Research in Astron. Astrophys. **2009** Vol. **9** No. **6**, 712 – 724
http://www.raa-journal.org *http://www.iop.org/journals/raa*

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

Automated estimation of stellar fundamental parameters from low resolution spectra: the PLS method *

Jian-Nan Zhang, A-Li Luo and Yong-Heng Zhao

National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China;
jnzhang@lamost.org

Received 2008 September 5; accepted 2009 January 7

Abstract PLS (Partial Least Squares regression) is introduced into an automatic estimation of fundamental stellar spectral parameters. It extracts the most correlative spectral component to the parameters ($T_{\rm eff}$, log g and [Fe/H]), and sets up a linear regression function from spectra to the corresponding parameters. Considering the properties of stellar spectra and the PLS algorithm, we present a piecewise PLS regression method for estimation of stellar parameters, which is composed of one PLS model for $T_{\rm eff}$, and seven PLS models for log g and [Fe/H] estimation. Its performance is investigated by large experiments on flux calibrated spectra and continuum normalized spectra at different signal-to-noise ratios (SNRs) and resolutions. The results show that the piecewise PLS method is robust for spectra at the medium resolution of 0.23 nm. For low resolution 0.5 nm and 1 nm spectra, it achieves competitive results at higher SNR. Experiments using ELODIE spectra of 0.23 nm resolution illustrate that our piecewise PLS models trained with MILES spectra are efficient for O $\sim$ G stars: for flux calibrated spectra, the systematic offsets are 3.8%, 0.14 dex, and –0.09 dex for $T_{\rm eff}$, log g and [Fe/H], with error scatters of 5.2%, 0.44 dex and 0.38 dex, respectively; for continuum normalized spectra, the systematic offsets are 3.8%, 0.12 dex, and –0.13 dex for $T_{\rm eff}$, log g and [Fe/H], with error scatters of 5.2%, 0.49 dex and 0.41 dex, respectively. The PLS method is rapid, easy to use and does not rely as strongly on the tightness of a parameter grid of templates to reach high precision as Artificial Neural Networks or minimum distance methods do.

Key words: methods: data analysis — methods: statistical — stars: fundamental parameters (classification, temperatures, metallicity) — techniques: spectroscopic — surveys

1 INTRODUCTION

The development of modern telescopes, such as SDSS, LAMOST, etc. gives researchers the opportunity to study large scale kinematic and chemical structure of the Galaxy by using great amounts of stellar spectra. Using these spectra, researchers can derive fundamental physical parameters: the effective temperature ($T_{\rm eff}$), gravity (log g) and metallicity ([Fe/H]), which are some of the most interesting properties of a star. For medium and low resolution spectra obtained by modern telescopic surveys, it is necessary to develop automatic spectral analysis technologies to obtain the parameters of stellar spectra from the data. At present, researchers have applied PCA (Principal Component Analysis), ANNs (Artificial Neural Networks), and other data analysis technologies to stellar spectral analysis.

* Supported by the National Natural Science Foundation of China.

The existing automated stellar spectral parameter measurement methods could be grouped into two kinds: MDM (minimum distance methods) and ANN methods. MDM firstly constructs a spectral template library in which their parameters have been identified accurately by traditional physical measurements. A spectrum to be estimated is then compared with each of the templates while a function is defined to describe the similarity between the estimated spectrum and the templates. Parameters of the template with the highest similarity are then assigned to the estimated spectrum. The k-nearest neighbor algorithm, weighted average algorithm, χ^2 technique and template matching method are variations of the MDM algorithm. The representative research is the ELODIE online stellar parameters estimation system by Katz et al. (1998) and Soubiran (1998). They constructed a stellar spectral template library composed of 211 F – K type stellar spectral templates with a resolution of 0.1 Å. With SNR = 100 data, the precision reached values of: $T_{\rm eff}$: 86 K, $\log g$: 0.28 dex, [Fe/H]: 0.16 dex; and with SNR = 10 data, $T_{\rm eff}$: 102 K, $\log g$: 0.29 dex, [Fe/H]: 0.17 dex. Fuentes & Gulati (2001) presented a distance-weighted-nearest-neighbor algorithm, and the input data included spectra and spectral indices. Recio-Blanco et al. (2006) designed the MATISSE algorithm to automatically derive the parameters and chemical abundances for the Gaia/RVS survey, with a library of synthetic spectra as the templates. An object's spectral parameters are linear combinations of the parameters of the most correlated template spectra. This algorithm is essentially a locally weighted average algorithm. Prieto (2003) performed MDM to determine stellar atmospheric parameters on $R \simeq 5000$ resolution spectra of A – K stars. To find the best fit with a template spectrum, he used a genetic algorithm when implementing MDM. Bonifacio & Caffau (2003) applied the χ^2 method to automatically determine the [Fe/H] of $R = 15\,000$ spectra for the giants in the Sagittarius dwarf spheroidal galaxy.

The other popularly applied algorithm is ANNs. It is a non-linear regression algorithm. We have little knowledge of the mapping process performed by ANNs between a spectrum and its estimated parameters, because an ANN is a black-box which cannot provide any information about the mapping relationship. ANNs for estimating stellar parameters has a shortcoming: the structure of an ANN (the number of layers, the parameters of mapping function, etc.) is often variable and is determined by experience and convention. Usually, the more sophisticated the ANN structure, the longer the training time for the ANN. Many researchers designed different ANNs to estimate stellar spectral parameters. Snider et al. (2001) explored a back-propagation ANN for the estimation of atmospheric parameters for Galactic F- and G-type stars. The ANN is fed with medium-resolution spectra ($\Delta\lambda = 1$ – 2 Å) and the accuracy of $\sigma(T_{\rm eff}) = 135$ – 150 K over the range $4250 \leq T_{\rm eff} \leq 6500$ K, $\sigma(\log g) =$ 0.25 – 0.30 dex over the range $1.0 \leq \log g \leq 5.0$ dex, and $\sigma([{\rm Fe/H}]) = 0.15$ – 0.20 dex over the range $-4.0 \leq [{\rm Fe/H}] \leq 0.3$ dex. Bailer-Jones (2000) developed a parameterization system based on a feedforward multilayer perceptron ANN with two hidden layers. It was shown to provide accurate three-dimensional physical parameterization of synthetic Kurucz model stellar spectra. Willemsen et al. (2005) employed a feedforward ANN, trained on synthetic spectra in a 1800 Å region around 4700 Å ($R \simeq 1500 - 2400$), to determine metallicities of main-sequence turn-off, subgiant and red giant stars in two globular clusters.

In this paper, PLS is introduced to the problem of estimating stellar parameters. PLS regression extracts the components from spectra that correlate with the parameters and builds up a linear regression between the spectral components and the corresponding parameters. This special property explores a new method of parameter estimation for a stellar spectrum. Considering the properties of stellar spectra, we designed a piecewise PLS method for parameterization of a stellar spectrum in this paper. Once the linear regression model is set up by PLS, for a spectrum in which the parameters need to be determined, it is only necessary to multiply it by the regression model's coefficients. We investigated the piecewise PLS with low resolution optical spectra. Section 2 introduces PLS and the framework of piecewise PLS for estimating stellar parameters; Section 3 describes the experiments and results on low resolution spectra at different SNR; Based on the experiments in Section 3, the paper gives discussion and conclusions in Section 4.

2 PLS REGRESSION AND PIECEWISE PLS METHOD

PLS regression is also called regression by means of projections to latent structures (PLS). It was developed between 1975 and 1982 by Herman Wold and co-workers, under the name 'partial least squares modeling in latent variables' (Wold 2001; Wang 1999). PLS regression generalizes and combines features from principal component analysis and multiple regression. Its goal is to predict a set of dependent variables from a set of explanatory variables. This prediction is achieved by extracting a set of latent variables (correlative component) from the explanatory variables which have the best predictive power. In this way, PLS sets up the linear regression from explanatory variables to dependent variables. PLS regression is particularly useful when we need to predict a set of dependent variables from a large set of explanatory variables. A number of important papers have discussed the objective function and statistical properties of PLS. Today, it is widely used in chemometrics (i.e. computational chemistry), economic analysis and other areas of data analysis, especially in the spectral analysis domain of chemometrics. Estimation of parameters from stellar spectra is similar to many solutions of PLS application in chemometrics. In this paper, we use the PLS method to set up the linear regression function from stellar spectra to stellar fundamental parameters ($T_{\rm eff}$, $\log g$, [Fe/H]). Once the regression function is set up, an object's spectral parameters are estimated by just inputting the spectrum into the regression model.

2.1 Regression Model for Estimating Stellar Fundamental Parameters from the Spectrum

Consider the linear regression model:

$$Y = X\beta + \varepsilon = \sum_{k=1}^{p} X_{.k}\beta_{k.} + \varepsilon \,, \tag{1}$$

where $X = (x_1, \cdots, x_n)^T \in R^{n\times p}$, n p-dimensional explanatory variables are collected in an $n \times p$ matrix (a p-dimensional spectrum data is denoted as x, which is a column variable). The kth column of X is expressed as $X_{.k}$ with $k = 1, \cdots, p$, and the ith row of X is expressed as $X_{i.}$ with $i = 1, \cdots, n$. $Y = (y_1, \cdots, y_n)^T \in R^{n\times m}$, n m-dimensional dependent variables are collected in an $n \times m$ matrix (the parameters of a stellar spectrum are denoted as a 3-dimensional variable $y = (T_{\rm eff}, \log\, g, [\text{Fe/H}])^T)$, where m is 3. $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \in R^{n\times m}$, are the residuals. $\beta = (\beta_1, \cdots, \beta_m) \in R^{p\times m}$, are the coefficients to be estimated. The kth row of β is expressed as $\beta_{k.}$ with $k = 1, \cdots, p$.

Given samples of stellar spectra as X with the parameters as Y, the linear regression coefficients are determined by the PLS regression method. For a spectrum x, its parameters are estimated through the regression function: $y^* = x\beta$.

PLS principle: PLS regression finds components from X that are also relevant for Y. Specifically, PLS regression searches for a set of components (called latent vectors) that performs a simultaneous decomposition of X and Y with the constraint that these components explain as much of the covariance between X and Y as possible. This step generalizes PCA. It is followed by a regression step where the decomposition of X is used to predict Y. For the mathematical principles of PLS regression, see Appendix. There are many algorithms for PLS regression. In our work, the NIPALS algorithm of Wold et al. is used for the estimation of stellar spectral parameters, which is given in the Appendix.

2.2 Piecewise PLS for Stellar Parameter Estimation

A stellar spectrum is composed of two components: continuum and lines. For the flux calibrated and flat field corrected spectrum, the shape of the continuum is determined by $T_{\rm eff}$, and the lines are mainly determined by the combination of $T_{\rm eff}$, $\log\, g$ and metallicity. Compared with $T_{\rm eff}$, it is difficult to derive precise values of $\log g$ and metallicity from low resolution spectra. In order to better separate the influence of each parameter on the spectra, we designed the piecewise PLS method to improve the $\log g$ and [Fe/H] estimation. The piecewise PLS framework is illustrated in Figure 1: one PLS regression

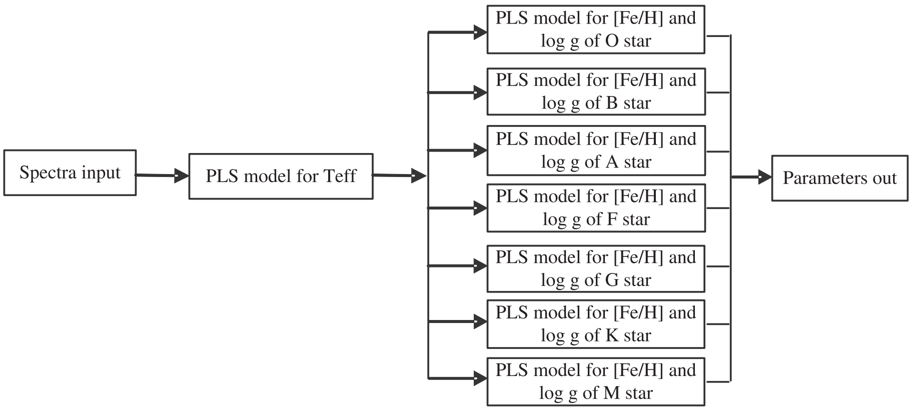
Spectra input
PLS model for Teff
PLS model for [Fe/H] and log g of O star
PLS model for [Fe/H] and log g of B star
PLS model for [Fe/H] and log g of A star
PLS model for [Fe/H] and log g of F star
PLS model for [Fe/H] and log g of G star
PLS model for [Fe/H] and log g of K star
PLS model for [Fe/H] and log g of M star
Parameters out

Fig. 1 Piecewise PLS framework for estimation of stellar spectral parameters.

model for $T_{\rm eff}$, and seven PLS regression models for $\log g$ and [Fe/H] estimation, each for one of the seven types of stars. The seven types of stars cover a temperature range: O, $>$ 25 000 K; B, 11 000–25 000 K; A, 7500–11 000 K; F, 6000–7500 K; G, 5000–6000 K; K, 3500 – 5000 K; M, $<$ 3500 K. First, the $T_{\rm eff}$ is estimated, then according to the $T_{\rm eff}$, the spectrum is input into the corresponding regression model to estimate $\log\ g$ and [Fe/H]. In this way, PLS extracts the most variation of spectra which correlates with the variation in $T_{\rm eff}$. For each type of spectra where $T_{\rm eff}$ is limited to a finite range, the resulting variation in $T_{\rm eff}$ is greatly minimized because PLS extracts the components of spectra with the largest variation which correlate to the components of $\log\ g$ and [Fe/H]. In order to illustrate the performance of the PLS method, the next section describes its applications to the estimation of stellar atmospheric parameters ($T_{\rm eff}$, $\log\ g$, [Fe/H]) on flux calibrated spectra, continuum normalized spectra and different source spectra.

3 APPLICATION OF PLS REGRESSION FOR STELLAR PARAMETER ESTIMATION

3.1 Application to Flux Calibrated Spectra

The piecewise PLS method was investigated using stellar spectra from the MILES library (Sánchez-Blázquez 2006). The MILES stellar library consists of 945 spectra: wavelength: λ = 350 nm $\sim$ 743 nm; resolution: $\Delta\lambda$ = 0.23 nm. The spectra were randomly split into two groups, one set (472 spectra) for PLS model training, and the other set (473 spectra) for testing.

Parameter coverage: $T_{\rm eff}$ from 3400 K to 36 000 K, $\log g$ from 0.2 to 5.86 dex and [Fe/H] from –2.8 to +0.7 dex. Data pre-processing: each spectrum was normalized to have unit standard deviation before training and testing. $\log_{10} T_{\rm eff}$ (rather than $T_{\rm eff}$) was used in the regression model to reduce the dynamic range of this parameter.

Figure 2 compares the result of the test set computed by the piecewise PLS model with the values from the MILES library, and the error histograms are also illustrated which are fitted by Gaussian curves respectively. The inset data of a, b and c in the figures of error histogram are the parameters of Gaussian curve function: $f(x) = a * e^{-(x-b/c)^2}$. The error of $T_{\rm eff}$ is expressed as the ratio of difference to the true $T_{\rm eff}$, the error of $\log g$ is the difference between the estimated $\log g$ and the true $\log g$, and the error of metallicity is the difference between the estimated [Fe/H] and the true [Fe/H]. Figure 2 shows that the three parameters ($T_{\rm eff}$, $\log g$, and [Fe/H]) estimated by the piecewise PLS method are in good agreement with the values of MILES. The regression models built up by piecewise PLS are feasible for parameter estimation.

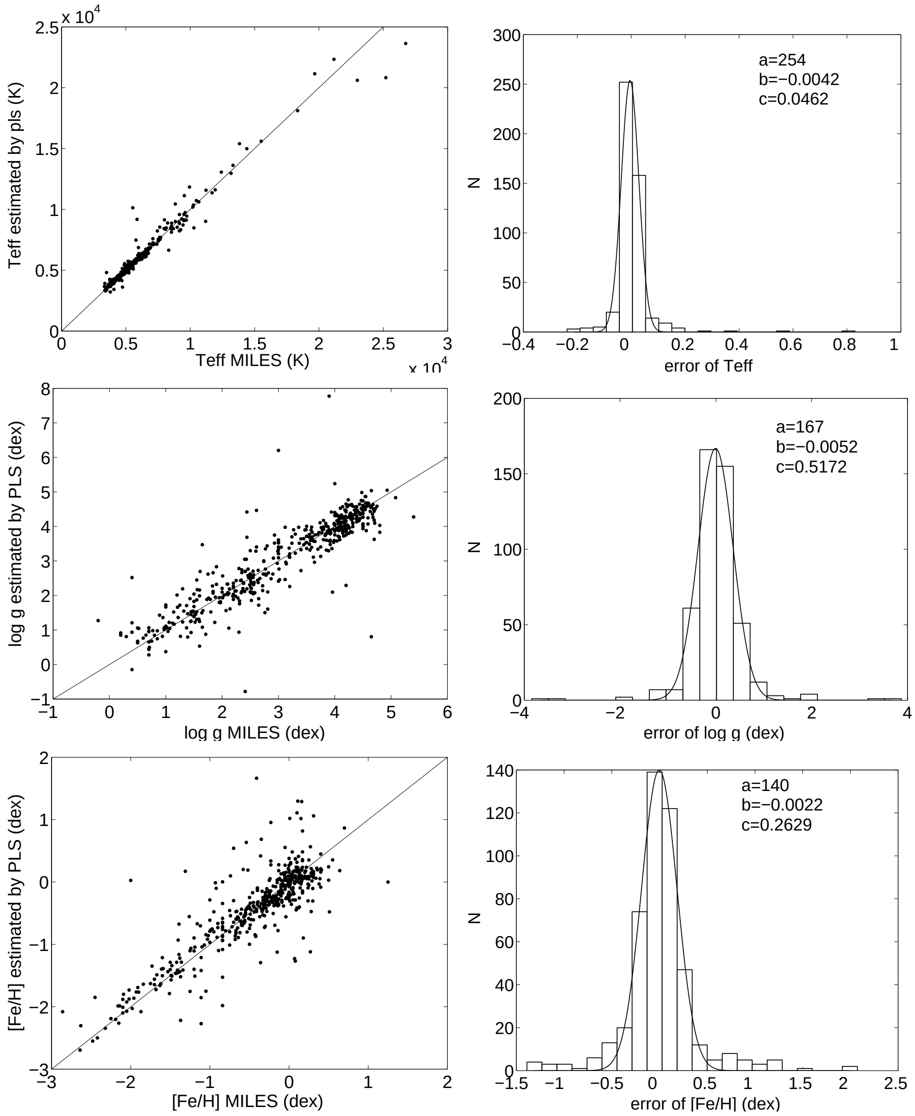

Fig. 2 Result of the piecewise PLS method for flux calibrated 0.23 nm resolution MILES spectra parameterization. The left side compares the set of test parameters estimated by the PLS method and true parameters provided by MILES where the solid line has a slope of 1. The right side shows histograms of the parameter errors and a Gaussian fit with the inset data: a, b, and c are the parameters of the Gaussian curve function: $f(x) = a * e^{-((x-b)^2/c^2)}$. Upper: $T_{\rm eff}$; Middle: $\log g$; Bottom: [Fe/H].

To check the stability of the piecewise PLS method, all the spectra were convolved with a Gaussian function to low resolution: from $\Delta\lambda = 0.23$ nm to $\Delta\lambda = 0.5$ nm and $\Delta\lambda = 1$ nm, respectively. The result of the analysis is listed in Table 1 with full SNR. To test the robustness of the PLS method, additive Gaussian noise was introduced into the set of test spectra with SNR of 100, 50, and 20, respectively. We computed the parameters of the noised test set and analyzed them. Table 1 presents the error statistic data for each test set spectra, μ is the bias (mean, error mean): $\mu = \frac{1}{n}\sum_{i=1}^{n} {\rm error}_i$; and σ is the standard deviation (std, error scatter): $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{error}_i - \mu)^2}$. Figure 3 presents the mean maximum errors (equal to $|\mu| + \sigma$) vs. SNR.

As Figure 3 and Table 1 illustrate, the accuracy and robustness of piecewise PLS tend to be affected by resolution. The parameterization of medium resolution spectra, such as $\Delta\lambda = 0.23\,\mathrm{nm}$, is almost unaffected by noise. When the resolution degrades from 0.23 nm to 0.5 nm and 1 nm, the accuracy and robustness degrade. For T_{eff} estimation, 0.5 nm resolution spectra have a mean maximum error curve close to those of the 0.23 nm resolution spectra when SNR is more than 50, while the 1 nm resolution spectra have larger mean maximum errors than those of 0.5 nm and 0.23 nm resolution spectra. For $\log g$, the mean maximum error curves of the three spectra set at different resolutions tend to blend together when SNR is higher than 100. Metallicity estimation has a similar performance with $\log g$ estimation: the medium resolution of $\Delta\lambda = 0.23\,\mathrm{nm}$ spectra are the most robust.

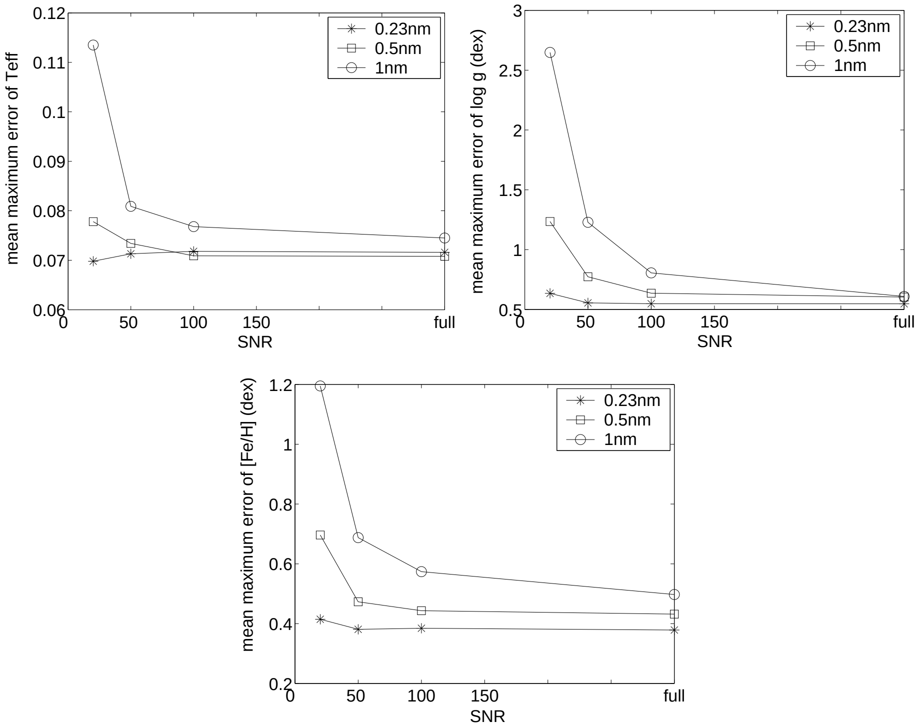

Fig. 3 Mean maximum error curves with SNR of 20, 50, 100 and full for flux calibrated spectra. We take $|\mu| + \sigma$ as the mean maximum error. The different symbols mean different resolution test data: line with circle for 1 nm resolution spectra, line with square for 0.5 nm resolution data, and the asterisk for 0.23 nm resolution data. Top left: T_{eff}; Top right: $\log g$; Bottom: metallicity.

3.2 Application to Continuum Normalized Spectra

In this section, the piecewise PLS method is tested on continuum normalized spectra with different SNR levels of noise. The pseudo-continuum of the MILES spectra is first extracted through a 7th degree polynomial fitting, then the pseudo-continuum is removed from the original spectra. This continuum normalization is preprocessed automatically by computer. Because M and late-K type stars are difficult to fit the continuum, the M and K stars were picked out from the training set and testing set. The remain-

Table 1 Error analysis for the PLS method estimating MILES spectral parameters (μ is the mean value of errors and σ is the standard deviation of errors).

		$\frac{\Delta T_{\rm eff}}{T_{\rm eff}}$		$\Delta \log g$		Δ [Fe/H]	
Resolution	SNR	μ (%)	σ (%)	μ (dex)	σ (dex)	μ (dex)	σ (dex)
0.23 nm	full	0.50	6.66	0.0006	0.5463	0.0111	0.3677
	100	0.51	6.67	0.0018	0.5457	0.0136	0.3710
	50	0.51	6.62	0.0024	0.5525	0.0111	0.3699
	20	0.40	6.58	0.0042	0.6304	0.0023	0.4125
0.50 nm	full	0.36	6.72	-0.0087	0.5947	0.0006	0.4314
	100	0.37	6.72	-0.0030	0.6329	0.0002	0.4431
	50	0.50	6.84	0.0148	0.7579	–0.0012	0.4719
	20	0.48	7.30	0.0319	1.2031	–0.0213	0.6752
1 nm	full	0.40	7.05	0.0003	0.6080	–0.0100	0.4874
	100	0.38	7.30	–0.0132	0.7921	–0.0246	0.5492
	50	0.53	7.56	0.0373	1.1911	–0.0085	0.6791
	20	1.08	10.27	0.1028	2.5454	0.0111	1.1838

Table 2 Error analysis for the PLS method of estimating the parameters of MILES spectra with continuum normalization that was preprocessed. The others are the same as Table 1.

		$\frac{\Delta T_{\rm eff}}{T_{\rm eff}}$		$\Delta \log g$		Δ [Fe/H]	
Resolution	SNR	μ (%)	σ (%)	μ (dex)	σ (dex)	μ (dex)	σ (dex)
0.23 nm	full	0.19	5.23	0.0289	0.5186	0.0225	0.3916
	100	0.29	5.27	0.0265	0.5271	0.0265	0.3874
	50	0.76	5.46	0.0574	0.5565	0.0279	0.4126
	20	2.05	6.45	0.1132	0.5999	0.0117	0.5271
0.50 nm	full	–0.26	5.50	–0.0101	0.5794	0.0041	0.4393
	100	–0.42	5.76	–0.0064	0.6171	–0.0100	0.4496
	50	–0.57	5.81	–0.0175	0.6893	–0.0119	0.4993
	20	–0.61	6.90	–0.0427	1.1744	–0.0044	0.6337
1 nm	full	0.19	5.78	–0.0192	0.6351	0.0078	0.4551
	100	0.21	5.95	–0.0256	0.8138	–0.0015	0.5179
	50	–0.16	7.03	–0.0290	1.1457	0.0551	0.6551
	20	–0.06	10.32	-0.0935	2.4937	–0.0559	1.1359

ing 273 spectra in the test set and the 285 spectra in the training set were smoothed to low resolution of 0.5 nm and 1 nm by Gaussian convolution. The results are presented in Figure 4, and the error analysis is listed in Table 2. To test the robustness of the method for continuum normalized spectra, Gaussian noise of SNR=100, SNR=50, and SNR=20 were added to the test spectra. The noised spectra were parameterized with the piecewise PLS models, and the results are listed in Table 2. Figure 5 shows the mean maximum error curves as a function of SNR. From Figure 4, we conclude that the regression models of continuum normalized spectra derived by piecewise PLS are feasible for the parameterization with small bias and error scatter. Resolution still influences the accuracy and robustness: medium resolution spectra of 0.23 nm have better performance than low resolution spectra of 0.5 nm and 1 nm. Owing to the normalization of the continuum, the accuracy of $T_{\rm eff}$ from medium resolution spectra is not as robust as that of flux calibrated spectra. Metallicity and $\log g$ estimation have a similar performance. We found that medium resolution spectra were more robust than the low resolution 0.5 nm and 1 nm spectra.

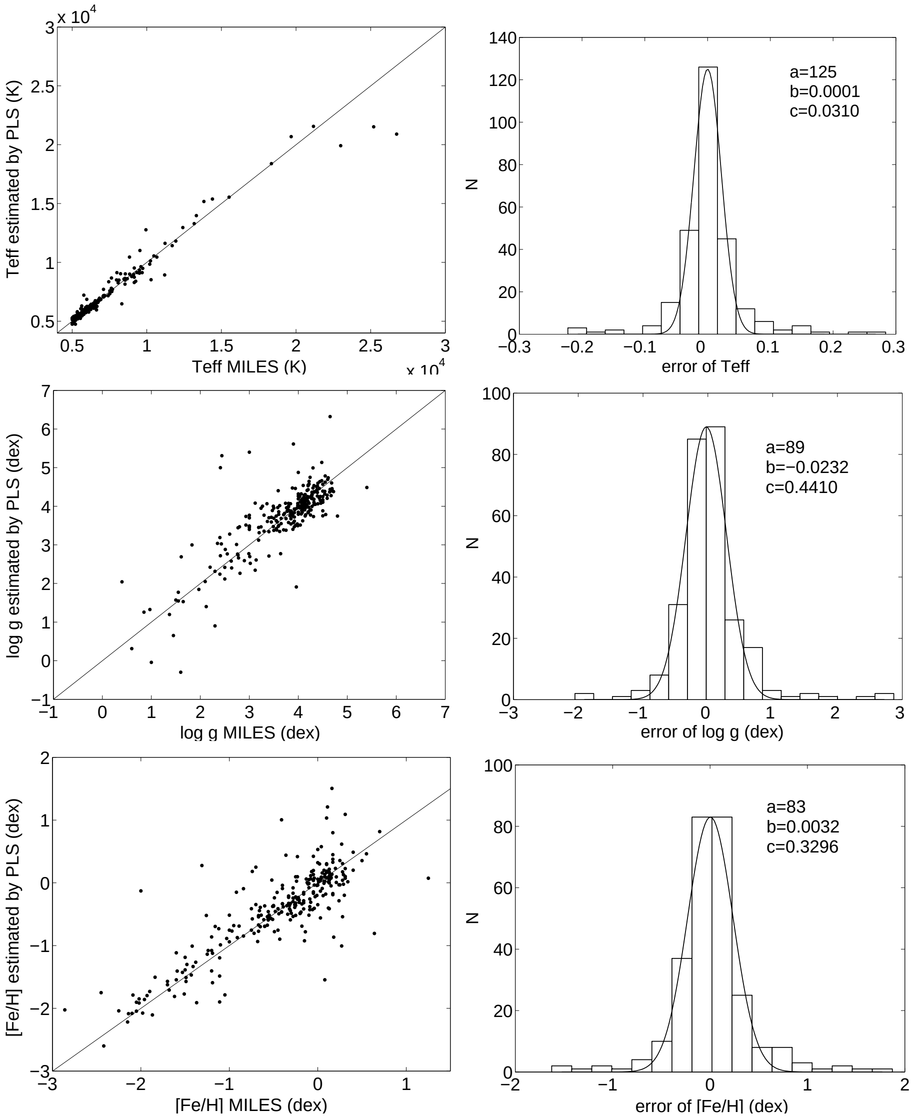

Fig. 4 Result of the PLS regression method for continuum normalized 0.23 nm resolution MILES spectral parameterization. The left side compares the test set parameters estimated by the PLS method and the true parameters provided by MILES where the solid line has a slope of 1. The right side shows histograms of the parameter errors and a Gaussian fit with the inset data the same as Fig. 2. Upper: $T_{\rm eff}$; Middle: $\log g$; Bottom: [Fe/H].

3.3 Test of ELODIE Stellar Library Spectra

In this section, spectra from the ELODIE release 3 (Moultaka et al. 2004; Prugniel 2001) library were selected for checking the piecewise PLS regression models which had been trained by the MILES spectra. In galaxy survey projects, researchers should construct stellar templates to measure parameters of stars. The best templates are produced from the project itself, but usually before the survey starts and many spectra with parameters have been calibrated, the templates have to be constructed by other stellar spec-

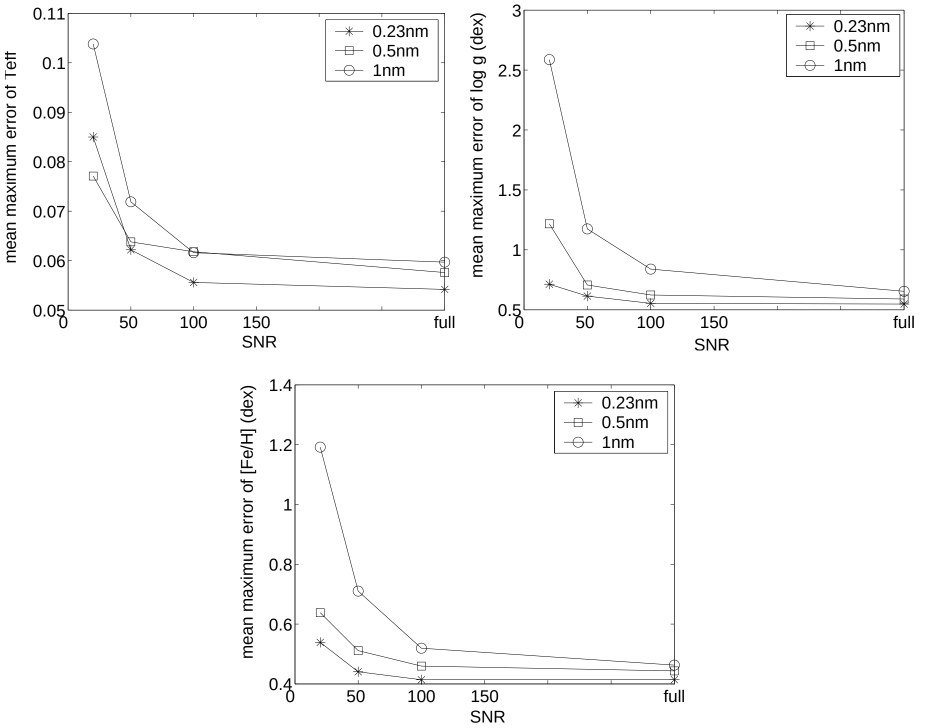

Fig. 5 Mean maximum error curves with SNR of 20, 50, 100 and full for flux calibrated spectra. We take $|\mu| + \sigma$ as the mean maximum error. The different symbols mean different resolution test data: line with circle for 1 nm resolution spectra, line with square for 0.5 nm resolution data, and the asterisk for 0.23 nm resolution data. Top left: $T_{\rm eff}$; Top right: log g; Bottom: metallicity.

Table 3 Error Analysis for the PLS Method of Estimating ELODIE Spectral Parameters

	$\frac{\Delta T_{\rm eff}}{T_{\rm eff}}$		$\Delta \log g$		Δ [Fe/H]	
Spectra	μ(%)	σ(%)	μ (dex)	σ (dex)	μ (dex)	σ (dex)
Flux calibrated	–3.81	5.16	0.1356	0.4416	–0.0928	0.3813
Continuum normalized	–3.78	5.20	0.1167	0.4864	–0.1285	0.4147

tra libraries, which are the observed spectra or synthesized spectra of a stellar atmospheric model, such as the Kurucz model, the MARCS model etc. This will influence the results of the measurements. To determine the system offset, we checked piecewise PLS models with spectra coming from the ELODIE library, where the models were trained by MILES spectra. The ELODIE library contains 1969 spectra of some 1390 stars with a resolving power of $R = 10\,000$. Each stellar spectrum provides the atmospheric parameters with a quality flag: quality flag for $T_{\rm eff}$ is on a $-1 \sim 4$ scale, $\log g$ is on a $-1 \sim 1$ scale, and [Fe/H] is on a $-1 \sim 4$ scale. We used 700 spectra with $5000 < T_{\rm eff} < 15\,000$ K and each parameter had a quality flag of at least 1. The spectra were smoothed to the resolution of 0.23 nm by convolution. We checked the PLS method with two kinds of spectra: flux calibrated and continuum normalized spectra. The error analysis is listed in Table 3. Figures 6 and 7 present the comparisons between the parameters from ELODIE and those estimated by the piecewise PLS method.

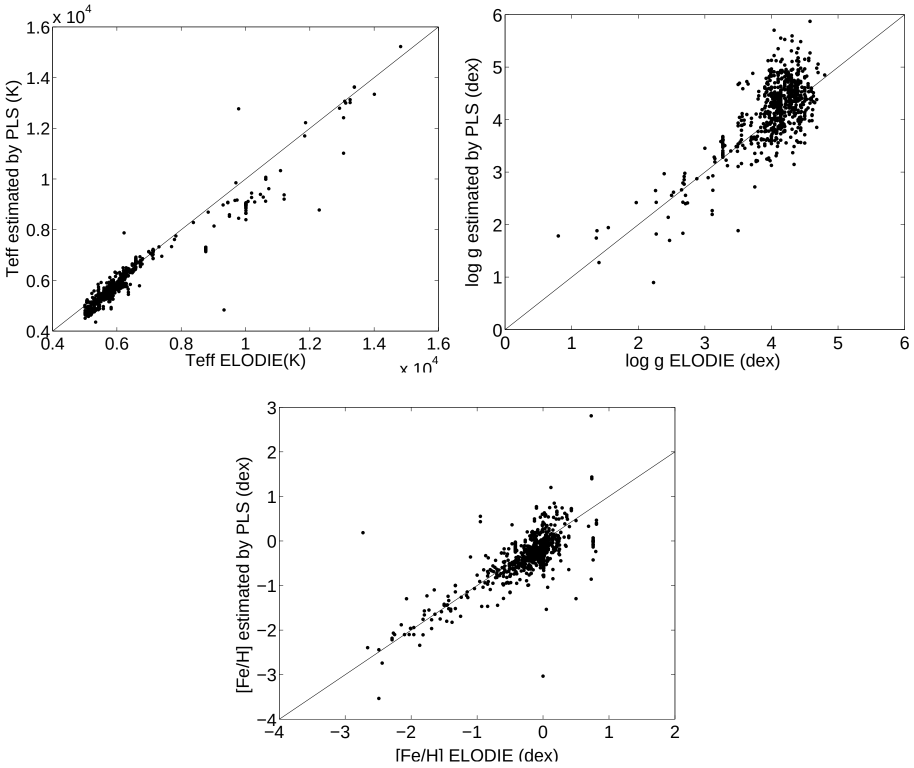

Fig. 6 Comparison between the parameters of ELODIE and the parameters estimated by the piecewise PLS method for flux calibrated spectra, where the solid line has a slope of 1. Top left: $T_{\rm eff}$; Top right: log g; Bottom: metallicity.

4 DISCUSSION AND CONCLUSIONS

Before evaluating the performance of the piecewise PLS method, we should stress that the accuracy of the results in the above experiments are limited by the nonuniform distribution of parameter coverage of training spectra. A majority of MILES spectra are F, G, K and M type stars, and A, O and B type stars are in the minority. The surface gravity mainly covers the range between 3.5 dex and 4.5 dex. The metallicities cluster in the range between –0.6 dex and 0.5 dex. The nonuniformity of training spectra lowered the regression ability of the PLS models for the spectra that were outside of the above dense parameter domain. With more spectra across all parameter ranges introduced into the training set in the future, better results could be achieved. Nevertheless, the PLS models still derived comparatively accurate results for the spectra of high $T_{\rm eff}$, or log g in the range of 0 – 3.5 dex, or [Fe/H] under – 0.6 dex, especially for metal-poor stars, as demonstrated in Figures 2, 4, 6 and 7.

PLS method is different from the present parameterization techniques of fundamental stellar parameter estimation. ANN method and MDM are interpolations of the training data. As Bailer-Jones (2001) pointed out, the ANN is an interpolation of the training data, and the more coarsely the parameter grid is sampled, the harder it is for the network to get a reliable interpolation. Compared with the methods above, the PLS method explores a new way for automated stellar spectral parameter estimation. The advantages of PLS method are:

- The precision of PLS regression does not rely as strongly on the tightness of parameter grid of training data as ANNs and MDM. PLS could derive a linear mapping function from spectrum to

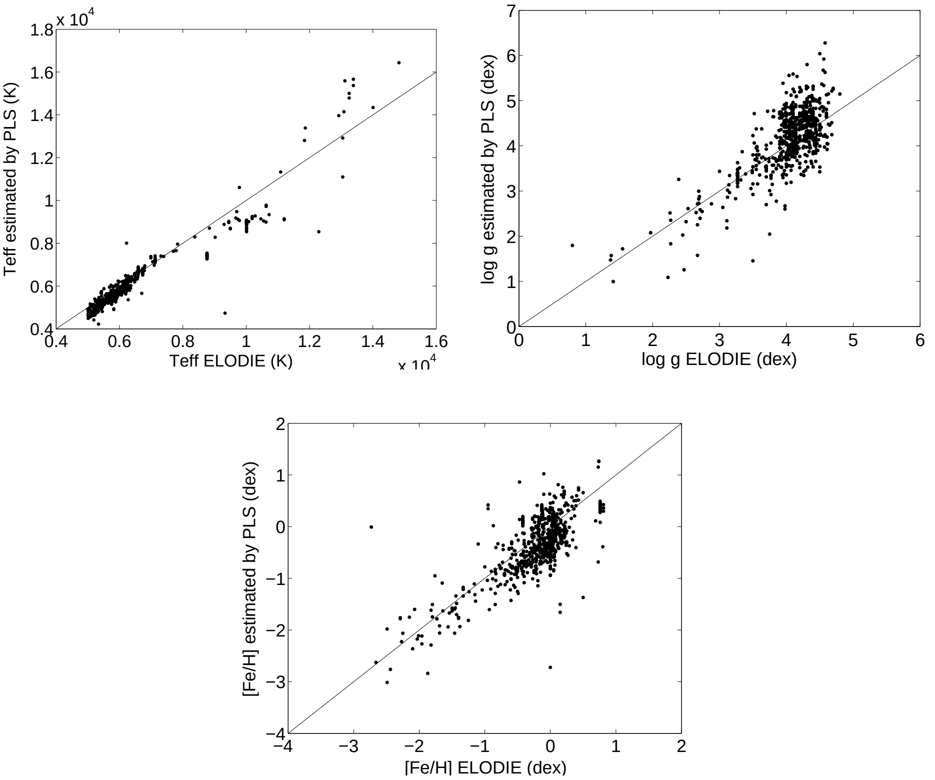

Fig. 7 Comparison between the parameters of ELODIE and the parameters estimated by the piecewise PLS method for continuum normalized spectra. Top left: $T_{\rm eff}$; Top right: $\log g$; Bottom: metallicity.

parameters given enough training data. We believe that this aspect makes PLS more efficient, while the grid of training data for MDM and ANNs may be coarse.

- Once the regression formula is set up, the estimation calculation is extremely rapid. The training procedure is also rapid. For example, the experiment on medium resolution spectra in Section 3.1 costs less than 2 min to set up eight PLS regression models for piecewise PLS on a Pentium D personal computer with a 3.4 GHz CPU clock frequency.
- It is very easy to use. As a mature statistical algorithm, many researchers provide PLS algorithm modules, and many mathematical software have inline PLS algorithm tools, such as SAS and Matlab.
- Piecewise PLS firstly estimates the temperature, then according to the spectral type, metallicity and gravity are estimated by one of seven PLS models. The results of experiments show that piecewise PLS is robust and efficient for medium resolution spectral parameter estimation.

We can conclude that the piecewise PLS method can be adapted to the automated parameterization of large surveys with huge amounts of low resolution spectra.

Acknowledgements We wish to thank Dr. James Wicker for detailed amendments of language, and the referee for insightful and critical comments, whose work improved the presentation of this paper. This work was funded by the National Natural Science Foundation of China (NSFC) under Nos. 10603001 and 10778626.

Appendix A: PLS REGRESSION PRINCIPLE AND ALGORITHM

Notion and Notation:

The n observations described by p explanatory variables are stored in an $n \times p$ matrix denoted by X, and the values of m dependent variables collected on these n observations are collected in the $n \times m$ matrix.

The Principle of PLS Regression

PLS regression is a multivariate data analysis technique which can be applied to relate dependent variables (y) to explanatory variables (x). The method aims to identify the underlying factors, that is, the linear combinations of the x-variables, that best fit and model the y dependent variable.

The PLS regression procedure can be viewed as a stepwise procedure, where at each step a score vector t is extracted from X and a score u is extracted from Y. For the first step:

$$\begin{aligned} t_1 &= Xw &&= w_1x_1 + \cdots + w_kx_k, \\ u_1 &= Yc &&= c_1y_1 + \cdots + c_my_m, \end{aligned}$$

where the unknown parameters are the weights $w = (w_1, \cdots, w_k)$ and $c = (c_1, \cdots, c_m)$, subject to $|w| = |c| = 1$. The optimization task of PLS regression is:

$$\begin{aligned} \mathrm{var}(t_1) &\to \max, \\ \mathrm{var}(u_1) &\to \max, \\ r(t_1, u_1) &\to \max\,. \end{aligned}$$

The regular mathematical expression is:

$$\max_{w_1c_1}\langle Xw_1, Yc_1\rangle \quad \text{s.t.} \quad \begin{cases} w_1^T w_1 &= 1, \\ c_1^T c_1 &= 1. \end{cases}$$

Using the Lagrange multiplier technique, one can show that w is the first eigenvector of the matrix X^TYY^TX, and c is the first eigenvector of the matrix Y^TXX^TY. Getting the latent vector t_1 and u_1, X and Y are modeled by the same latent vectors:

$$\begin{aligned} X &= t_1p_1^T + X_1, \\ Y &= u_1q_1^T + Y_1^*, \\ Y &= t_1r_1^T + Y_1, \end{aligned}$$

where the regression coefficient vectors are:

$$p_1 = \frac{X^Tt_1}{\|t_1\|^2}, \qquad q_1 = \frac{Y^Tu_1}{\|u_1\|^2}, \qquad r_1 = \frac{Y^Tt_1}{\|t_1\|^2}.$$

X_1, Y_1^*,Y_1 are the three residual matrices when X, Y are explained by the latent vectors.

For the second step, X and Y are replaced by the residual matrices X_1 and Y_1, respectively, then the second weighting vectors w_2, c_2 are computed in the same way and the corresponding latent vectors t_2, u_2 are extracted from X_1 and Y_1. Consequently, X_1 and Y_1 are deflated by the same latent vectors:

$$\begin{aligned} X_1 &= t_2p_2^T + X_2, \\ Y_1 &= t_2r_2^T + Y_2. \end{aligned}$$

Step by step in the same way, if the rank of matrix X is A, X and Y can be expressed by a series of latent vectors:

$$\begin{aligned} X &= t_1p_1^T + \cdots + t_Ap_A^T, \\ Y &= t_1r_1^T + \cdots + t_Ar_A^T + Y_A. \end{aligned}$$

Algorithm of PLS

There are several algorithms for calculating the PLS model. The NIPALS algorithm of Wold et al. (Wold 2001; Philippe 2005) is shown below.

The X and Y are first processed with their columns centered.

1. Getting a starting vector of u, usually one of the Y columns.
2. The X-weights, w: $w = X^T u/u^T u$, set norm w to $\|w\| = 1$.
3. Calculate X-scores, t: $t = Xw$, set norm t to $\|t\| = 1$.
4. The Y-weights, c: $c = Y^T t/t^T t$, set norm c to $\|c\| = 1$.
5. Update Y-scores, u: $u = Yc/c^T c$.
6. If t is not converged, i.e., $\|t_{\rm old} - t_{\rm new}\|/\|t\| > \varepsilon$, where ε is a small threshold, then go to step 2. If t converged, compute the value of b which is used to predict Y from t as $b = t^T u$, and compute the loading vector p for X. Then deflate the effect of t from both X and Y:
$$p = X^T t,$$
$$X = X - tp^T.$$
$$Y = Y - btc^T.$$
The vectors t, u, w, c, and p are then stored in the corresponding matrices, and the scalar b is stored as a diagonal element of B.
7. If X is a null matrix, then the whole set of latent vectors has been found, otherwise the procedure can be re-iterated from step 1.

The dependent variables Y are predicted using the multivariate regression formula as:

$$Y^* = TBC^T = XB_{\rm PLS},$$

where $B_{\rm PLS} = P^{T+}BC^T$ with P^{T+} being the Moore-Penrose pseudo-inverse of P^T.

When only a subset of latent vectors is used, the prediction of Y is optimal for this number of predictors. The obvious question is to find the number of latent vectors needed to obtain the best generalization for the prediction of new observations. This is, in general, achieved by cross-validation techniques.

References

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

Bailer-Jones, C. A. L. 2001, in Automated Data Analysis in Astronomy, ed. R. Gupta et al. (New Delphi, India: Narosa Publishing House), 83

Bonifacio, P., & Caffau, E. 2003, A&A, 399, 1183

Fuentes, O., & Gulati, R. K. 2001, in The Seventh Texas-Mexico Conference on Astrophysics: Flows, Blows and Glows, Revista Mexicana de Astronomía y Astrofísica, eds. H. L. William, & T.-P. Silvia (Serie de Conferencias) 10, 209

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

Moultaka, J., Ilovaisky, S. A., Prugniel, P., & Soubiran, C. 2004, PASP, 116, 693

Philippe, B., Vincenzo, E. V., & Michel, T. 2005, Computatinal Statistics & Data Analysis, 48, 17

Prieto, C. A. 2003, MNRAS, 339, 1111

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

Recio-Blanco, A., Bijaoui, A., & de Laverny, P. 2006, MNRAS, 370, 141

Sánchez-Blázquez, P., Peletier, R. F., Jiménez-Vicente, J., et al. 2006, MNRAS, 371, 703

Snider, S., Prieto, C. A., Hippel,T. V., Beers, T. C., Sneden, C., Qu, Y., & Rossi, S. 2001, ApJ, 562, 528

Soubiran, C., Katz, D., & Cayrel, R. 1998, A&AS, 133, 221

Wang, H. W. 1999, Partial Least-squares Regression-method and Applications (Beijing: National Defence Industry Press of China)

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

Wold, S., Trygg, J., Berglund, A., & Antti, H. 2001, Chemometrics and Intelligent Laboratory Systems, 58, 131

Wold, S., Sjostrom, M., & Eriksson, L. 2001, Chemometrics and Intelligent Laboratory Systems, 58, 109