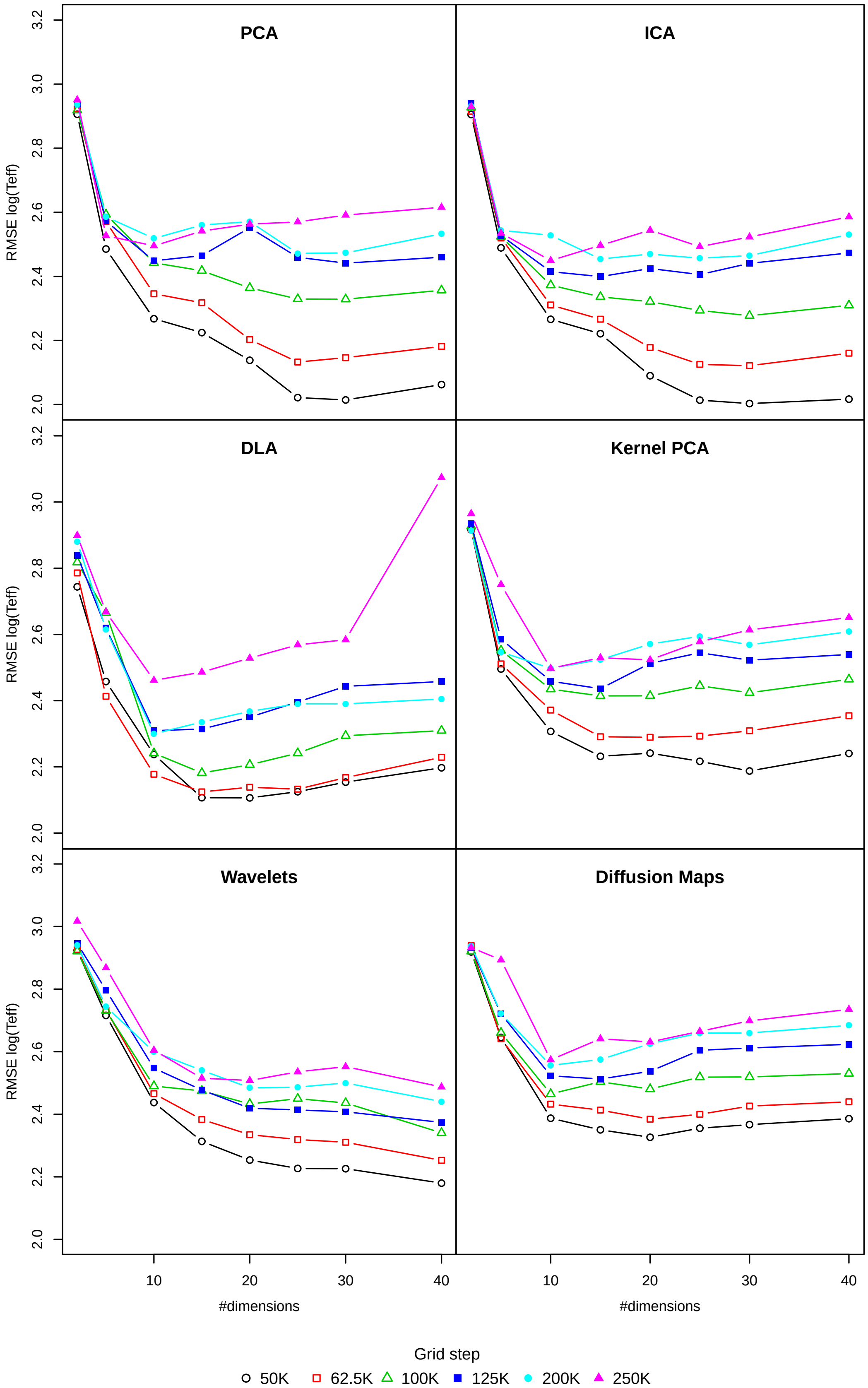


**Figure A10.** Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 100)

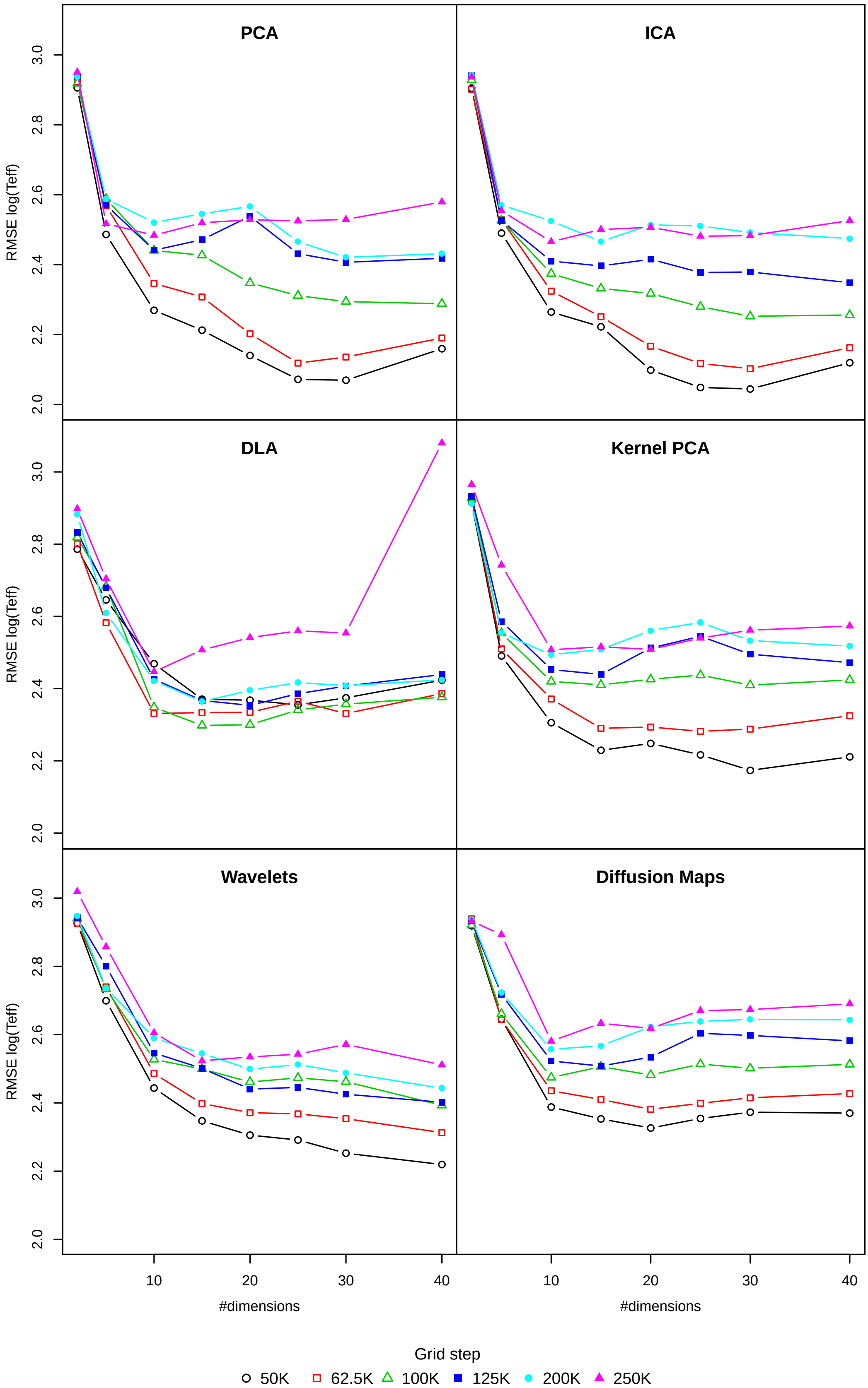


**Figure A11.** Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 25)