RAA 2014 Vol. **14** No. **4**, 423–432 doi: 10.1088/1674–4527/14/4/005
http://www.raa-journal.org *http://www.iop.org/journals/raa*

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

Estimating stellar atmospheric parameters based on Lasso features

Chuan-Xing Liu[1], Pei-Ai Zhang[1] and Yu Lu[2]

[1] Department of Mathematics, Jinan University, Guangzhou 510632, China; *qzhzhang@163.com*
[2] School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China

Received 2013 May 21; accepted 2014 January 28

Abstract With the rapid development of large scale sky surveys like the Sloan Digital Sky Survey (SDSS), GAIA and LAMOST (Guoshoujing telescope), stellar spectra can be obtained on an ever-increasing scale. Therefore, it is necessary to estimate stellar atmospheric parameters such as $T_{\rm eff}$, $\log g$ and [Fe/H] automatically to achieve the scientific goals and make full use of the potential value of these observations. Feature selection plays a key role in the automatic measurement of atmospheric parameters. We propose to use the least absolute shrinkage selection operator (Lasso) algorithm to select features from stellar spectra. Feature selection can reduce redundancy in spectra, alleviate the influence of noise, improve calculation speed and enhance the robustness of the estimation system. Based on the extracted features, stellar atmospheric parameters are estimated by the support vector regression model. Three typical schemes are evaluated on spectral data from both the ELODIE library and SDSS. Experimental results show the potential performance to a certain degree. In addition, results show that our method is stable when applied to different spectra.

Key words: methods: data analysis — stars: fundamental parameters — techniques: spectroscopic — surveys

1 INTRODUCTION

A fundamental problem in astrophysics is to explore the evolution and nature of stars in the Milky Way. Traditionally, this process has usually been done manually or interactively by investigating stellar populations, ages and chemical abundance. Recently, however, stellar spectra can be obtained on an ever-increasing scale with the rapid development of large sky survey projects, such as the Sloan Digital Sky Survey (SDSS, York et al. 2000), the follow-up Sloan Extension for Galactic Understanding and Exploration (SEGUE, Yanny et al. 2009), GAIA (Lobel 2011) and the Guoshoujing Telescope— also called the Large sky Area Multi-Object fiber Spectroscopic Telescope survey (LAMOST, Su et al. 1998; Zhao 2000; Zhu et al. 2006; Cui et al. 2012; Luo et al. 2012). Therefore, in order to achieve scientific goals and make full use of potential value contained in these observations, it is necessary to automatically estimate stellar atmospheric parameters such as $T_{\rm eff}$, $\log g$ and [Fe/H], a process which attracts much attention in the astronomical information processing community. For example, Koleva et al. (2009) designed the analysis software ULySS to automatically determine stellar atmospheric parameters. The fundamental parameters mainly include effective temperature ($T_{\rm eff}$), surface gravity ($\log g$) and metallicity ([Fe/H]), which are the basis for investigating the properties of stellar atmospheres. Therefore, research in this field has become a practical but still challenging task with the development of large-scale sky surveys.

During the last twenty years, progress in estimating parameters of stellar spectra has been developing rapidly in astronomy. Researchers have designed many algorithms to estimate fundamental atmospheric parameters based on stellar spectra. In essence, these methods can be roughly classified into two categories: minimum distance methods (MDMs) and Artificial Neural Networks (ANNs). MDM is a nearest neighbor algorithm. It constructs a spectral library in which parameters have been accurately identified according to physical measurements and compares the spectra being tested with a catalog of templates. Then the parameters of the templates with the highest similarity are regarded as the atmospheric parameters of the spectra being tested. The cross-correlation, weighted average algorithm and the k-nearest neighbors method can be regarded as examples of MDMs. The representative works include Katz et al. (1998); Fuentes & Gulati (2001); Zwitter et al. (2005); Jofré et al. (2010); Allende Prieto et al. (2006). By contrast the ANN method is a non-linear regression algorithm, which provides a mapping from the spectra to their atmospheric parameters as their outputs. There are many studies in this field, including Snider et al. (2001); Willemsen et al. (2005); Re Fiorentin et al. (2007); Manteiga et al. (2010); Bailer-Jones (2000).

Nowadays, researchers are further exploring automatic methods of estimating stellar atmospheric parameters in order to obtain more accurate results from spectra. Feature selection is a key procedure in these methods. Because feature selection can reduce irrelevant components which usually degrade the accuracy of estimation, eliminating irrelevant features also reduces the computing time required for estimating the model. Broadly speaking, feature selection is a procedure that decomposes a spectrum into different components, reorganizes them and selects the most appropriate feature based on some rules. Many researchers have applied principal component analysis (PCA) (Jolliffe 2002), wavelet (Lu et al. 2012), forward stagewise regression (Hastie et al. 2007) and other methods to select features from stellar spectra. We investigate the application of the least absolute shrinkage selection operator (Lasso) algorithm (Tibshirani 1996) to select features for estimating atmospheric parameters. Lasso is an L_1-norm regularization technique for linear regressions which is widely applied in machine learning and statistics. In this paper, our scheme for estimating atmospheric parameters consists of the following two steps: selecting spectral features by the Lasso algorithm and estimating the atmospheric parameters based on the support vector regression (SVR) model. In the first step, we choose the Lasso algorithm because it can find relevant features fast. In the second step, we apply the SVR model to estimate stellar parameters because it has a global optimum and exhibits better prediction accuracy due to its implementation of the structural risk minimization principle.

This paper is organized as follows. Section 2 introduces the data representing stellar spectra. Section 3 introduces the Lasso algorithm and SVR model. Section 4 presents the results and analysis. Finally, Section 5 gives the discussion and conclusions.

2 DATA

In this section, we describe the experimental data arising from ELODIE spectra and SDSS spectra. All the spectra are flux calibrated. Our model estimates the stellar atmospheric parameters on flux calibrated spectra. We conduct experiments on both synthetic spectra and real spectra.

In the first experiment, we choose a sample set from the ELODIE (Prugniel & Soubiran 2001, 2004; Prugniel et al. 2007) library, which is a synthetic spectral library based on stellar atmosphere models. The sample of stellar spectra consists of 1800 spectra with wavelength in the range $\lambda = 421 \sim 650$ nm, resolution $\Delta\lambda = 1$ nm, and parameter coverage $T_{\rm eff}$ from 3700 K to 13 386 K, $\log g$ from 0 dex to 4.80 dex and [Fe/H] from –2.94 dex to 1.0 dex.

In the second experiment, the set of stellar spectra is composed of 5000 stellar spectra which are from SDSS. All the 5000 stellar spectra are from 102 plates (0266–0367) of Date Release 7 (DR-7). Six hundred and forty spectra are simultaneously observed on one plate. About 1 200 000 spectra can be observed in DR-7. The 5000 spectra consist of different spectral types because they are randomly selected from these plates. The wavelength range is $10^{2.6} \sim 10^{2.9}$ nm, resolution $\Delta\lambda$ = 0.1 nm, and atmospheric parameter coverage $T_{\rm eff}$ from 4163 K to 9685 K, $\log g$ from 1.260 dex to 4.994 dex and [Fe/H] from –3.437 dex to 0.182 dex.

In order to reduce the range of $T_{\rm eff}$, we utilize $\log_{10} T_{\rm eff}$ rather than $T_{\rm eff}$ in this paper. Estimated values are accurate in noise-free data based on the results of synthetic spectra. In fact, the stellar spectra are influenced by noise, so our model is trained on real spectra from SDSS. The model's consistency and stability are demonstrated by testing different spectra.

The performance of the estimation method is evaluated based on the mean absolute error

$$\mu = \frac{1}{n}\sum_{i=1}^{n} |\text{error}_i|,$$

the mean error

$$\omega = \frac{1}{n}\sum_{i=1}^{n} (\text{error}_i),$$

and the standard deviation of error

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\text{error}_i - \omega)^2},$$

where error$_i$ is the deviation between the estimated value and the true value. Here i represents the ith spectrum, and n is the number of spectra being tested. The mean absolute error and the mean error indicate the degree of deviation between the estimated values and the true values. The standard deviation of error represents the level of dispersion of errors. We choose three kinds of errors to evaluate the method in order to analyze the results of the estimation more comprehensively.

3 LASSO AND THE SVR MODEL

In this section, we introduce the Lasso algorithm to select features and apply the SVR model to estimate fundamental stellar atmospheric parameters.

3.1 The Lasso Algorithm for Selecting Features

Tibshirani (1996) proposed the Lasso algorithm, which does not focus on selection of subsets but rather on defining a continuous shrinking operation that can produce coefficients of redundant components that are exactly zero. It has been shown in literature that the algorithm can effectively select the relevant features in high-dimensional data space.

The objective function of the Lasso algorithm is

$$(\hat{\alpha}, \hat{\beta}) = \text{arg min}\left\{\sum_{i=1}^{N}(y_i - \alpha - \sum_j \beta_j x_{ij})^2\right\} \quad \text{subject to} \sum_j |\beta_j| \leq t.$$

Suppose that we have data (x^i, y_i), $i = 1, 2, \ldots, N$, where $x^i = (x_{i1}, \ldots, x_{ip})^T$ is the predictor variable, y_i is the response, β_j is the regression coefficient of the jth variable and $t \geq 0$ is a tuning parameter. In this model, we assume that x_{ij} is standardized such that $\sum_i x_{ij}/N = 0$, with

$\sum_i {x_{ij}}^2/N = 1$. The Lasso problem is a quadratic programming problem with a constraint from the linear inequality. The number of coefficients are controlled by the parameter t. Let β_j^0 be the full set of least squares estimates and let $t_0 = \sum |\beta_j^0|$. Values of $t < t_0$ will cause shrinkage of the solutions towards 0, so some coefficients β_j may be exactly equal to 0. If the coefficient $\beta_j \neq 0$, then the jth row is a feature that can be included in the model. For example, we extract 12-dimensional features from ELODIE spectra according to the surface temperature in this paper. The extracted feature vector is $X_i = (x_{i6}, x_{i10}, x_{i11}, x_{i28}, x_{i65}, x_{i66}, x_{i94}, x_{i162}, x_{i169}, x_{i172}, x_{i204}, x_{i226})$. It means that the corresponding $\beta_j \neq 0$ for $j =$ 6,10,11,28,65,66,94,162,169,172,204,226. The selected features represent more relevant information from the spectra. We use the same method to select features for other parameters. The parameters can be effectively estimated according to a small number of features. As is well known, stellar spectra contain much noise and irrelevant information. It is not necessary to estimate stellar parameters with the full spectra. Therefore, feature selection is important. For the Lasso algorithm, the features are selected by stacking the data components with nonzero coefficients into a vector.

Feature selection in Lasso depends on the regularization parameters, and the set of solutions for all values of these free parameters is provided by the regularization path (Hastie et al. 2004). There are many algorithms that can be used to solve the Lasso problem. In this paper, we use a classical algorithm, the Least Angle Regression (LAR, Efron et al. 2004), to solve the Lasso problem, which can adaptively estimate the regularization parameters. It has been shown that the Lasso algorithm is robust and fast.

3.2 Support Vector Regression

SVR is a non-linear kernel regression method which tries to find the best regression hyperplane with the smallest structural risk in a high-dimensional feature space (Yeh et al. 2011). Today, it is widely used in business applications and data analysis. In this paper, we apply the ε-insensitive support vector regression (ε-SVR) (Drucker et al. 1997; Smola & Schölkopf 2004) to estimate the fundamental stellar atmospheric parameters.

Suppose that we have a set of training spectra $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $x_i \in R^d$ is a spectrum, and $y_i \in R$ is the target value (the stellar atmospheric parameters here) for the spectrum x_i, where n and d represent the number of training spectra and the number of variables in each spectrum respectively. The SVR model is set up by training spectra and it is used to estimate the parameters of stellar spectra. At present, ε-SVR is widely used in many fields. The ε-SVR is formulated as follows.

$$\min_{\omega,\xi_i,\xi_i^*} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{1}$$

$$\text{s.t.} \begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i \geq 0, \xi_i^* \geq 0, i = 1, \ldots n, \end{cases}$$

where ω is a vector of weights for the variables, and b is a constant. The constant C expresses a trade-off between penalty for errors beyond ε and the size of the weights. The parameters ξ_i and ξ_i^* are slack variables. We define the dual optimization problem (2) of the ε-SVR formulation as follows.

$$\max_{\alpha,\alpha^*} \left\{ -\varepsilon \sum_{i=1}^{l}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) \right\}, \tag{2}$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0\\ 0 \leq \alpha_i^* \leq C, i = 1,\ldots,l\\ 0 \leq \alpha_i \leq C, i = 1,\ldots,l.\end{cases}$$

The optimal problem (2) is easier to solve due to researchers providing many nonlinear kernel functions, such as the polynomial kernel function, the radial basis function (RBF), the sigmoid kernel function and so on. The estimation function $f(x)$ is given as

$$f(x) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(x_i, x_j) + b, \tag{3}$$

where α_i^* and α_i are nonzero Lagrangian multipliers for the dual problem. The kernel function $K(x_i, x_j)$ represents the inner product. In this paper, we use the RBF as the kernel function.

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \tag{4}$$

where $\gamma > 0$. The RBF can avoid difficulties associated with very large numbers because its value ranges between 0 and 1. The RBF offers the ability to learn a variety of nonlinear relationships. The spectral data have millions of observations and more than a thousand variables. Therefore, we apply the ε-SVR model to deal with the large scale spectral data.

3.3 Estimating Scheme for Fundamental Parameters of Stellar Atmospheres

In this section, we introduce how the spectral data are pre-processed. Each spectrum is normalized before it is trained and tested, and we describe the process used by the estimation algorithm.

A stellar spectrum contains two components: continuum and lines. In spectroscopy, the shape of the continuum is determined by T_{eff}, and the lines are mainly determined by the combination of T_{eff}, $\log g$ and metallicity. Stellar spectra can be divided into seven types of stars according to different temperatures. The seven types of stars cover different temperature ranges: A, 7500–11000 K; B, 11 000–25 000 K; F, 6000–7500 K; G, 5000–6000 K; K, 3500–5000 K; M, <3500 K and O, >25 000 K. At present, researchers can obtain relatively accurate values of T_{eff}, but it is difficult to achieve precise values of $\log g$ and [Fe/H]. In this paper, we use the Lasso method to select features in order to estimate parameters more accurately.

The stellar spectral data are divided into a training set and a testing set. According to the Lasso algorithm, we obtain a feature vector of spectra, as given by (X_i, Y_i). Variable X_i is a feature vector and Y_i is the corresponding parameter.

Since the flux of each spectrum has a different scale, we preprocess experimental data after feature selection. That is a key operation for estimating fundamental stellar parameters. The data preprocessing is shown as follows.

$$X_{ij} = X_{ij}\Bigg/\sqrt{\sum_{j=1}^{p} X_{ij}^2} \qquad (i = 1, 2, \ldots, n, j = 1, 2, \ldots, p), \tag{5}$$

where i is the number of the spectrum and j is the number of the feature, and X_{ij} is an $n \times p$ matrix. We use Equation (5) to preprocess the training data and the testing data. For each spectrum x_i being tested, parameters are estimated through the regression function $f(x)$.

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section we show the estimated results of the SVR model and related analysis.

Table 1 Three kinds of statistical errors for the method of estimating ELODIE spectral parameters with different resolutions and noises (μ is the mean absolute error, ω is the mean error and σ is the standard deviation of errors).

Resolution (nm)	SNR	$\mu_{\log T_{\rm eff}}$	$\omega_{\log T_{\rm eff}}$	$\sigma_{\log T_{\rm eff}}$	$\mu_{\log g}$	$\omega_{\log g}$	$\sigma_{\log g}$	$\mu_{\rm [Fe/H]}$	$\omega_{\rm [Fe/H]}$	$\sigma_{\rm [Fe/H]}$
1	full	0.0142	−0.0046	0.0267	0.2132	−0.0107	0.3463	0.1772	−0.0236	0.3226
1	80	0.0145	−0.0056	0.0269	0.2293	−0.0076	0.3664	0.1877	−0.0194	0.3302
1	50	0.0328	−0.0056	0.0493	0.5205	0.0745	0.7451	0.3049	−0.0548	0.4507
1	30	0.0670	0.0122	0.0897	0.6212	−0.0772	0.8697	0.3312	−0.1626	0.4089
2	full	0.0163	−0.0058	0.0275	0.2462	−0.0049	0.3774	0.2026	0.0097	0.3285
2	80	0.0163	−0.0056	0.0271	0.2696	−0.0159	0.4061	0.2035	0.0033	0.3282
2	50	0.0391	0.0099	0.0587	0.5305	0.0304	0.7936	0.2888	−0.0391	0.4398
2	30	0.0771	0.0035	0.1035	0.6409	0.0739	0.8781	0.3292	−0.1581	0.4132
5	full	0.0188	−0.0049	0.0301	0.2878	−0.0166	0.4684	0.1869	−0.0173	0.3211
5	80	0.0191	−0.0067	0.0305	0.3015	−0.0192	0.4903	0.1953	−0.0053	0.3273
5	50	0.0415	−0.0041	0.0671	0.5486	0.0055	0.9231	0.2706	−0.0043	0.4482
5	30	0.0993	−0.0007	0.1334	0.6880	0.0903	0.9829	0.3532	−0.1131	0.4802

Table 2 The mean absolute error of $\log T_{\rm eff}$, $\log g$ and [Fe/H] estimated by the KLSR model, the KPCR model, the non-parametric regression model and our method (SVR model).

Model	$\log T_{\rm eff}$	$\log g$	[Fe/H]
KLSR model	0.0224	0.9378	0.2848
KPCR model	0.0170	0.3509	0.2471
Non-parametric model	0.0153	0.3768	0.2695
SVR model	0.0142	0.2132	0.1772

4.1 Application to ELODIE Stellar Spectral Library

In this experiment, our spectral data come from the ELODIE library. Some researchers have designed many models to estimate stellar parameters with the ELODIE library. Representative works have included the non-parametric model (Zhang et al. 2005), the KLSR model (Zhang et al. 2009) and the KPCR model (Zhang et al. 2009). Our sample consists of 1800 spectra from the ELODIE library. All spectra are divided into two parts. One set (900 spectra) is selected for training the SVR model, while the other set (900 spectra) is used for testing. We extract the 12-dimensional features according to the surface temperature, the 16-dimensional features considering the surface gravity and 15 dimensional features using metallicity by the Lasso algorithm. Finally, the parameters are estimated by the SVR model.

In order to test the stability of our method, all the spectral data are changed in resolution from $\Delta\lambda = 1\,$nm to $\Delta\lambda = 2\,$nm and $\Delta\lambda = 5\,$nm. The experimental results are listed in Table 1. To check the robustness of estimation for the model, additional noise is introduced into the spectra with signal-to-noise ratio (SNR) of 80, 50 and 30. We estimate the parameters with different SNRs and various resolutions. All the statistical errors are listed in Table 1. In Table 2, the experimental results of the KLSR model, the KPCR model and the non-parametric model are extracted from the corresponding references Zhang et al. (2009) and Zhang et al. (2005) respectively. Table 3 presents three kinds of statistical errors for different types of stars.

The mean absolute error of three parameters is very small, especially for $T_{\rm eff}$. Figure 1 compares the estimated results with the true values from the ELODIE library.

Table 2 and Figure 1 show that the SVR model is accurate for estimating parameters on synthetic spectra. Table 1 shows the mean errors and the standard deviations of errors are more robust than the

Table 3 Three kinds of statistical errors for the different stellar types with the estimated results using ELODIE spectral parameters (μ is the mean absolute error, ω is the mean error and σ is the standard deviation of error).

$T_{\rm eff}$ (K)	$\mu_{\log T_{\rm eff}}$	$\omega_{\log T_{\rm eff}}$	$\sigma_{\log T_{\rm eff}}$	$\mu_{\log g}$	$\omega_{\log g}$	$\sigma_{\log g}$	$\mu_{\rm [Fe/H]}$	$\omega_{\rm [Fe/H]}$	$\sigma_{\rm [Fe/H]}$
K:3500,5000	0.0118	0.0002	0.0157	0.3222	−0.0444	0.4278	0.1408	−0.0181	0.2167
G:5000,6000	0.0064	−0.0004	0.0097	0.1725	−0.0316	0.2564	0.0958	−0.0101	0.1315
F:6000,7500	0.0124	0.0044	0.0164	0.1923	0.0665	0.3668	0.2407	−0.0349	0.4484
A:7500,11000	0.0225	−0.0152	0.0362	0.2010	−0.0832	0.3651	0.2865	−0.0150	0.4484
B:11000,25000	0.0566	−0.0566	0.0558	0.1342	−0.0787	0.2151	0.2789	−0.1114	0.4661

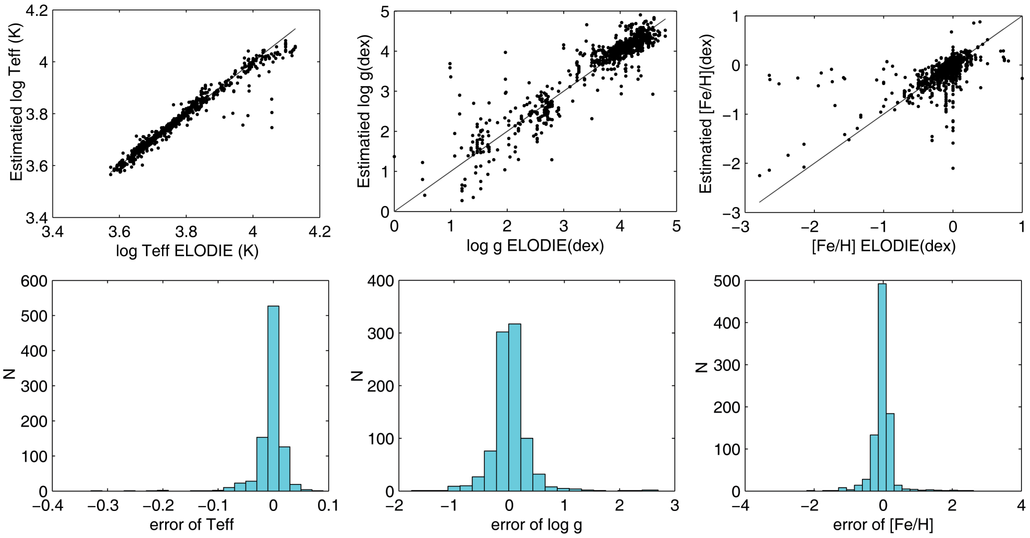

Fig. 1 Comparison of the true parameters with estimated results of the atmospheric parameters using the ELODIE spectra. The solid lines show a perfect correlation. The bottom panels show histograms of errors associated with the parameters.

mean absolute errors when using different resolutions and noises. From Table 1, we conclude that parameters are estimated more accurately when there is low noise, but the resolutions still influence the accuracy and robustness.

Table 2 shows the accuracy of our model is better than other models, especially for $\log g$. Moreover, the three kinds of statistical errors show that the three parameters ($T_{\rm eff}$, $\log g$ and [Fe/H]) estimated by the SVR are in agreement. The histograms show our model is feasible since the discrepancies are shown in the lower panels.

The above results are obtained with synthetic data. In the following, we will use the real spectra from SDSS to estimate parameters in the next section.

4.2 Application to SDSS Spectra

In this experiment, the real stellar spectra from SDSS are selected for verifying the SVR model. We select a sample of 5000 stellar spectra in the experiment. Here a set of 2500 spectra is used as the training set and the other 2500 spectra are the testing set. We extract the 22-dimensional features according to the surface temperature, the 47-dimensional features using the surface gravity and the

Table 4 Three kinds of statistical errors for the method of estimating SDSS spectral parameters with different resolutions (μ is the mean absolute error, ω is mean error and σ is the standard deviation of error).

Resolution (nm)	$\mu_{\log T_{\rm eff}}$	$\omega_{\log T_{\rm eff}}$	$\sigma_{\log T_{\rm eff}}$	$\mu_{\log g}$	$\omega_{\log g}$	$\sigma_{\log g}$	$\mu_{\rm [Fe/H]}$	$\omega_{\rm [Fe/H]}$	$\sigma_{\rm [Fe/H]}$
0.1	0.0097	–0.0002	0.0142	0.2724	–0.0253	0.3834	0.2728	–0.0195	0.3629
0.2	0.0094	0.0002	0.0140	0.2672	0.0238	0.3777	0.2848	–0.0187	0.3762
0.5	0.0136	–0.0013	0.0184	0.2940	0.0386	0.4173	0.3141	–0.0253	0.4107

Table 5 Three kinds of statistical errors for the different stellar types with the estimated results of SDSS spectral parameters (μ is the mean absolute error, ω is the mean error and σ is the standard deviation of error).

$T_{\rm eff}$ (K)	$\mu_{\log T_{\rm eff}}$	$\omega_{\log T_{\rm eff}}$	$\sigma_{\log T_{\rm eff}}$	$\mu_{\log g}$	$\omega_{\log g}$	$\sigma_{\log g}$	$\mu_{\rm [Fe/H]}$	$\omega_{\rm [Fe/H]}$	$\sigma_{\rm [Fe/H]}$
K: 3500, 5000	0.0114	0.0071	0.0133	0.1867	–0.0053	0.2931	0.2575	0.0927	0.3317
G: 5000, 6000	0.0088	–0.0003	0.0123	0.2342	0.0341	0.3456	0.2233	–0.0140	0.2878
F: 6000, 7500	0.0090	–0.0005	0.0126	0.3079	0.0217	0.4172	0.2914	–0.0533	0.3779
A: 7500, 11000	0.0131	–0.0068	0.0207	0.3609	0.0395	0.4519	0.3778	–0.0414	0.4917

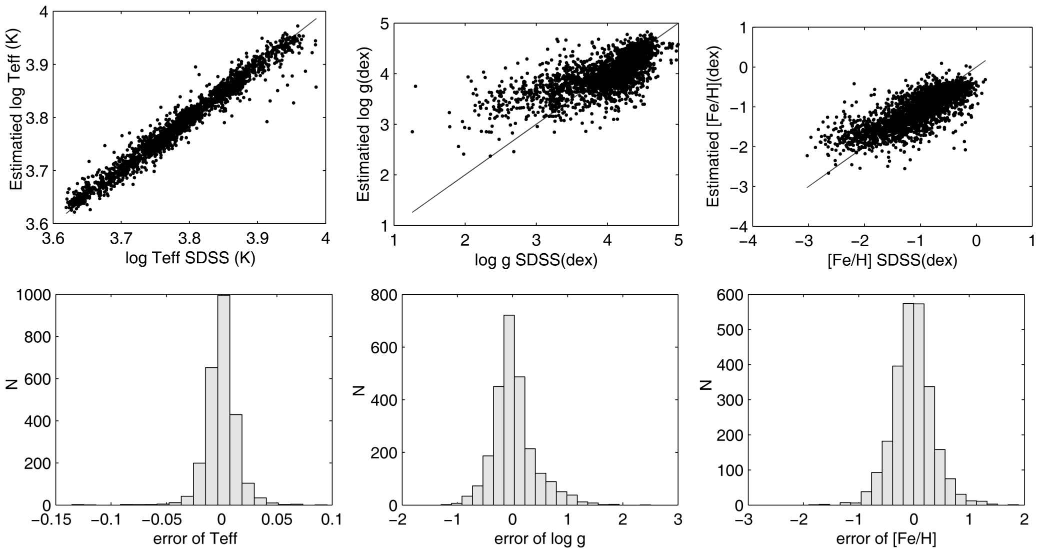

Fig. 2 Comparison of the target parameters with the estimated atmospheric parameters of SDSS spectra. The solid lines show a perfect correlation. The bottom panels show histograms of the parameter errors.

21-dimensional features considering metallicity. Then the parameters are calculated by the SVR model. The resolution is degraded from 0.1 nm to 0.2 nm and 0.5 nm in order to test robustness of our model. Three kinds of statistical errors when using different resolutions are listed in Table 4. Table 5 shows three kinds of errors for different types of stars. Figure 2 presents a comparison between estimated values and target values.

From Table 4 and Figure 2, we conclude that our model is feasible and robust for lower resolution spectra at 0.2nm and 0.5nm. Table 5 shows the errors in different types of stars.

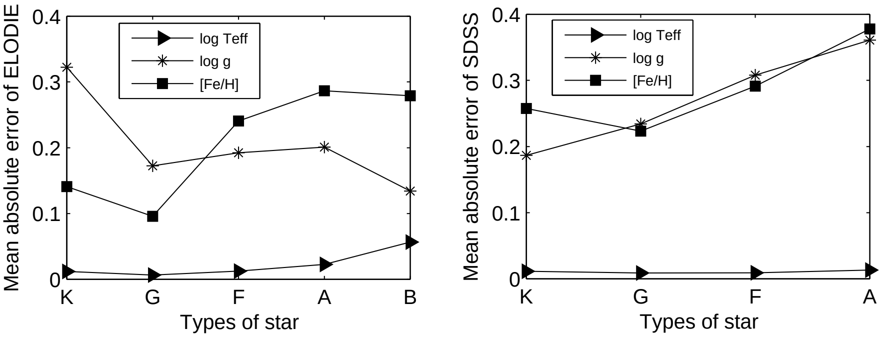

Fig. 3 Curves showing the average absolute error with different types of stars.

Figure 3 shows curves representing the average absolute error as a function of different stars. Based on the above results, we conclude that our model can estimate the parameters of large surveys with various resolutions. Our method is stable and efficient, especially for $T_{\rm eff}$.

5 DISCUSSION AND CONCLUSIONS

In this paper, we apply the SVR model to estimate three fundamental parameters ($T_{\rm eff}$, $\log g$ and [Fe/H]) using the method of Lasso feature selection. The initial process of feature selection determines the accuracy of results. In the first experiment, we select 6%–8% of features as the support vectors. In the second experiment, we select only 0.7%–1% of features as the support vectors. The experimental results show that our model estimates stellar parameters relatively accurately.

From Table 1 and Table 4, we can conclude that the model is robust for different resolutions. From Table 3, Table 5 and Figure 3, we conclude the results are consistent with different types of stars. For the effective temperature, the results are relatively accurate for all stars. For surface gravity and metallicity, estimated values of G type stars are better than those of other types of stars. From experimental results, the main advantages of our model are as follows:

(1) The complexity of the SVR is lower because the Lasso algorithm selects fewer features as support vectors such that the reliability and efficiency of the model are enhanced. The results show that the model is stable and robust for estimation of stellar parameters, especially for SDSS spectra.
(2) The training and estimation procedures are rapid after feature selection. The first experiment on ELODIE spectra in Section 4.1 costs less than 30 s. The second experiment on SDSS spectra in Section 4.2 costs less than 4 min on an AMD A6 personal computer with a 1.4 GHz processor.
(3) Our model is very convenient to apply, because there are many algorithms that can be used to solve the Lasso problem. Here we choose the SVR model to estimate stellar parameters because it is widely applied in the fields of statistics and machine learning, and is regarded as a mature algorithm.

In conclusion, our model can effectively estimate fundamental parameters in synthetic spectra and real spectra. The model estimates each parameter independently. In fact, the three fundamental parameters have intrinsic correlations due to their physics. Therefore, we may consider their interaction in order to estimate parameters more accurately and effectively. The following work will be multi-task regression applied to estimating fundamental stellar parameters.

Acknowledgements We are very grateful to the referees for their helpful comments and suggestions.

References

Allende Prieto, C., Beers, T. C., Wilhelm, R., et al. 2006, ApJ, 636, 804

Bailer-Jones, C. A. L. 2000, A&A, 357, 197

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1197

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. 1997, Advances in Neural Information Processing Systems, 9, 155

Efron, B., Hastie, T., Johnstone, I., et al. 2004, The Annals of Statistics, 32, 407

Fuentes, O., & Gulati, R. K. 2001, in Revista Mexicana de Astronomia y Astrofisica Conference Series, 10, The Seventh Texas-Mexico Conference on Astrophysics: Flows, Blows and Glows, eds. William H. Lee, & Silvia Torres-Peimbert, 209

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. 2004, Journal of Machine Learning Research, 5, 1391

Hastie, T., Taylor, J., Tibshirani, R., et al. 2007, Electronic Journal of Statistics, 1, 1

Jofré, P., Panter, B., Hansen, C. J., & Weiss, A. 2010, A&A, 517, A57

Jolliffe, I. 2002, Principal Component Analysis (2nd edn.; New York: Springer-Verlag)

Katz, D., Soubiran, C., Cayrel, R., Adda, M., & Cautain, R. 1998, A&A, 338, 151

Koleva, M., Prugniel, P., Bouchard, A., & Wu, Y. 2009, A&A, 501, 1269

Lobel, A. 2011, in Journal of Physics: Conference Series, 328, 012027 (IOP Publishing)

Lu, Y., Li, C. L., Li, X. R., et al. 2012, Spectroscopy and Spectral Analysis, 32, 2583

Luo, A.-L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1243

Manteiga, M., Ordóñez, D., Dafonte, C., & Arcay, B. 2010, PASP, 122, 608

Prugniel, P., & Soubiran, C. 2001, A&A, 369, 1048

Prugniel, P., & Soubiran, C. 2004, arXiv: astro-ph/0409214

Prugniel, P., Soubiran, C., Koleva, M., & Le Borgne, D. 2007, VizieR Online Data Catalog, 3251, 0

Re Fiorentin, P., Bailer-Jones, C. A. L., Lee, Y. S., et al. 2007, A&A, 467, 1373

Smola, A. J., & Schölkopf, B. 2004, Statistics and Computing, 14, 199

Snider, S., Allende Prieto, C., von Hippel, T., et al. 2001, ApJ, 562, 528

Su, D. Q., Cui, X., Wang, Y., & Yao, Z. 1998, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 3352, Advanced Technology Optical/IR Telescopes VI, ed. L. M. Stepp, 76

Tibshirani, R. 1996, Journal of the Royal Statistical Society. Series B (Methodological), 58, 267

Willemsen, P. G., Hilker, M., Kayser, A., & Bailer-Jones, C. A. L. 2005, A&A, 436, 379

Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, AJ, 137, 4377

Yeh, C. Y., Huang, C. W., & Lee, S. J. 2011, Expert Systems with Applications, 38, 2177

York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579

Zhang, J. N., Wu, F. C., Luo, A. L., & Zhao, Y. H. 2005, Acta Astronomica Sinica, 46, 406

Zhang, J. N., Wu, C. F., & Luo, A. L. 2009, Spectroscopy and Spectral Analysis, 29, 1131

Zhao, Y. 2000, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 4010, Observatory Operations to Optimize Scientific Return II, ed. P. J. Quinn, 290

Zhu, Y., Hu, Z., Zhang, Q., Wang, L., & Wang, J. 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 6269, 62690M

Zwitter, T., Munari, U., & Siebert, A. 2005, in The Three-Dimensional Universe with Gaia (ESA Special Publication), 576, eds. C. Turon, K. S. O'Flaherty, & M. A. C. Perryman, 623