MNRAS **000**, 1–11 (2015) Preprint 15 December 2015 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 ($\log g$) and metallicity ([M/H]). 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)

© 2015 The Authors

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 (Allende Prieto 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 normalization can be mitigated by projecting the observed spectra onto bases of functions spaces such as in the wavelet or Fourier decompositions (see **?** or **?** 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. **?**, **?**, or **?**). 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 performance in atmospheric parameters estimation. The investigation is performed by an empirical evaluation of the selected techniques on specifically designed synthetic datasets. In 2 we review the data compression techniques evaluated in this work and their properties. In Sect. 3 we describe the numerical experiments carried out to evaluate these techniques, and in Sect. 4 we present the main results from the experiments. Finally, in 5 we summarize the most relevant findings from the experiments and discuss their validity and limitations.

2 DIMENSIONALITY REDUCTION

For the sake of computational efficiency, and thinking 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 anal-

ysis (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).

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 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.

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 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 algorithm (Zarzoso & Comon 2010) represents a simple modification of FastICA. This algorithm 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) algorithm (Cardoso & Souloumiac 1993). This approach exploits the fourth-order moments in order to separate the source signals from mixed signals. This work uses 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, one patch is built by the given sample and its neighbours. On each patch DLA preserves the local discriminative information through integrating two criteria: that the distances between the intra-class samples will be as small as possible and the distance between the inter-class samples will be as large as possible. In the second stage, each part optimisation is weighted by *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. 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.

DM 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, DM attempt to retain the pairwise diffusion distances as good 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 *localness* 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 dividing a signal into different frequency and time intervals. Wavelet transform is a popular feature technique that has been developed to improve the shortcomings of the Fourier transform. Wavelets are better for modelling than Fourier analysis because they are not loosing the space information when moving to the frequency domain.

Wavelets can be constructed from a function (named *mother wavelet*), which is confined in a finite interval. 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 using inner products to decompose any given signal like in Fourier analysis. This method offers multi-resolution analysis in time and frequency domain and it can be useful to reveal trends, breakdown points or discontinuities.

Dimensionality reduction with wavelets consists on keeping a reduced number of wavelet coefficients. There are two common ways of coefficient selection: (1) to keep the first coefficients (the signal is represented by a rough sketch, because these coefficients correspond to the low frequencies of the signal) and (2) to keep the k most significant coefficients (this yields a representation of the signal with less variance) (Li et al. 2010). The first approach is used in this work for reducing the spectra.

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 ('kernel trick'). The application of PCA in the kernel space provides Kernel PCA the property of constructing nonlinear mappings.

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.

The Gaussian kernel was used in this work as the kernel function to provide the high-dimensional space to PCA. In this case, the inverse kernel width (σ) is the parameter used to optimise the results.

3 EXPERIMENT DESCRIPTIONS

3.1 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.

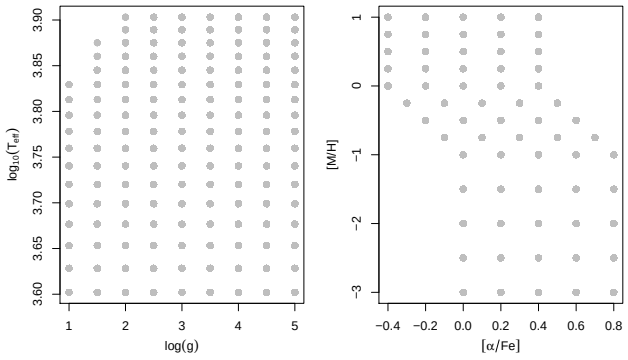

Figure 1. Coverage in parameter space of the dataset

The dataset contains a grid of 8780 synthetic high-resolution spectra (R = 19800) between 5339 and 5619 Å 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 solar abundances are those used by (Gustafsson et al. 2008). Fig. 2 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 dimension (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.

3.2 The experiments

We investigate the utility of six dimensionality reduction techniques for feature extraction with a view to improving the performance of the estimation of atmospheric parameters. Furthermore, 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 selecting the compression technique and final dimension that minimizes the atmospheric parameter ($T_{\rm eff}$, log g or [M/H]) estimation error; in the second stage we analyse the estimation error for each compression technique as a function of the SNR of the original spectrum; and finally, we select the best performing compression technique and study the generalisation performance of an SVM model trained with different sets of spectra of various SNRs.

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

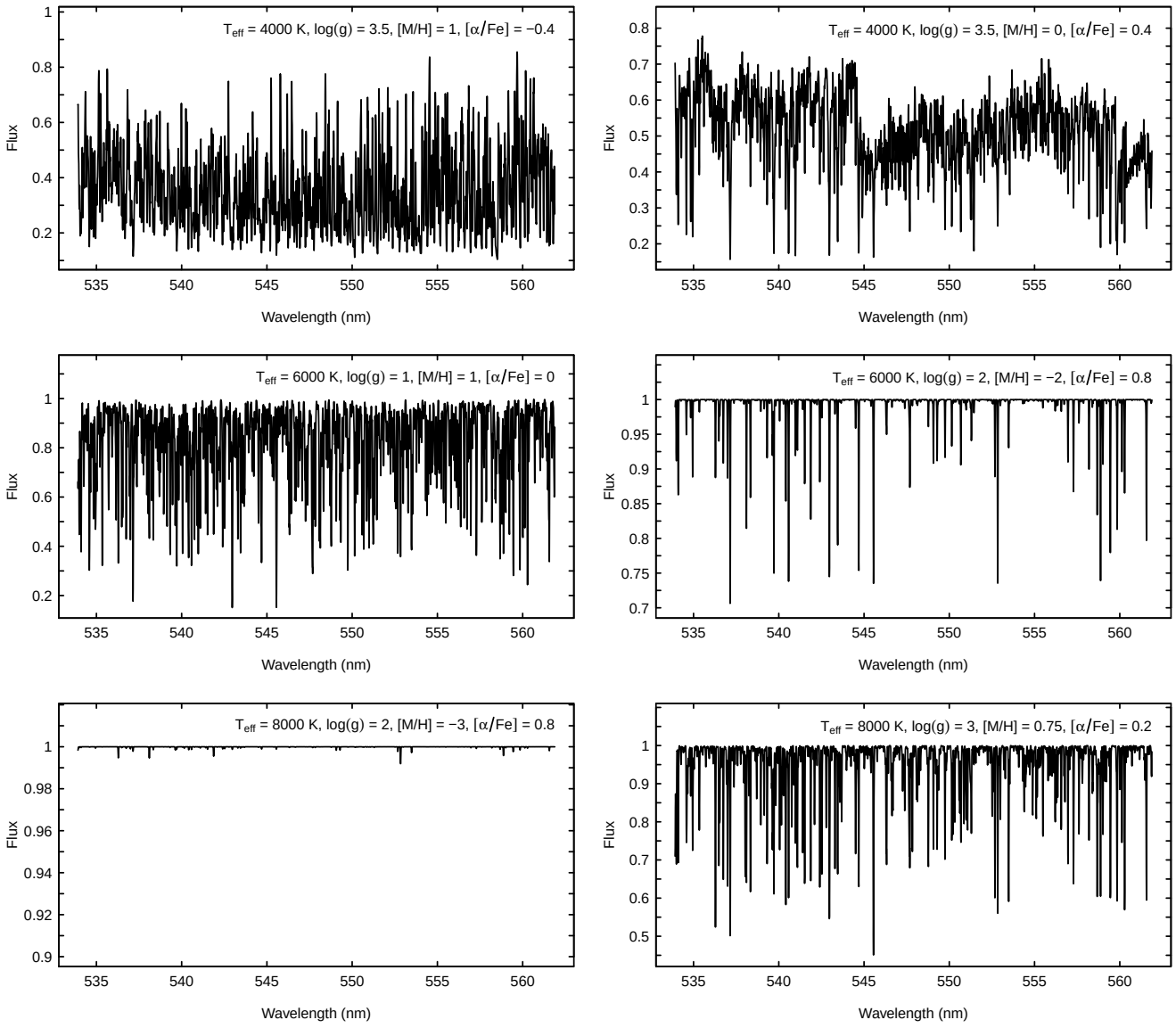

Figure 2. Example spectra from the dataset

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 or [M/H]) prediction as measured by the root mean squared error (RMSE) (equation (1)).

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

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

The dataset was randomly split into two sets, one for training (66% of the available spectra) and one for evaluation (the remaining 34%), to investigate the performance of the selected dimensionality reduction techniques on stellar spectra. 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. 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. Table 1 presents the values that were analysed in this work as well as the best parameter value obtained in each case.

(ii) Construct SVM models based on a varying number of dimensions (2, 5, 10, 15, 20, 25, 30 and 40) from the reduced space and the training set. SVM parameters (kernel size and

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
DMAP	eps.val	[0.01 - 700]	600
KPCA	σ	[0.0001 - 0.01]	0.001

noise contribution) are fine-tuned to minimize the prediction error of the atmospheric parameter ($T_{\rm eff}$, log g or [M/H]).

(iii) Project the evaluation set spectra onto the low-dimensional space computed in step 1.

(iv) Obtain atmospheric parameter ($T_{\rm eff}$, log g or [M/H]) estimations with the SVM models trained in step 2 based on the new features obtained in step 3.

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

The same procedure described above was used to analyse the robustness properties of the selected dimensionality reduction techniques against increasing noise rates. In this case, Gaussian white noise of different variances (SNR equal to 100, 50, 25 and 10) was added to the original synthetic spectra.

The optimal compression method and the optimal reduced dimension were selected according to the results of the previous experiments to investigate the optimal SNR of the training samples with a view to obtaining the best generalization performance of the atmospheric parameters estimators. Twenty-five datasets were generated for a wider range of SNR levels (150, 125, 100, 75, 50, 25, 10 and 5) and SVM models were trained to estimate the atmospheric parameter ($T_{\rm eff}$, log g or [M/H]) to assess the consistency of the results. Each SVM model was used to estimate the atmospheric parameters ($T_{\rm eff}$, log g or [M/H]) for spectra with SNR ratios different from those used during the learning phase. The performance measure of the models was evaluated using 10-fold cross validation:

(i) The dataset was split into 10 smaller sets or *folds*.

(ii) The optimal compression technique was applied to nine of the these folds, and the original spectra compressed to a space of the dimensionality found optimal in the experiments described in previous paragraphs.

(iii) A SVM model was trained using the training spectra projected onto the new space.

(iv) The tenth fold was used as the evaluation set. The spectra were projected onto the new low-dimensional space and the SVM model used to estimate the corresponding atmospheric parameter ($T_{\rm eff}$, log g or [M/H]). Then the quality of the prediction was measured.

(v) Step 4 was repeated using spectra with different SNR to that used for computing the ICA projector and for training the SVM model.

(vi) Steps 2 to 5 were repeated 10 times (using each time a different fold for evaluation) and the performance measure was calculated by averaging the values obtained in the loop.

Finally, an analysis of the effect of the grid density over the performance in atmospheric parameters estimation was carried out. To do this, six new datasets of synthetic spectra were computed by considering different grid densities. Thus, 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 equal zero. Table 2 presents the step-sizes used in this study as well as the number of synthetic spectra available in each grid. Furthermore, different SNR levels (100, 50, 25, 10) were added to the original synthetic spectra.

Table 2. 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

The performance of the $T_{\rm eff}$ models trained with the new datasets was evaluated using 10-fold cross validation.

4 RESULTS AND DISCUSSION

In this section, we present the results of the experiments described in the Sect. 3.

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. Fig. 4 to 9 show the RMSE obtained with the evaluation set (the 34% of the full set of spectra that was not not used to define the compression transformation or to train SVM models). Overall, wavelets combined with SVM models have the highest errors regardless of the number of retained dimensions, with the exception of [M/H] estimation where DLA performed worse for noisy synthetic spectra. DLA achieved the lowest prediction errors for the unrealistic noise-free data. Furthermore, the performance of this technique was the best for the highest compression rates (two or five dimensions) to estimate $T_{\rm eff}$ and log g. However, it was outperformed by the other techniques in almost any other case. PCA and diffusion maps yielded similar prediction errors in $T_{\rm eff}$ estimation. As for the estimation of surface gravity and metallicity, results obtained with PCA were better than results from diffusion maps. An important characteristic of diffusion maps combined with SVM models is that similar estimation errors were obtained regardless of the SNR of the data.

Finally, the best strategies to compress the spectra are kernel PCA and ICA, with ICA outperforming kernel PCA in most of the parameter space. RMSE errors increased 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

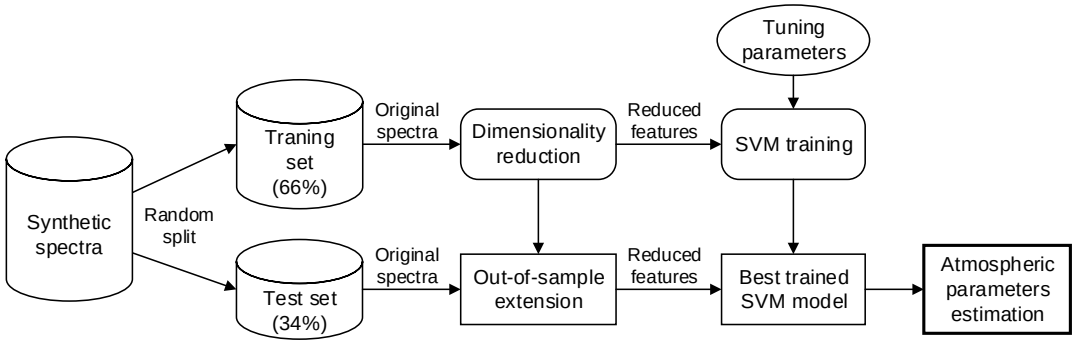

Figure 3. Process flow chart for investigating the performance of the selected dimensionality reduction techniques.

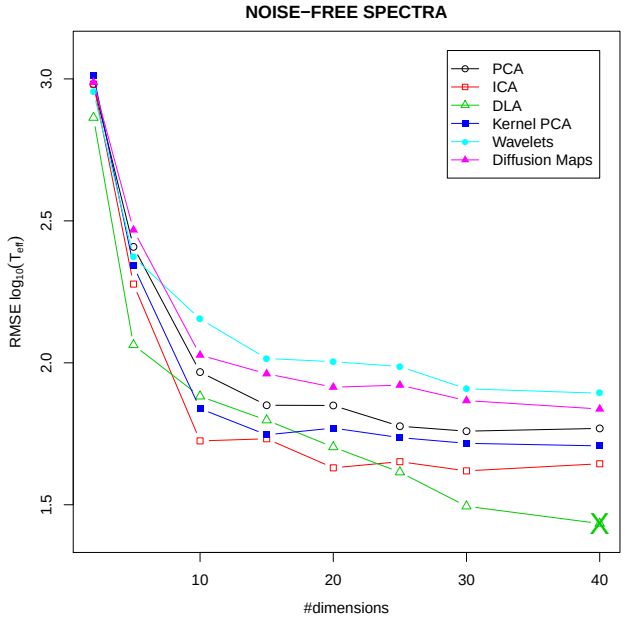

Figure 4. Temperature estimation error as a function of the number of dimensions used for data compression, for synthetic spectra

some of the nonlinear techniques used in this study, such as diffusion maps, DLA or wavelets.

Table 3 quantifies the prediction errors of the best models trained 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 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 or [M/H]) predictions degrades for lower SNR (see Fig. 10, 11 and 12). However, RMSE errors were relatively low even for low SNR ($\sim$ 10). It is worth mentioning that diffusion maps and kernel PCA give quite stable results regardless of the SNR of the spectra.

According to the results shown above, the best performing dimensionality reduction technique was ICA with 20 independent components. Thus, we selected this compression method to investigate the noise properties of the training set in order to obtain the best generalization performance of the $T_{\rm eff}$, log g and [M/H] estimators. As described in section 3.2, these experiments were carried out using 25 different noise realisations for each SNR analysed (150, 125, 100, 75, 50, 25, 10 and 5) in order to assess the consistency of the results. Fig. 13 shows mean 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 parameter ($T_{\rm eff}$, log g and [M/H]) predictor. Main conclusions of the analysis of this figure are the following:

- There are no large discrepancies amongst the estimations obtained applying the same model to different datasets.
- For spectra of SNR of about 100 or greater, there are minimal differences in the precision achieved with models trained with spectra of SNR of 50 or greater.
- The accuracy is exponentially reduced for estimators constructed from noise-free spectra when the SNR of the spectra is about 75 or lower.
- The overall performance is very similar when the SNR of the evaluation spectra is between 50 and 75, except for the noise-free model in the estimation of $T_{\rm eff}$.
- For the evaluation sets with SNR values of 25, the best accuracy is obtained with the model constructed from spectra with SNR of 50. In the prediction of log g, the best model corresponds to that trained with SNR of 25.
- For SNR lower than 10, the model with best generalization performance is that trained with SNR equal to 10 for $T_{\rm eff}$ and [M/H]. The lowest prediction errors for log g estimation were obtained with the model trained with SNR equal to 25.

This analysis yields the very important consequence that models trained with noise-free spectra are not adequate to estimate atmospheric parameters of spectra with relatively high SNRs (up to 75). Moreover, in order to improve the generalization performance of the models, there is no need to match the SNR of the training set to that of the real spectra. Based on this results, two ICA+SVM models would be enough to estimate $T_{\rm eff}$ and [M/H]: one for spectra with SNR of about 25 or greater (model trained with SNR of 50) and one for spectra with SNR of 10 or lower (model trained with SNR of 10). On the contrary, only one ICA+SVM model trained with SNR of 25 would be enough to estimate the surface gravity.

Finally, Fig. 14 to 18 present the $T_{\rm eff}$ estimation errors

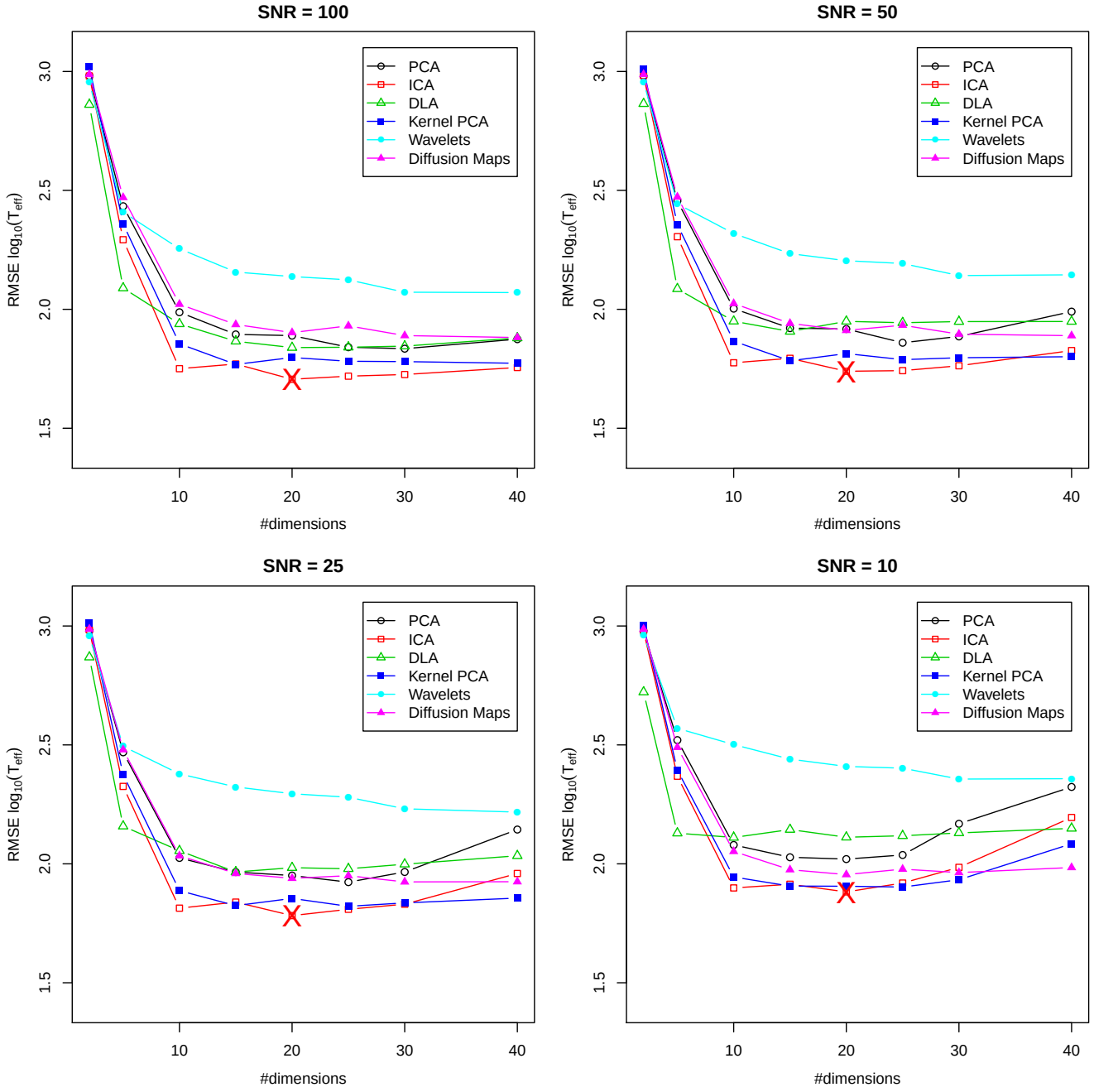

Figure 5. Temperature estimation error as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

that were obtained with different grid densities and SNR levels. 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 to the grid density variation.
- The grid density appeared to have less effect over the performance of Wavelets and diffusion maps.

5 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

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

SNR	Method	Dimension	$RMSE(T_{\rm eff}, K)$	$RMSE(log g)$	$RMSE([M/H], dex)$
Noise-free	DLA	40 / 30[1]	27.16	0.13	0.017
100	ICA	20	50.81	0.15	0.033
50	ICA	20	54.91	0.17	0.038
25	ICA	20	60.59	0.18	0.043
10	ICA	20	76.21	0.21	0.057

[1] The best performance for [M/H] was obtained with 30 dimensions instead of 40.

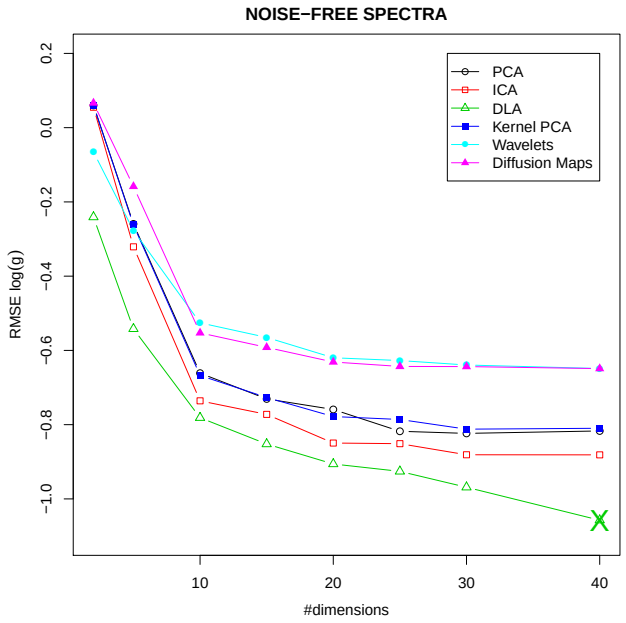

Figure 6. Surface gravity estimation error as a function of the number of dimensions used for data compression, for synthetic spectra

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...

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
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., JÃ¿rgensen 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
Mallat S., 1998, A Wavelet Tour of Signal Processing. Academic Press
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
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
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

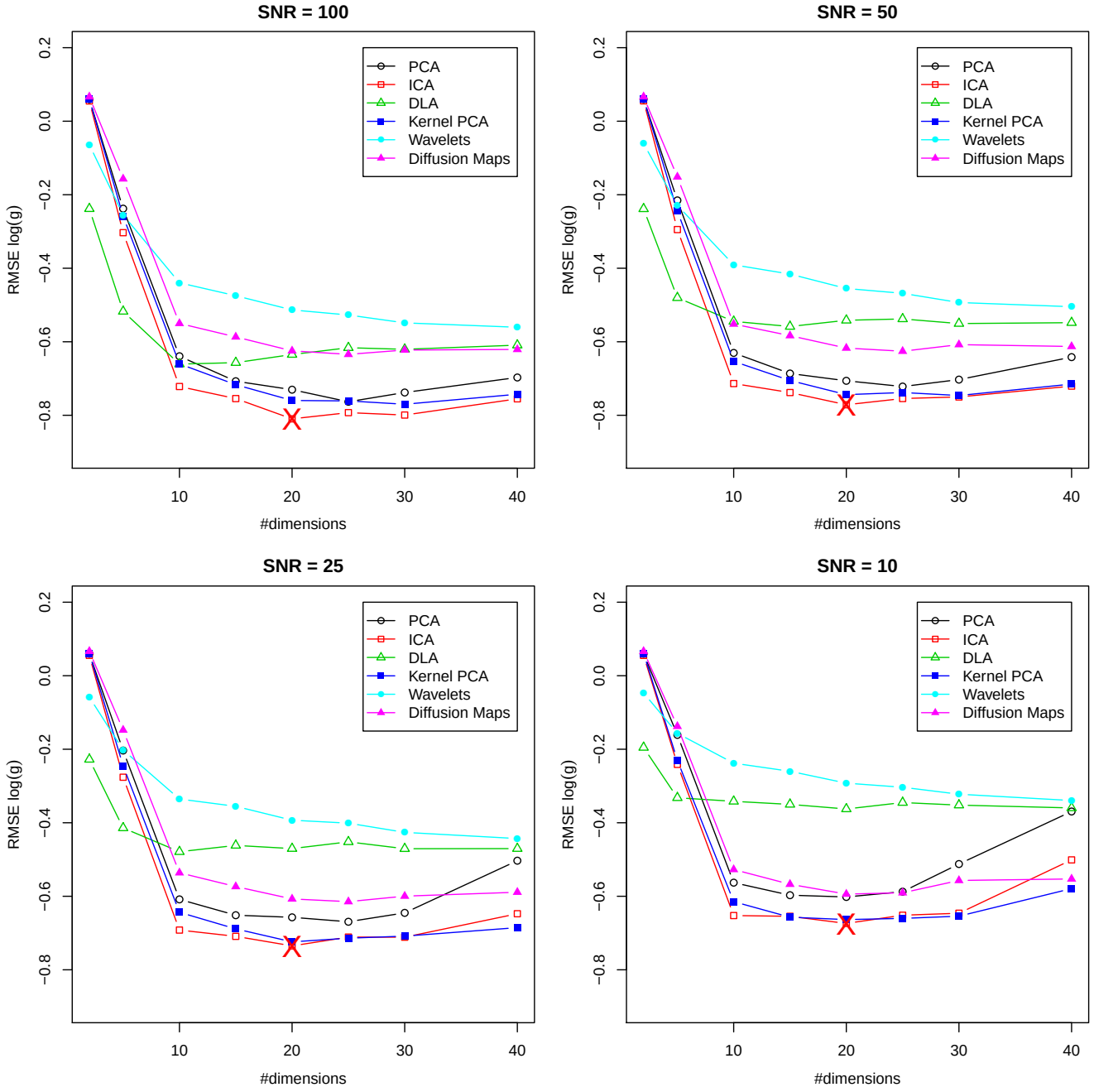

Figure 7. Surface gravity estimation error as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

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
Zarzoso V., Comon P., 2010, IEEE Transactions on Neural Networks, 21(2), 248
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

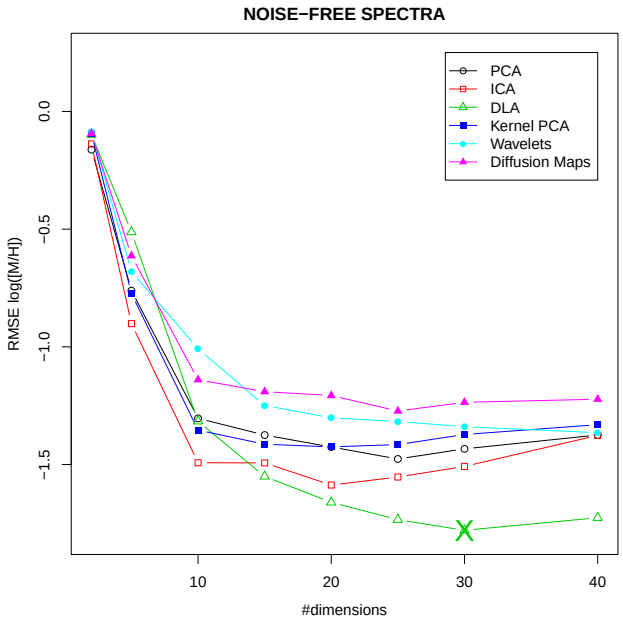

Figure 8. Metallicity estimation error as a function of the number of dimensions used for data compression, for synthetic spectra

APPENDIX A: SOME EXTRA MATERIAL

If you want to present additional material which would interrupt the flow of the main paper, it can be placed in an Appendix which appears after the list of references.

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

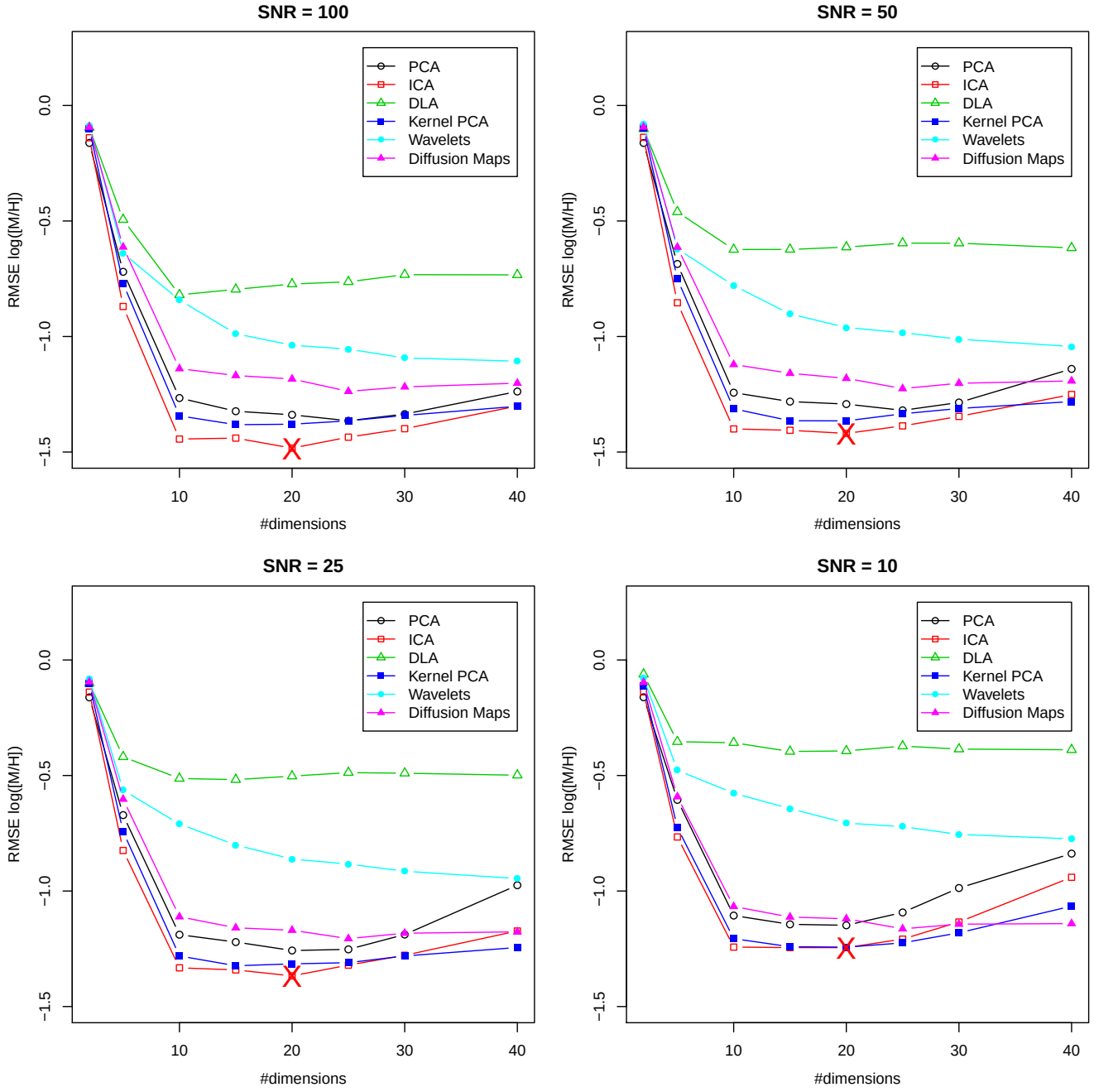

Figure 9. Metallicity estimation error as a function of the number of dimensions used for data compression, for noisy synthetic spectra.

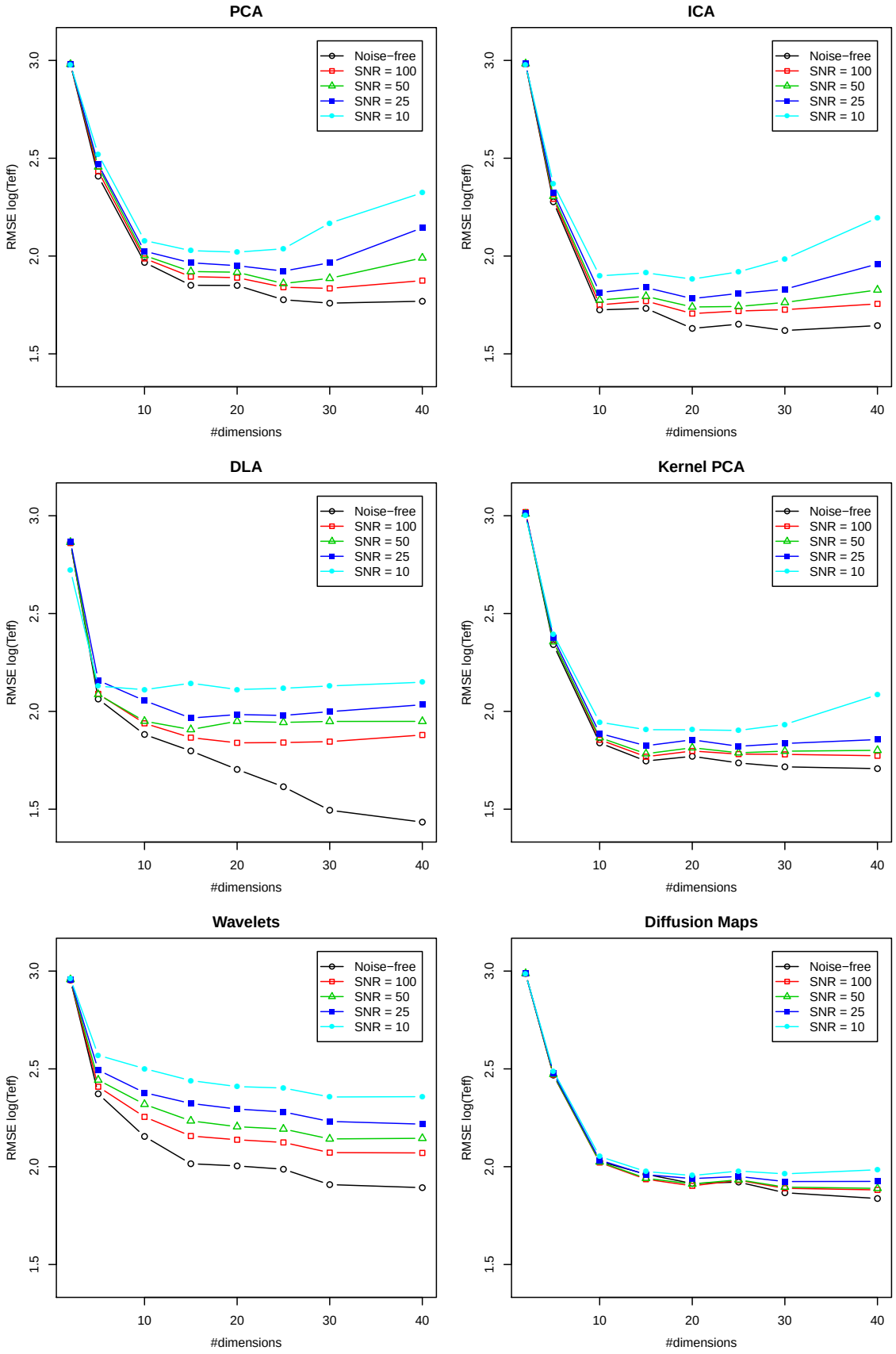

Figure 10. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

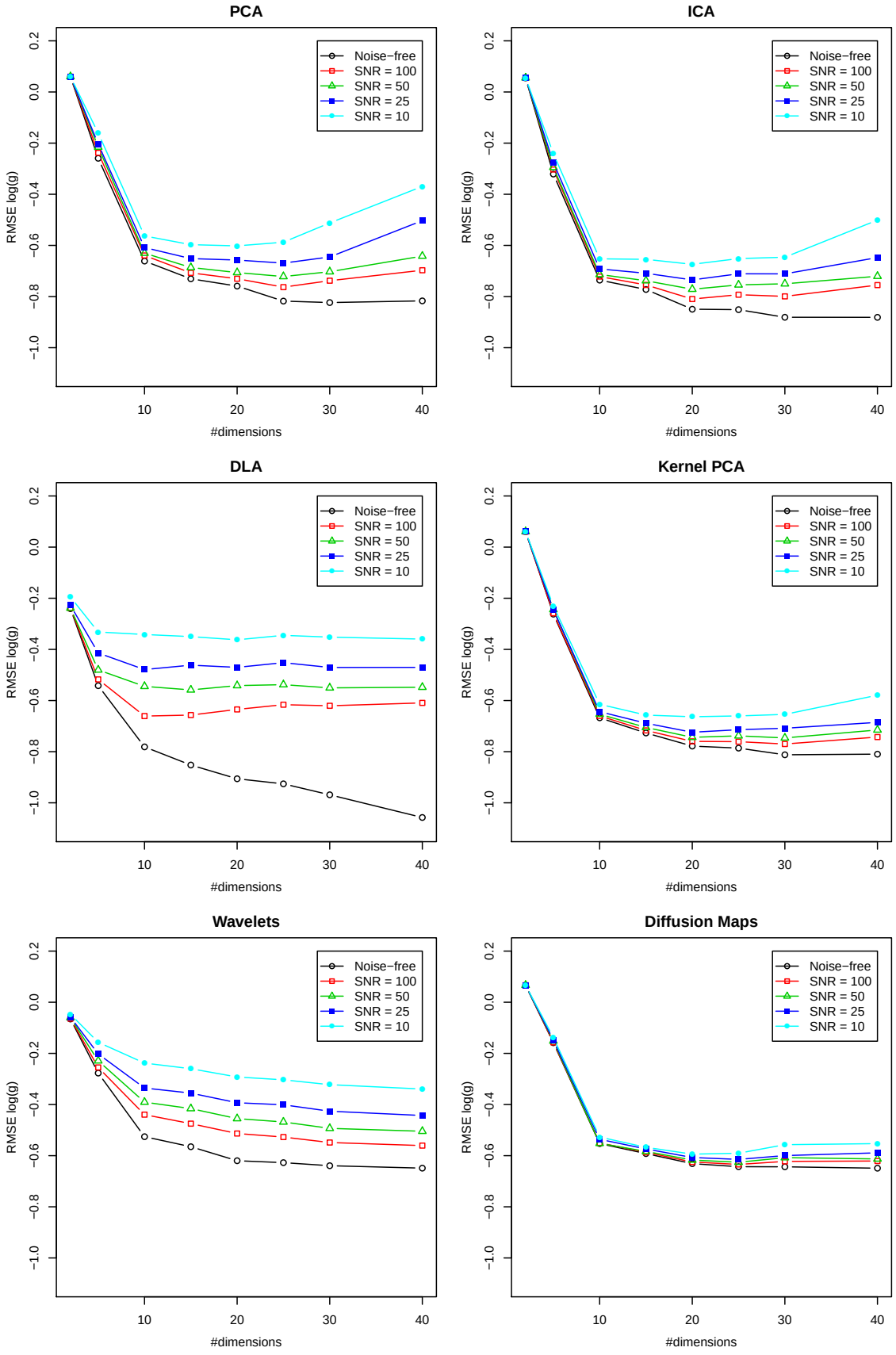

Figure 11. Surface gravity estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

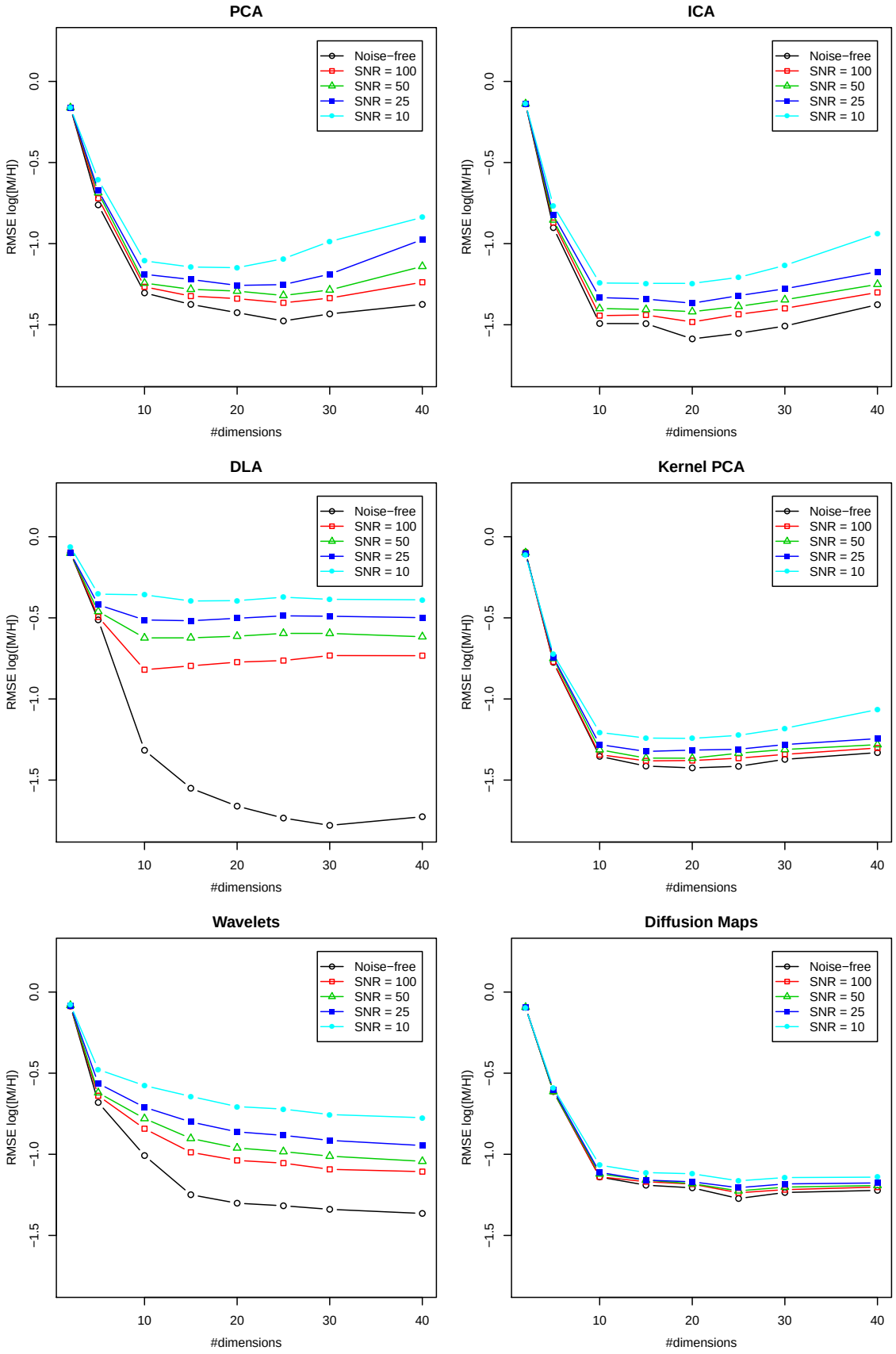

Figure 12. Metallicity estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific SNR

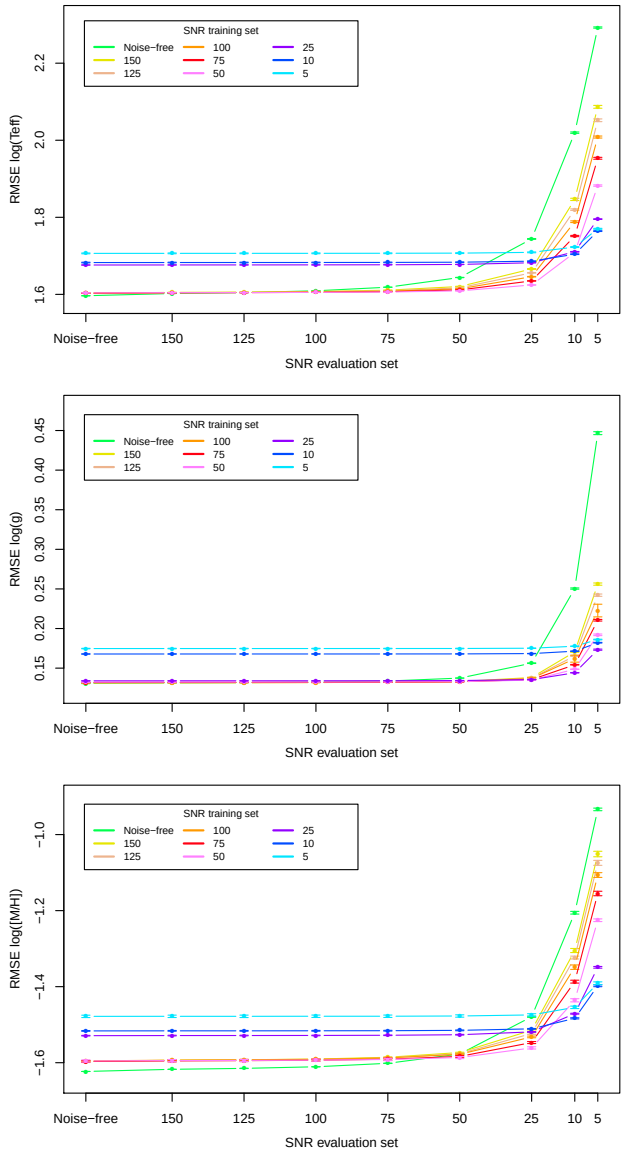

Figure 13. Temperature estimation error against the SNR of the evaluation set. Each line corresponds to a model trained with a specific SNR

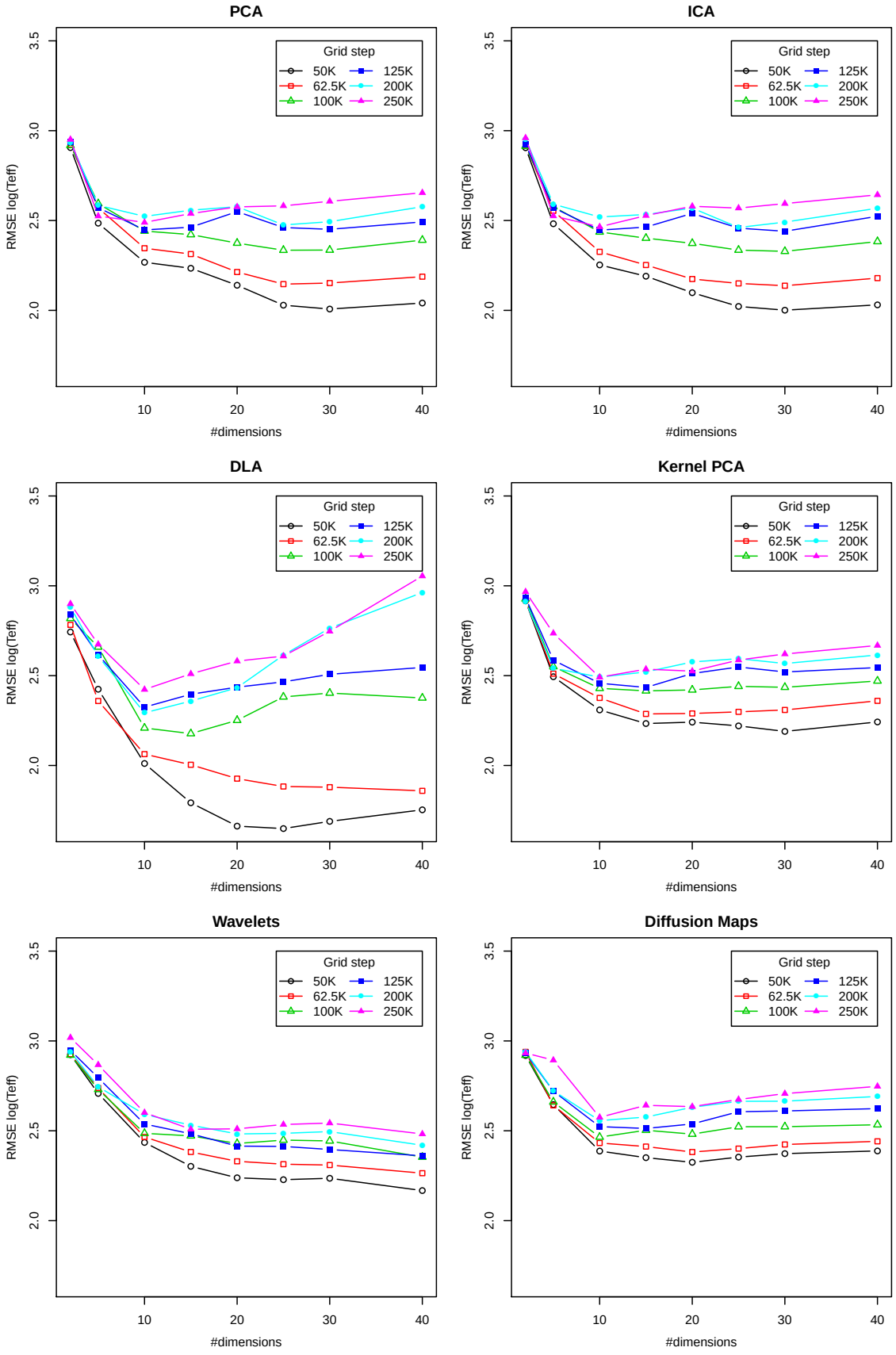

Figure 14. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (Noise-free spectra)

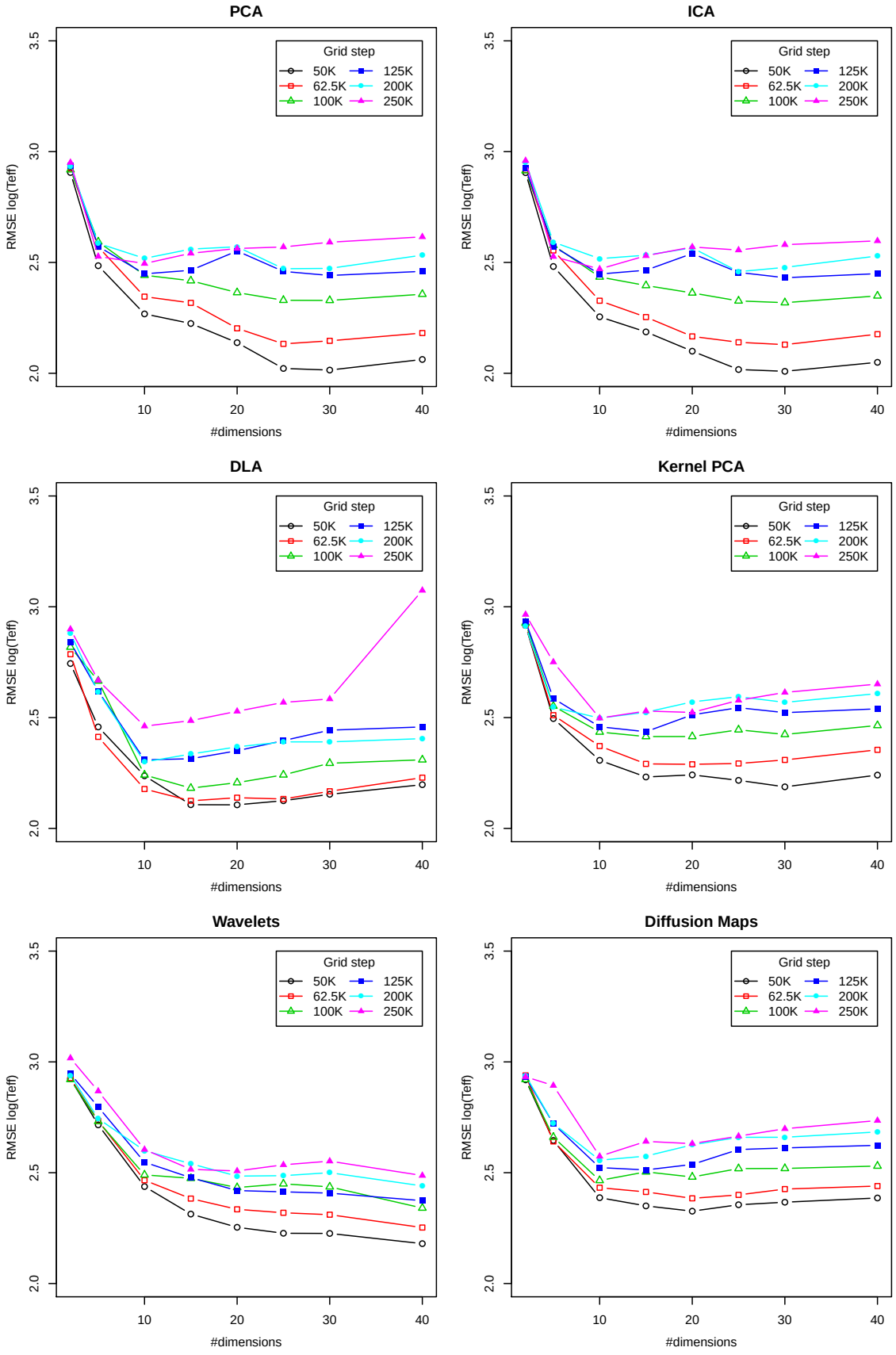

Figure 15. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 100)

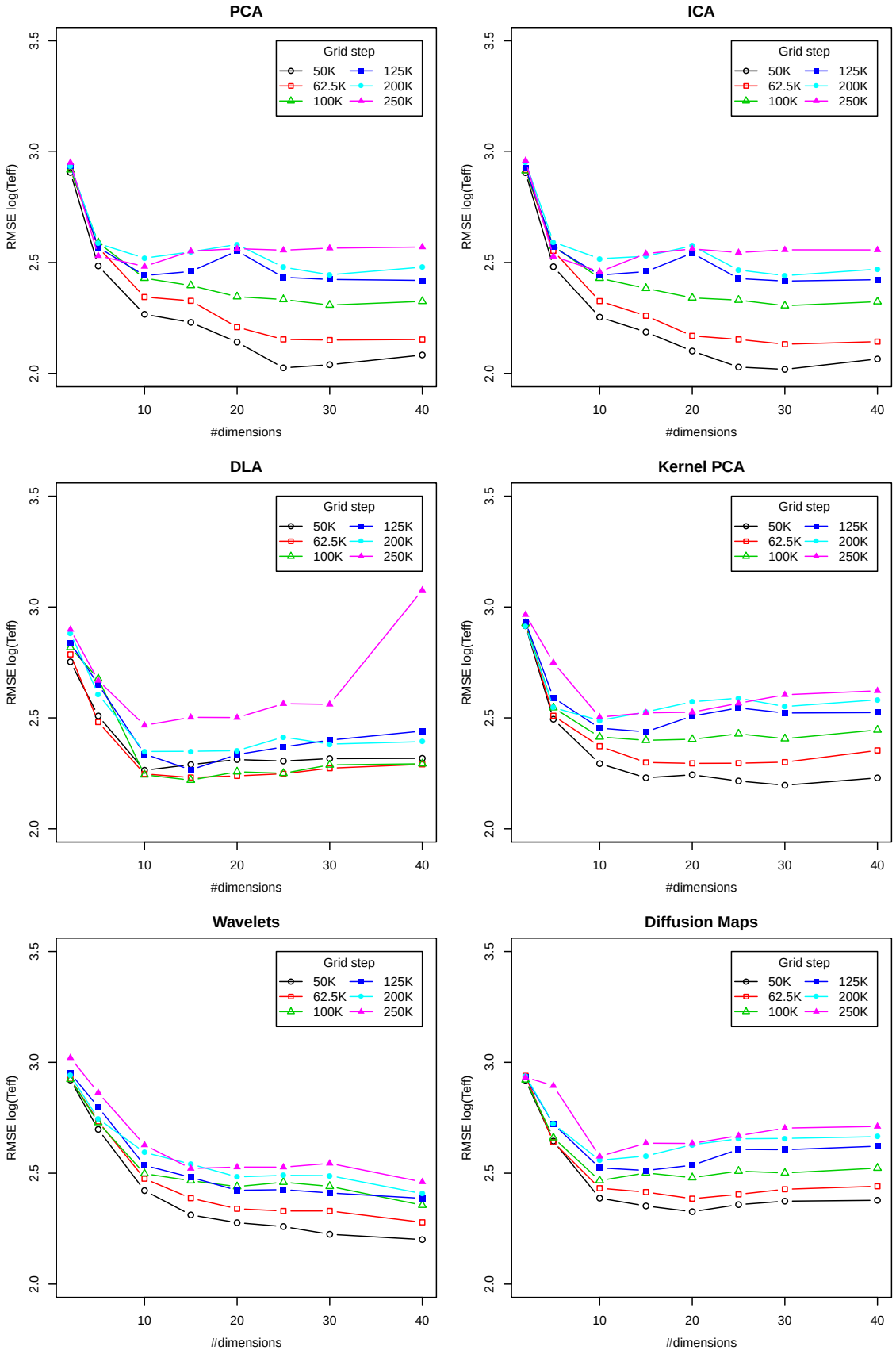

Figure 16. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 50)

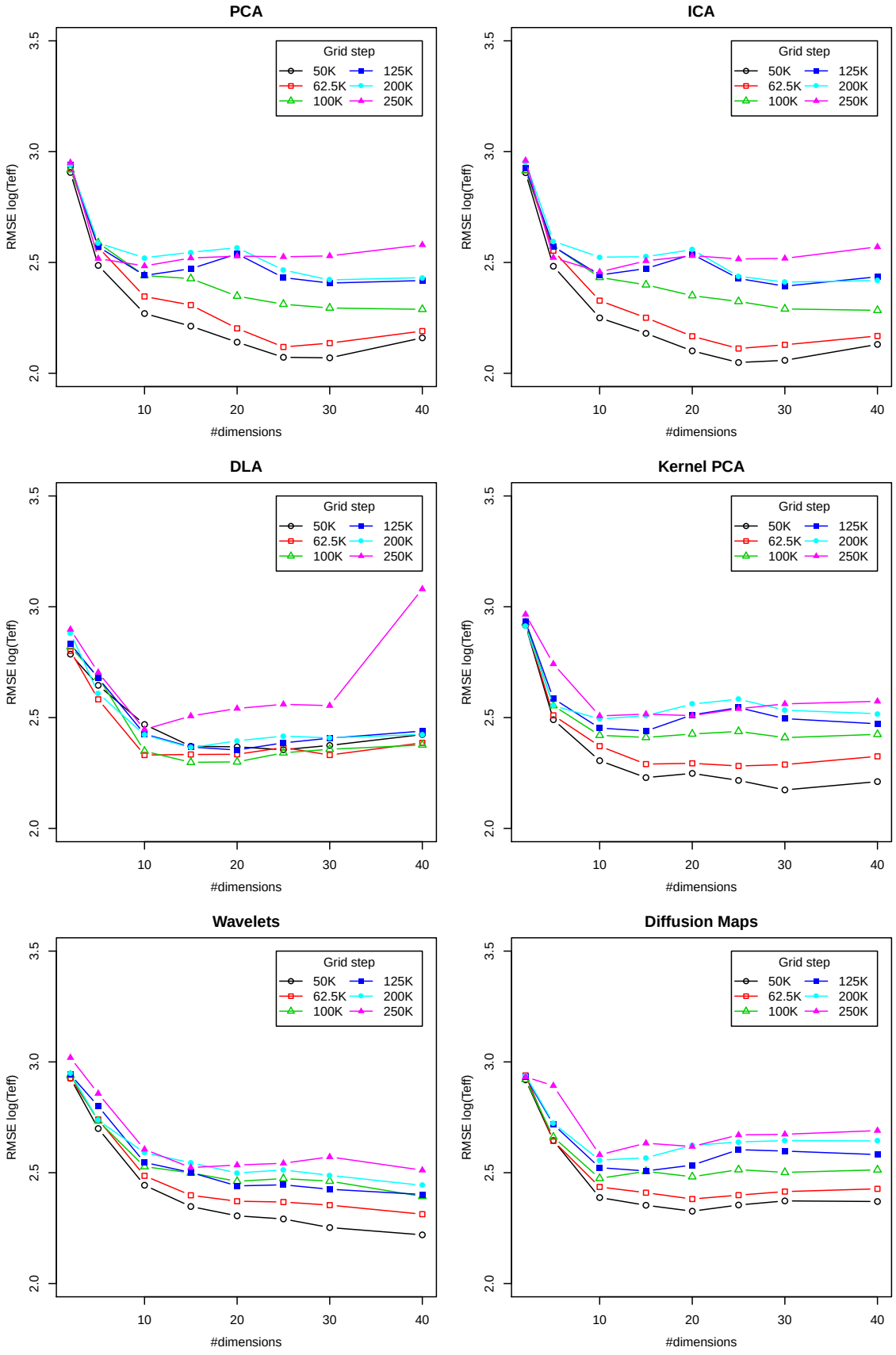

Figure 17. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 25)

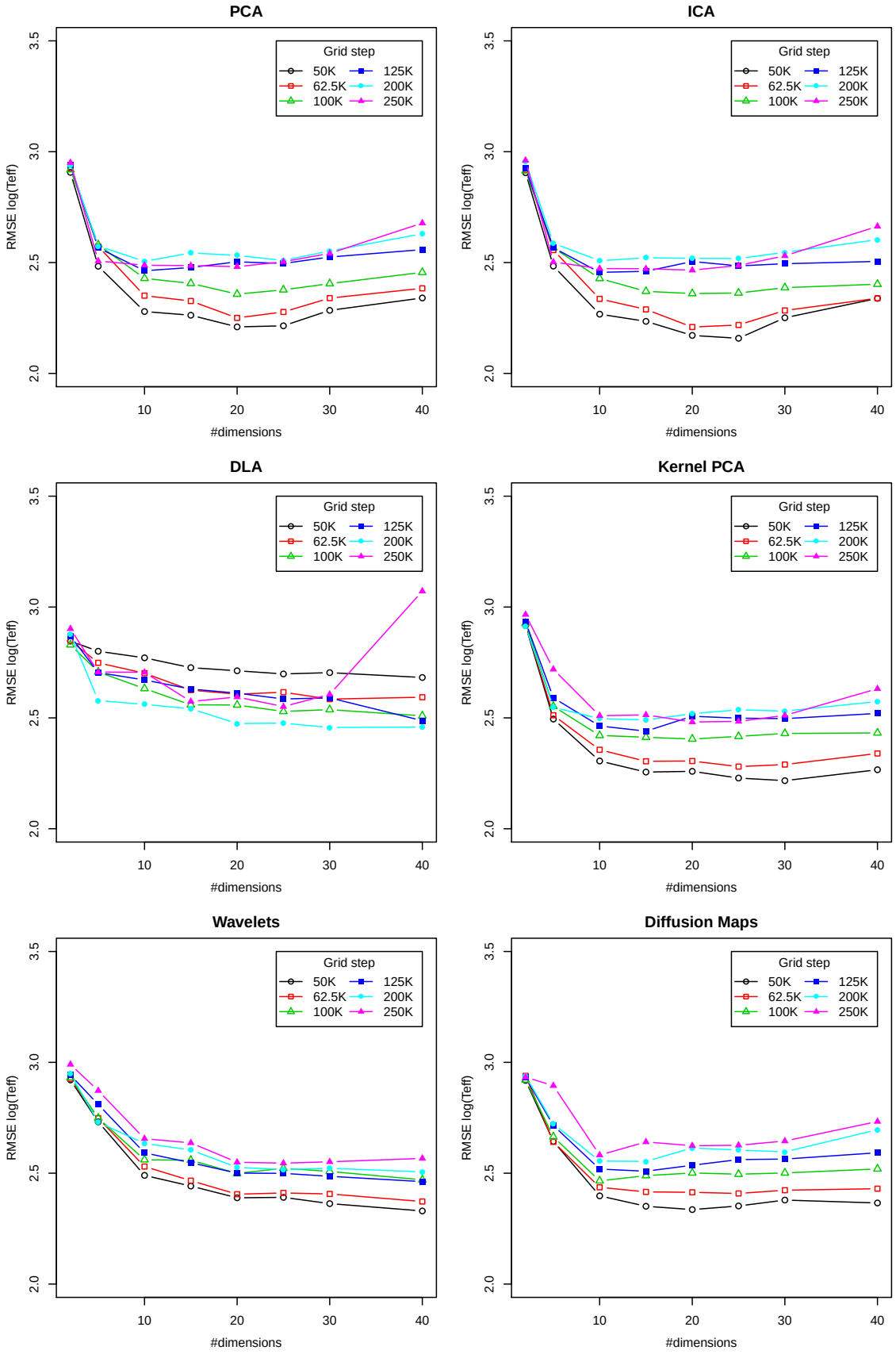

Figure 18. Temperature estimation error against the number of dimensions used for data compression. Each line corresponds to a model trained with a specific grid step (SNR = 10)