RAA 2019 Vol. **19** No. **10**, 140(8pp) doi: 10.1088/1674–4527/19/10/140

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

A PCA approach to stellar abundances I. testing of the method validity

Wei He[1,2] and Gang Zhao[1,2]

[1] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; *whe@nao.cas.cn*
[2] School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049, China

Received 2019 March 28; accepted 2019 April 30

Abstract The derivation of element abundances of stars is a key step in detailed spectroscopic analysis. A spectroscopic method may suffer from errors associated with model simplifications. We have developed a new method of deriving the various element abundances of stars based on the calibration established from a group of standard stars. We perform principal component analysis (PCA) on a homogeneous library of stellar spectra, and then use machine learning to calibrate the relationship between principal components and element abundances. By testing with spectral libraries S4N and MILES, we find that our procedure provides good consistency when spectra from a homogeneous set of observations are used, and it could be expanded to stars with quite a wide range of stellar parameters, with both dwarfs and giants. Moreover, we discuss the four key factors that have a significant impact on the results of derived element abundances, including the resolution of the spectra, wavelength range, the signal-to-noise ratio (S/N) of spectra and the number of principal components adopted.

Key words: stars — stars: abundances — techniques: spectroscopic — methods: data analysis

1 INTRODUCTION

The spectra we sample from a star provide information on the atmosphere of the star. Stellar spectra are described with three basic atmospheric parameters including effective temperature, surface gravity and overall metallicity. The most notable information inferred from the spectrum of a star is the chemical composition, from the strength of the absorption lines associated with different elements. By measuring the element abundances of stars, we seek to account for the production of chemical elements that we identify in the universe, its time dependence and for many of the features of galaxies that we observe. Understanding stellar evolution, the birth and death of stars and how they interact with their environments is critical to comprehending the evolution of galaxies.

There are several ways to determine the element abundances of a star. The two most common approaches are the direct comparison of observed and synthetic spectra, and the equivalent width technique based on excitation equilibrium and ionization balance (Blanco-Cuaresma et al. 2014). The synthetic spectral fitting technique tries to minimize the difference between observed and synthetic spectra by directly comparing the whole observation or some delimited regions. The equivalent width method does not use all the information contained in the shape of the absorption-line profiles, but only their area, to deduce the element abundances from different lines. Also, to determine element abundances, the stellar atmospheric parameters – effective temperature, surface gravity, overall metallicity, microturbulent and macroturbulent velocities, and rotation – must be known. Moreover, the model atmospheres and line list selection will all contribute to the error of measured element abundances.

Ongoing new surveys have resolutions that allow more precise determination of not only the stellar parameters of stars, but also of the chemical abundances of several individual elements. Examples of such projects are the Gaia-ESO Survey (Gilmore et al. 2012; Randich et al. 2013), RAVE (Steinmetz et al. 2006), APOGEE (Allende Prieto et al. 2008), GALAH (da Silva et al. 2012), and the future with millions of stars from the Radial Velocity Spectrograph (RVS) provided by *Gaia*. A sufficient quantity of accurate stellar abundances will reveal the complete chemical evolution of the Milky Way, and most automated pipelines have evolved from traditional manual

methods. However, there has been relatively little work done to analyze the spectra using unconventional methods. The Cannon is a data-driven model that is effective for inferring physical attributes of stars from spectra (Ness et al. 2015; Casey et al. 2016). The basic assumption by which Cannon operates is that the spectrum of a star is a function of its stellar parameters and the element abundance is a smooth function of the independent variable, while our hypothesis is that principal component analysis (PCA) could help to retain important information in the spectrum, and the nonlinear function between the principal components and the element abundances can be reversed by machine learning. Our procedure is inspired by the work of Bermejo et al. (2013), who derived stellar effective temperatures using PCA and Bayesian calibration.

In this paper, we introduce a new method of inferring stellar abundances. We firstly condense the information contained in stellar spectra using PCA. Then we map the principal components onto stellar abundances based on a set of calibration stars. Once the calibration for some standard stars is performed by deep learning, we can derive stellar abundances for other stars based on the calibration. By applying PCA, it is optimized to make use of the most information inherent in a spectrum.

We describe the spectral library and methods of calibrating the relationship between spectra and element abundances in Section 2. The test results and analysis of factors influencing the derivation of element abundances are presented in Section 3. Finally, the conclusions can be found in Section 4.

2 DATA PREPARATION

2.1 S4N Library

The S4N library comes from a high-resolution spectroscopic survey of all the stars more luminous than $M_{\rm V} = 6.5$ within 14.5 pc from the Sun (Allende Prieto et al. 2014). A preliminary abundance and kinematic analysis of FGK stars in the sample was performed. Also, abundances of 16 chemical elements were discussed based on transitions of majority species. These spectra have a signal-to-noise ratio (S/N) of 150−600 and a resolving power of $R{\sim}50\,000$. They cover the range of -0.9 to 0.5 in metallicity. The final spectra were velocity corrected using hundreds of solar atomic lines as reference, and this helps to produce a highly homogeneous archive. We are left with 104 stars as we only include stars with the element abundances well calibrated in the library. Formerly, the determination of abundances, and micro- and macroturbulence, was accomplished by minimizing the chi-square value between synthetic and observed line profiles. A line-by-line approach was used to identify outliers and estimate internal uncertainties. Moreover, this approach to calculating the element abundances is strictly differential as it incorporates every single abundance estimate from a line to the solar abundance from the same line.

2.2 MILES Library

The MILES database contains flux calibrated optical spectra with high S/N for 985 stars covering $\lambda\sim3525-7000$ Å with a homogeneous resolution full width at half maximum (FWHM) $= 2.3$ Å. The parametric coverage of sample stars is quite wide: $2800 \leq T_{\rm eff} \leq 50\,400$ K, $0.0 \leq \log(g) \leq +5.0$ and $-2.7 \leq$ [Fe/H] $\leq +1.0$ dex. Milone et al. (2011) have obtained [Mg/Fe] measurements for 76.3% of stars in the MILES spectral library. These abundance ratios were procured through a compilation of values from the literature on high-resolution spectroscopic studies and analysis using MILES mid-resolution spectra. The calibrated [Mg/Fe] values have small average uncertainties, which are 0.09 with high-resolution spectra and 0.12 with mid-resolution spectra. Though spectra in the MILES database had been flux calibrated, a homogeneous 1-D normalization was applied through the `continuum` task of the `onedspec` Package of the NOAO Optical Astronomy Packages of `Image Reduction and Analysis Facility` (`IRAF`).

2.3 Method

2.3.1 PCA approach

PCA is an algebraic and statistical tool which aims to find directions of the largest variance in the data. To use principal components, we adopt a new basis set formed by eigenvectors of the correlation matrix and order them by decreasing eigenvalues. With projection of the data onto the first tens of elements of the basis, we can acquire the principal components of the projected spectra and reproduce optical low-resolution spectra with minimal error. Moreover, it implies that the PCA method leaves us with enough information about the data. So with minimal effort, PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the hidden and simplified structures. We processed the data using PCA with the following steps:

1. Calculate the mean spectrum and subtract it from each spectrum.

2. Organize data matrix **D** which has dimensions **M**×**N**, where **M** is the number of spectra and **N** is the number of data points from each spectrum.

3. Apply singular-value decomposition (SVD) to the data matrix, $D = UWV^T$, where **W** is a diagonal matrix with positive elements, and **V** is a matrix of eigenvectors (each column is an eigenvector, the eigen-spectra in this work). The eigenvectors are called principal axes or the principal directions of the data. Projections of the data on the principal axes are called principal components.

By using PCA, we condense the calibration set of spectra into numbers of eigen-spectra. Then by projecting the test set of spectra onto the eigen-spectra, the principal components of each test spectrum would be derived. These principal components of each test spectrum should contain information on how different elements affect the shape of the spectra. Finally, we seek a calibration between the principal components of the spectra and their elemental abundances using machine learning.

2.3.2 Machine learning

Machine learning implements statistical techniques to enable computer systems to "learn" data. Machine learning tasks are usually divided into two categories, unsupervised learning and supervised learning, depending on whether there is a final goal of learning.

Unsupervised learning should find its own structure in the input. However, a supervised learning algorithm analyzes training data and produces an inferred function. With this function, we can predict the output variables from the new input data. To infer the mapping function, we used a class of feedforward artificial neural networks called a multilayer perceptron (MLP). An MLP consists of at least three layers of nodes. Between the input and output layer, there can be one or more non-linear layers, called hidden layers.

The input layer contains a set of neurons representing the input features, here referring to the principal components of the spectra. The output layer will be the element abundances. Each neuron in the hidden layer transforms values from the previous layer with a weighted linear summation, followed by a non-linear activation function; here we use the rectified linear unit function.

For regression, the MLP uses the Square Error loss function, written as

$$\mathrm{Loss}(\hat{y}, y, W) = \frac{1}{2}||\hat{y} - y||_2^2 + \frac{\alpha}{2}||W||_2^2 \,. \qquad (1)$$

The learning problem in neural networks is formulated with respect to minimization of the loss function. This function is, in general, composed of an error term and a regularization term. The error term evaluates how well the neural network fit the data set. On the other hand, by controlling the actual complexity of the neural network, the regularization term is used to prevent overfitting. Starting with initial random weights, the MLP minimizes the loss function by repeatedly updating these weights (Rumelhart et al. 1986). After the loss is calculated, it is passed back to propagate it from the output layer to the previous layers, providing an updated value for each weighting parameter that is intended to reduce the loss.

The necessary condition states that if the neural network is at a minimum of the loss function, then the gradient is the zero vector. Newton's method is a second order algorithm because it makes use of the Hessian matrix. The purpose of this method is to find a better training direction by using the second derivatives of the loss function. The application of Newton's method is computationally expensive because it requires many operations to evaluate the Hessian matrix and calculate its inverse. The algorithm we adopted here is Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS), an optimization algorithm in the quasi-Newton method family that uses a finite amount of computer memory to approximate the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Liu & Nocedal 1989). It is a simplified algorithm for parameter estimation in machine learning. L-BFGS utilizes an estimate of the inverse Hessian matrix to guide its search through variable space. Whereas BFGS stores a dense $n \times n$ approximation to the inverse Hessian (with n being the number of variables in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. Due to its resulting linear memory requirement, the L-BFGS method is particularly well suited for optimization problems with a large number of variables. The algorithm stops when the preset maximum number of iterations is reached, or when the improvement in loss is below a certain fractional number.

Before we apply the machine learning method, we need to preprocess the data. MLP is sensitive to feature scaling, so we need to scale our data before modeling. For instance, we scale each attribute in the input vector X, the principal components of each spectrum to $[-1, +1]$, and the element abundances to [0, 1]. Of course, the same scaling must be applied to the calibration set and the test set at the same time to obtain meaningful results. Our cal-

culation is implemented in libraries from the *scikit-learn* Python package (Pedregosa et al. 2011).

3 RESULTS AND DISCUSSION

We now present the results of applying the previous formalism to the calibration of various element abundances. The inferred functions are computed by deep learning using the calibration set. Then we apply the model to the test set in the same library. To evaluate the performance of fitting the element abundances, we will focus on the recovery of [X/Fe].

With the S4N library, we randomly selected 80 stars as the calibration set and the remaining 24 stars as the test set. In Figure 1, we show the recovery performance of different kinds of element abundances including alpha elements, iron-peak elements and heavy elements. The calculation is based on spectra with downgraded resolution of 10 000. The recovery errors of most samples are less than 0.1 dex. Also, we find that the ratio of two elements from different categories is easier to recover than that of two elements from the same category, since two elements from the same category tend to be correlated and quite similar, especially [Ti/Sc], [Ni/Fe], [Nd/Ce], etc.

As with the S4N library, we randomly selected 532 stars in the MILES library as the calibration set and 155 stars as the test set. In Figure 2, we show both the recovery results from our procedure and error bars from (Milone et al. 2011). The root mean square (RMS) of our recovery error is $\sim$0.15 dex, which is close to the mean error of [Mg/Fe] $\sim$0.12 from mid-resolution spectra in MILES.

We will discuss the four key factors that have a significant impact on the recovery results, including the resolution of the spectra, wavelength range, S/N of the spectra and the number of principal components adopted.

3.1 Resolution

For the stellar element abundances we calculated, we tested with both $R = 10\,000$ and 1000. Though degrading the spectra means destroying information, it does help to reduce the impact of high-frequency instrumental distortions in the data. When we smoothed the spectra from $R =$ 10 000 to 1000, the recovery results of element abundances for the test set of stars became worse at $R = 1000$, which means that lower resolution would cause the loss of some critical information about the element abundances.

For example, we compare the recovery performance of [Ca/Si] under different resolutions of $R = 1000$ and 10 000 in Figure 3. The error range of the testing set naturally becomes expanded with lower resolution. Other than [Ca/Si], we also tested the influence of resolution on [Y/Si], though the RMS error seems to be close. Apparently the error itself shows some tendency against the actual value of element abundances. For each element, the recovery accuracy depends on how much information could be extracted by principal components from the spectra, and the suitably higher resolution means that the spectra have kept more information about the shape of the elements' curve on the spectra.

3.2 Wavelength Range

The wavelength range plays an essential role in the calculation of element abundances since different elements would affect the shape of the spectra to different extents at various ranges, though Bermejo et al. (2013) found that the spectral range was not critical for obtaining reliable results for effective temperature. However, for different elements, their absorption lines are concentrated in different ranges of the whole spectrum.

For simplicity, we split the spectra into the red part (5800 Å$-$6800 Å) and the blue part (4800 Å$-$5800 Å) to compare the fitting results under different wavelength ranges. For most of the elements we study here, the red part is more conducive to the recovery of the abundances. Also, the three alpha elements: Mg, Si and Ca are most sensitive to the wavelength range. Notably, we compare the recovered [Mg/Si] under the two different parts of the spectrum in Figure 4. It demonstrates that the red part of the spectrum can distinguish the relative abundances of Mg and Si, but the blue part cannot lead to the correct results since it overestimates the lower values and underestimates the higher values.

Moreover, the iron-peak elements are almost unaffected by this factor. Also, among the heavy elements, Ba, Ce, Nd and Eu are recovered more precisely under the red part of the spectrum, although Y is good with both parts of the spectrum. Apparently, for FGK stars, the red part of the spectrum contains more information about the element abundances of a star, and such related information could be extracted by the PCA method.

3.3 Truncation of Principal Components

In fact, the number of principal components has a great impact on the calibration accuracy of the element abundances. Bermejo et al. (2013) stated that an accurate calibration of the effective temperature with the principal components from spectra could be obtained independently

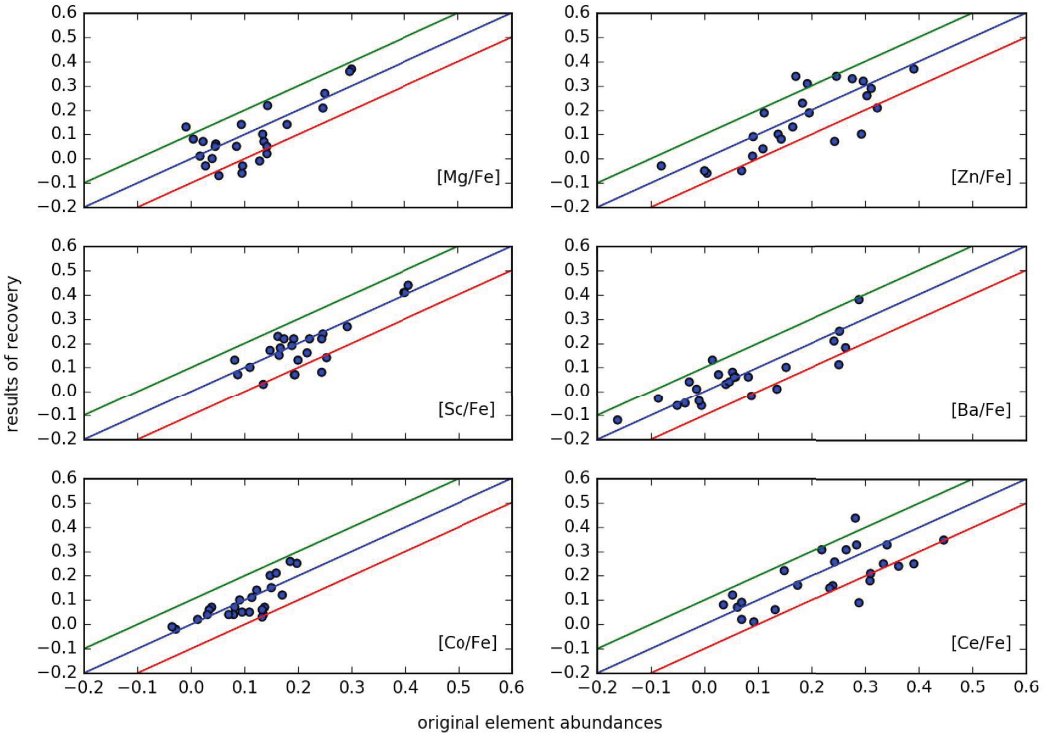

Fig. 1 Recovery results of the test set from the S4N library with [Mg/Fe], [Sc/Fe], [Co/Fe], [Zn/Fe], [Ba/Fe] and [Ce/Fe]. The *three solid lines* with slope = 1 have intervals of 0.1 dex.

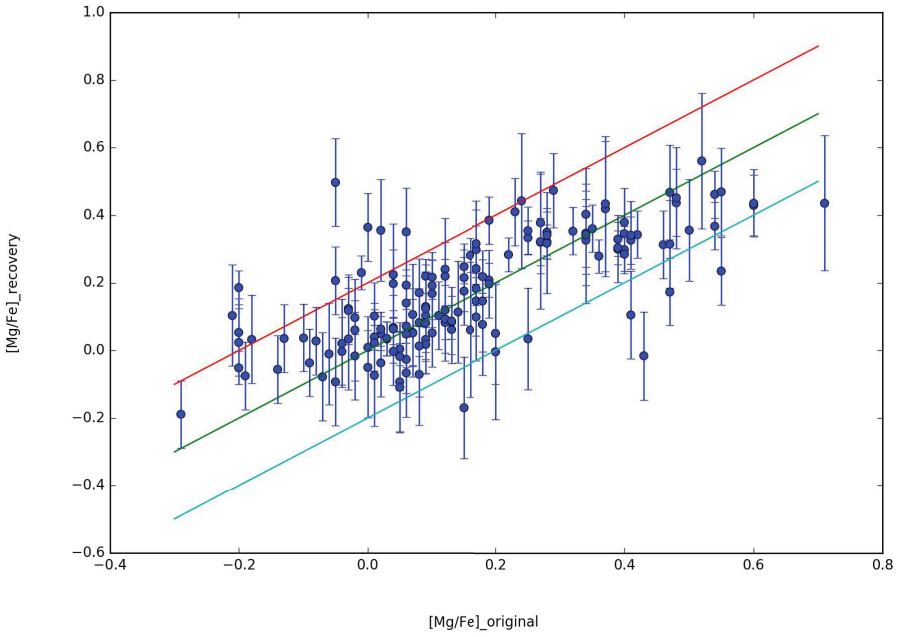

Fig. 2 The recovery results of the test set from the MILES library with [Mg/Fe]. The error bars come from Milone et al. (2011). The three *solid lines* have the same meanings as in Fig. 1.

from the number of principal components. However, it is evident that the element abundances have a more complicated relationship with the shape of the spectra than the effective temperature, so we have to decide the proper truncation of the principal components.

After our tests, for the $R = 10\,000$ spectra, increasing the number of principal components would be helpful for

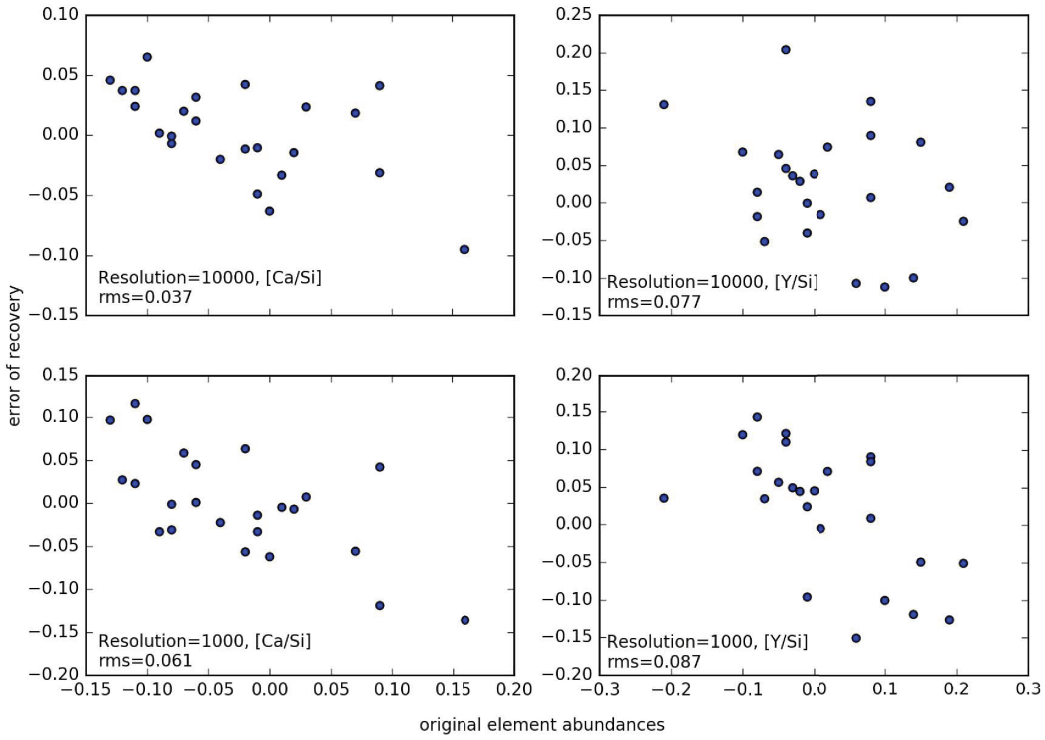

Fig. 3 Comparison of the recovery performance of [Ca/Si] under resolutions of $R = 1000$ and 10 000.

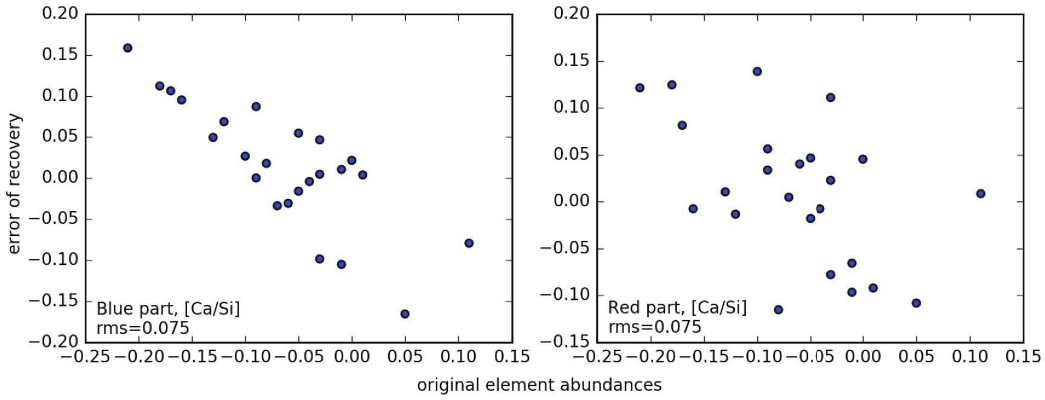

Fig. 4 Comparison of the recovered [Mg/Si] under the two different parts of the spectrum.

deriving more accurate element abundances. In Figure 5, we compare the fitting error of [Mg/Si] for the test stars set with the number of principal components ranging from 3 to 80. It is apparent that the truncation of principal components severely affects the fitting results. The PCA method determines that the most critical information be focused on principal components with the most eigenvalues, and keeping most of the extracted features would help to deduce the element abundances more precisely. However, for the $R = 1000$ spectra, the resolution itself limits the useful information, and even with all the principal components kept, the fitting accuracy cannot reach an acceptable level.

3.4 S/N of the Spectra

The value of S/N can limit the calibration of element abundances. Initially, a spectrum in the S4N spectral library has a sufficiently high S/N. Also, convolving these spectra to a lower resolution would result in extremely high S/N per pixel. Therefore, as a test, we directly superimposed random numbers with a normal distribution onto the degraded spectra. Recalculating these spectra with given errors can manifest the impact of spectral S/N on our method.

There are two ways in which S/N affects the measurement of element abundances. On one hand, the S/N of the calibration set determines whether the relationship be-

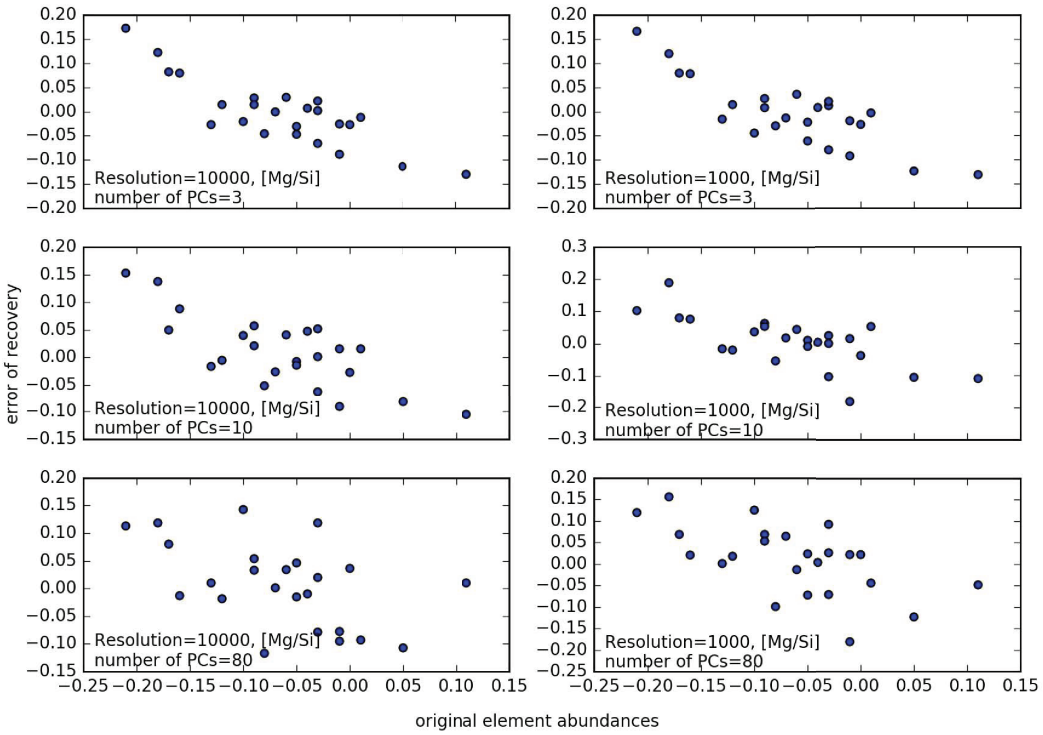

Fig. 5 Comparison of the fitting error of [Mg/Si] for the test stars set with the number of PCs ranging from 3 to 80. The left panels display the results under a spectral resolution of 10 000, while the right panels are those under a resolution of 1000.

tween element abundance and principal component can be accurately determined. On the other hand, the S/N of the test set will affect the accuracy of the calculated element abundances.

After our tests, the low S/N can lead to overfitting. In the case of a resolution of 10 000, the S/N per pixel of the calibration set needs to exceed 300 to ensure fairly correct fitting of element abundance and principal component relationships, whereas, for the spectra with resolution of 1000, the S/N per pixel of the calibration set needs to be more than 500.

It is not surprising that we require the S/N of the calibration set of spectra to be quite high to ensure that the element abundances can be inferred correctly. However, the S/N of the test set does not need not to be that high to achieve acceptable results. With a spectral resolution of 10 000, unless the S/N of the test star is less than 30, the measurements of most element abundances would not have uncertainties larger than 0.15 dex. However, for the spectral resolution of 1000, the S/N of the test set needs to be at least 100 to ensure accuracy.

4 CONCLUSIONS

We have developed a method of deriving element abundances from spectroscopic data by projecting observed spectra onto the eigenvectors and applying the calibration derived by machine learning based on a set of stars with reliably measured element abundances. Our method converts the spectra into a fairly small number of principal components, and it mainly studies the relationship between overall morphology of the spectra and element abundances. It also helps to calculate elemental abundance for medium and low resolution spectra.

Unlike traditional methods of determining element abundances, our procedure does not suffer from problems of atmospheric models, line selection or line profile fitting. We checked the method internally for S4N spectra with excellent results for alpha elements, iron-peak elements and heavy elements, and we also show that this method could be expanded to stars with a wide range of stellar parameters ($0.0 \leq \log(g) \leq +5.0$ and $-2.7 \leq$ [Fe/H] $\leq +1.0$ dex) by testing with [Mg/Fe] from the MILES library.

Moreover, we tested the parameters which could affect the calibration accuracy. We discovered that suitably high resolution helps to extract useful information from the spectra. Different elements suffer different impacts if we change the range of spectra being analyzed. Also, the S/N of the calibration set needs to be high enough to derive the relationship between element abundances and principal components. The Cannon method mainly employs spectra

with precise stellar parameters and elemental abundance as reference samples for machine learning, and then calculates the elemental abundance of spectra with lower S/N. At this point in line with our approach, previous discussion in this paper also shows that we have higher S/N requirements for the reference group's spectra, while the calculated spectrum requires lower S/N.

In further research, we expect to include more samples of high quality stellar spectra with precisely measured element abundances to cover a wider range of stellar parameters. As this method has demonstrated great potential in extracting chemical information from mid-resolution spectra, we plan to apply this method to the LAMOST spectral survey (Zhao et al. 2006, 2012) in our second paper in this series to study the structure and chemical evolution of the Milky Way.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Grant Nos. 11890694 and 11390371).

References

Allende Prieto, C., Barklem, P. S., Lambert, D. L., & Cunha, K. 2004, A&A, 420, 183

Allende Prieto, C., Majewski, S. R., Schiavon, R., et al. 2008, Astron. Nachr., 329, 1018

Munoz Bermejo, J., Asensio Ramos, A., & Allende Prieto, C. 2013, A&A, 553, A95

Blanco-Cuaresma, S., Soubiran, C., Heiter, U., & Jofre, P. 2014, A&A, 569, 111

Casey, A. R., Hogg, D. W., Ness, M., Rix, H.-W., Ho, A. Q., & Gilmore, G. 2016, preprint (arXiv:1603.03040)

da Silva, R., Porto de Mello, G. F., Milone, A. C., et al. 2012, A&A, 542, A84

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25

Liu, D. C., & Nocedal, J. 1989, Mathematical Programming, 45, 503, https://doi.org/10.1007/BF01589116

Milone, A. D. C., Sansom, A. E., & Sanchez-Blazquez, P. 2011, MNRAS, 414, 1227

Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533

Randich, S., Gilmore, G., & Gaia-ESO Consortium, 2013, The Messenger, 154, 47

Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645

Zhao, G., Chen, Y.-Q. & Shi, J.-R., et al. 2006, ChJAA (Chin. J. Astron. Astrophys.), 6, 265

Zhao, G., Zhao, Y.-H. & Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, RAA (Research in Astronomy and Astrophysics), 12, 723