Research in Astronomy and Astrophysics manuscript no.
(LaTeX: sample.tex; printed on September 5, 2013; 22:43)

On the Construction of New Stellar Classification Templates Library for LAMOST Spectra Analysis Pipeline

Wei Peng[1,2], Luo Ali*[1,3], Wang Fengfei[1], Li Yinbi[1], Zhang Jiannan[1], Pan Jingchang[3], Tu Liangping[1,4], Jiang Bin[3], Zhao Yongheng[1], Chen Jianjun[1,2], Chen Xiaoyan[1], Du Bing[1], Hou Wen[1,2], Kong Xiao[1,2], Liu Jie[3], Song Yihan[1], Wu Yue[1] and Yi Zhenping[1,3]

[1] Key Laboratory of Optical Astronomy,National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100012, China *lal@lamost.org*

[2] University of Chinese Academy of Sciences, Beijing, 100049, China

[3] School of Mechanical, Electrical and Information Engineering, Shandong University,Weihai, 264209, China

[4] School of Science, Liaoning University of Science and Technology, Anshan, 144051, China

Abstract The aim of LAMOST Spectra Analysis Pipeline (also called 1D Pipeline) is to classify and measure the spectra of objects observed in the survey. Stars are classified into different sub-classes by matching to spectra templates. To make the classification better and reliable, a new LAMOST stellar spectral classification templates library is constructed. This paper presents the construction of the templates through the Principal Component Analysis(PCA) of about one million LAMOST stellar spectra. These spectra were selected from all released 1,946,429 stellar spectra in LAMOST Data Release One (DR1). Convolved with the SDSS *ugriz* filter curve, each spectrum is assigned a pseudo g-r color (hereafter g*-r*). All spectra are divided into different groups by the g*-r*. Additional groups are formed by different subclasses classified with the current pipeline. In addition, some special types of psectra are manually picked out to add into the groups. In each group, we exclude some outliers and then apply PCA method to the remaining spectra. All spectra are reconstructed using the first few principal components and the weighted average spectra are constructed as the template spectra in the groups. Each template spectrum is labeled a MK class by comparing with some label-known templates and the spectra of MK standard stars . All the templates are visually inspected and some low-quality spectra are excluded. Additionally, some unlabeled or wrongly labeled spectra are relabeled or abandoned. The finally left spectra compose the template library. The first version of the library contains 164 spectra and 65 different MK classes, which is available to public on the web site.

Key words: methods: data analysis, methods: statistical, surveys

1 INTRODUCTION

The LAMOST survey(Cui et al. 2012; Zhao et al. 2012) contains two main parts: the LAMOST ExtraGAlactic Survey (LEGAS) and the LAMOST Experiment for Galactic Understanding and Exploration survey of Milky Way stellar structure(LEGUE, see Deng et al. (2012)). The unique design of LAMOST enables it to take 4000 spectra in a single exposure to a limiting magnitude as faint as r=19 at the resolution R=1800, which is equivalent to the design goal of r=20 for the resolution R=500. The LAMOST therefore has great potential to efficiently survey a large volume of space for stars and galaxies.

The LAMOST spectral analysis pipeline(Luo & Zhao 2001; Luo et al. 2004, 2008; Wang et al. 2010; Luo et al. 2012) is based on the specBS pipeline(Aihara et al. 2011) for the analysis of SDSS spectra. This analysis code carries out χ^2 fits of the spectra to templates in wavelength space (in the spirit of Glazebrook et al. (1998)), fitting spectra with linear combinations of eigen-spectra and low-order polynomials.

Currently, a set of SDSS spectra with carefully chosen zero points of a vagriety of spectral types, which also contain galaxies, QSOs and stars (categorized in terms of subtypes), are used to construct the templates. With more spectra with high quality observed, it is necessary and better to construct a new template library based on the spectra observed by LAMOST.

In this paper, we described in detail the construction of the LAMOST stellar classification template library. The paper is processed as follows: Section 2 describes the used spectra from LAMOST Data Release One (DR1). The detailed description of the template library construction steps, the related methods and the group dividing are outlined in section 3. The results and discussions are given in 4. A brief summary is given in section 5.

2 THE SPECTRA FROM LAMOST DATA RELEASE ONE(DR1)

The first data release of LAMOST survey contains the spectra in the pilot survey and the first year of general survey. The pilot survey of LAMOST was launched on 2011 Oct 24, and ended in June 2012.The first year of LAMOST general survey began from September 28th 2012 and ended on June 3rd 2013.

The survey has obtained spectra of stars in the Milky Way, which includes fainter objects on dark nights (Yang et al. 2012; Carlin et al. 2012), brighter objects on bright night (Zhang et al. 2012), objects in the disk of the Galaxy with low latitude (Chen et al. 2012) and objects in the region of the Galactic Anti-Center(Liu et al. 2013). The DR1 totally contains 2,204,860 spectra, including 717,660 spectra of pilot survey and 1,487,200 spectra of regular survey. In addition, the atmospheric parameters of 1,085,404 stars are calculated, which becomes the largest stellar spectral parameters catalog in the world at present.

The spectral resolution R is about 1800 around g band with a 2/3 silt width(Wang et al. 2013). The wavelength coverage is from 3700 Å to 9100 Å. Two arms of each spectrograph covers the wavelength range and overlaps in 200Å. The raw data have been reduced with LAMOST 2D pipeline(Bai 2012) including bias subtraction, cosmic-ray removal, spectral trace and extraction, flat-fielding, wavelength calibration sky subtraction, and combination. For the spectra with SNR$>$5, the 1D pipeline(Wang et al. 2010) gives spectral type and redshift. For LAMOST spectra with low confidence in measurement or low SNR, human checking is also applied to data quality(Luo et al. 2013).

We exclude those spectra in the Galactic Anti Center and M31 to avoid the high effect of interstellar dust extinction on the spectra and 742,669 spectra are left. Some other spectra are excluded in the construction of the library, which will be discussed in the section 3.

3 THE CONSTRUCTION STEPS

3.1 The pseudo g-r color

LAMOST is a spectroscopic oriented survey telescope and doesn't have its own photometric data. The photometric data of objects are from different surveys. Meanwhile, the flux calibration is relative not absolute(Song et al. 2012). Consequently, we can not get accurate and uniform colors for LAMOST spectra. Here, we propose a pseudo g-r color(hereafter g*-r*) obtained by convolving each observed spectra with the SDSS $ugriz$ filter response curves . We describe the calculation in detail as follows:

1. Suppose that the sampling points of SDSS g & r filter response curves are P_g, P_r respectively and the response curve values are C_g, C_r respectively .
2. Interpolate the flux of the observed spectra in the points of P_g and P_r to get F_g and F_r respectively.
3. Get the pseudo color g^*-r^*:

$$g^* - r^* = -2.5 * log\frac{\sum F_g * C_g}{\sum C_g} + 2.5 * log\frac{\sum F_r * C_r}{\sum C_r} \quad (1)$$

We select these objects with SDSS $ugriz$ filter magnitude(from different surveys) and $SNR > 20$. The diagram of g-r color and $g^* - r^*$ of these objects is shown in Fig.1. There is a obvious linear relationship between these two colors. We derive the best-fit expression as:

$$g - r = 0.807 * (g^* - r^*) + 0.655 \quad (2)$$

Among these spectra, we also choose these spectra classified as O, B, A, F, G, K, M-type. For these spectra, the average value and standard deviation of $g^* - r^*$ in each class are also calculated. As shown in Figure 2, the $g^* - r^*$ color varies obviously in each class.

The g-r color is a very good indicator of effective temperature. For SDSS spectra, Željko Ivezić et al. (2008) derived a relation between effective temperature and the color g-r in the $-0.3 < g - r < 1.3$ color range:

$$Log_{10}(T_{eff}/K) = 0.0283 * (g - r)^3 + 0.0488 * (g - r)^2 - 0.316 * (g - r) + 3.882 \quad (3)$$

We can then derive a expression between effective temperature T_{eff} and the color g*-r* from the formula 3 and the formula 2 as:

$$Log_{10}(T_{eff}/K) = 0.0283 * (g^* - r^*)^3 + 0.0318 * (g^* - r^*)^2 - 0.203 * (g^* - r^*) + 3.696 \quad (4)$$

For some spectra classified as A, F, G, K-type, the effective temperatures(T_{eff}), surface gravities(Log g) and metallicities([Fe/H]) determined by the LASP(LAmost Stellar Parameter pipeline, see Wu et al. (2011)) are provided. The relationship between $g^* - r^*$ color and T_{eff} is shown in Figure 3. We also derive a best fit 3-order polynomial expression as:

$$Log_{10}(T_{eff}/K) = 0.0432 * (g^* - r^*)^3 + 0.0107 * (g^* - r^*)^2 - 0.165 * (g^* - r^*) + 3.746 \quad (5)$$

As shown in Fig.3, these two formulas 4 and 5 nearly coincide with each other in the Teff range [5500,7000].

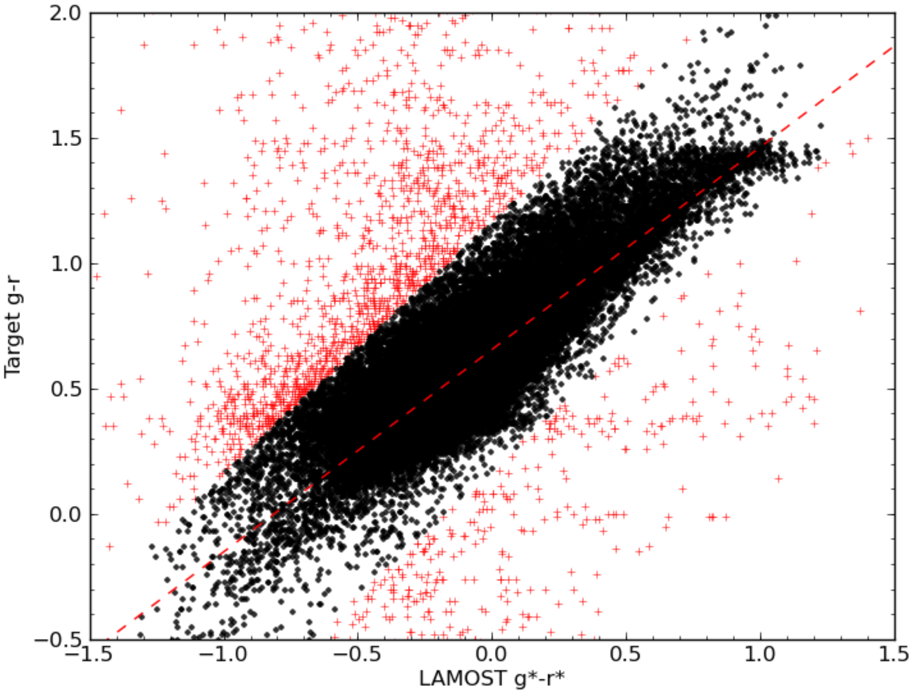

Fig. 1 The diagram of g-r color and $g^* - r^*$. The X-axis is our proposed $g^* - r^*$ and the Y-axis is the g-r color obtained from target catalogue. The red line is our derived best-fit expression as formula 3. And the points in red are excluded outliers while deriving the expression.

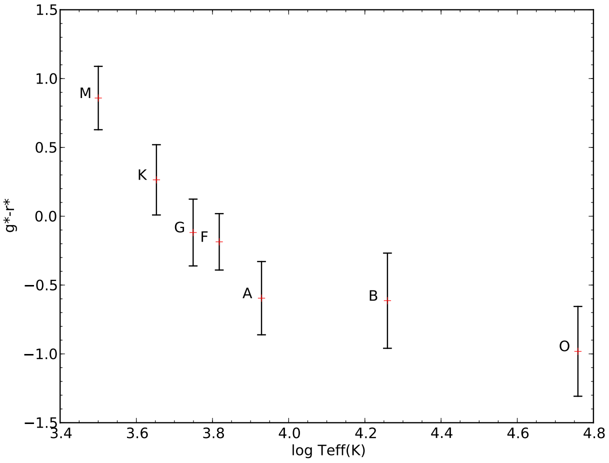

Fig. 2 The average value and standard deviation of $g^* - r^*$ for each class. For each class, the X-value are the median effective temperature in theory. Meanwhile, the center of each error-bar is the average $g^* - r^*$ color of spectra classified as the corresponding class and the half length is the standard deviation.

3.2 Group Dividing

We select these spectra with $g^* - r^*$ in the range[-1.5,2.0] and divide all spectra into 175 groups with 0.02 mag width interval. The number distribution is shown in Figure 4. We mark these these groups with group-id from 1 to 175.

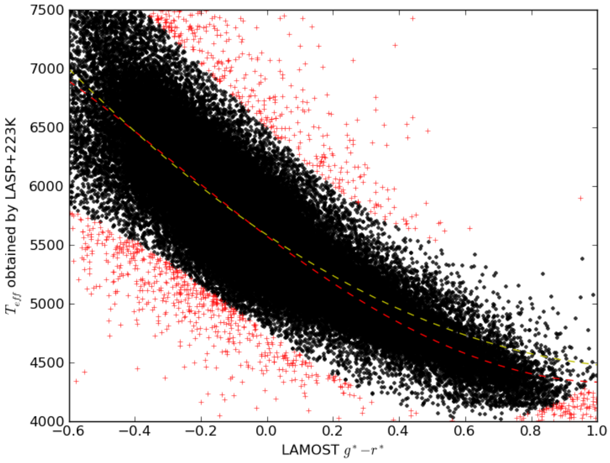

Fig. 3 The relationship between $g^* - r^*$ color and T_{eff}. The T_{eff} is added by 223K to decrease the system inconsistency between SSPP and LASP(Wu et al. 2011). The yellow line is the expression as formula 4. The red line is our derived best-fit expression as formula 5. And the points in red are excluded outliers while deriving the expression.

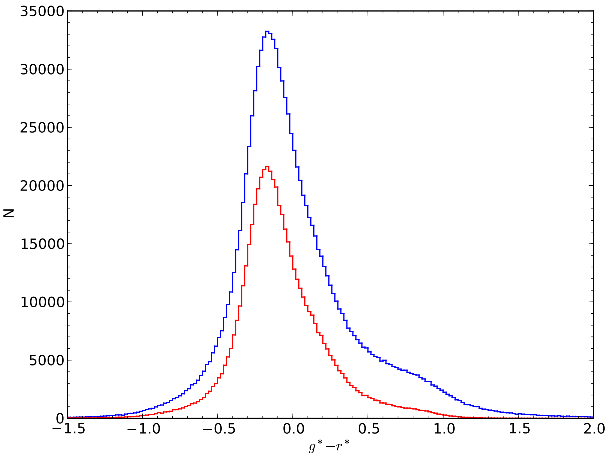

Fig. 4 The number distribution of spectra in each $g^* - r^*$ bin. The blue line is all spectra while the red line is the spectra with $SNR > 10$.

After the automated processing of LAMOST Spectra Analysis Pipeline and visual inspection, there are 76 different stellar subclasses in our selected spectra. The distribution is given is Table 1. Some A-type spectra in MILES library(Falcón-Barroso et al. 2011) are picked out to add into the templates. Consequently, there are more A-type subclasses. We mark these these groups with group-id from 176 to 251.

Yi et al. (2013) present a spectroscopic catalog of 67,082 M dwarfs from LAMOST Pilot Survey. All spectra are divided into 10 subclasses from M0 to M9. We have divided these spectra into groups with group id as shown in Table 1. Zhao et al. (2013) present a spectroscopically identified catalog of 70 DA white dwarfs (WDs). Meanwhile, Zhang et al. (2013) identified 230 other DA white dwarfs. We combine these two catalogs and add the spectra into group 250. Jiang et al. (2013) report the identification of 10 cataclysmic variables. We add these 10 spectra into group 207.

Table 1 The number distribution of different subclasses

Group ID	Subclass	Amount	Group ID	Subclass	Amount	Group ID	Subclass	Amount
176	A0	67	202	B9	443	227	L	2
177	A0I	27	203	Binary	170	228	L0	1
178	A0III	407	204	Carbon	168	229	L1	1
179	A0p	27	205	CarbonWD	6	230	L2	1
180	A1IV	653	206	Carbon_lines	10	231	L5	4
181	A1V	527	207	CV	17	232	L5.5	8
182	A2I	10	208	EM	42	233	L9	5
183	A2IV	1692	209	Emission	21	234	M0	19139
184	A2V	5761	210	F0	27808	235	M0V	13
185	A3I	37	211	F2	44192	236	M1	19953
186	A3IV	2084	212	F5	119328	237	M2	17231
187	A3V	2080	213	F9	292830	238	M2V	12
188	A4III	926	214	G0	47697	239	M3	9749
189	A4V	773	215	G2	92229	240	M4	3860
190	A5	26	216	G4	1	241	M5	855
191	A5I	213	217	G5	81202	242	M6	412
192	A5V	1253	218	G7	3650	243	M7	259
193	A6IV	1240	219	K	1	244	M8	47
194	A6V	322	220	K0	1998	245	M9	66
195	A7III	4033	221	K1	85218	246	Non	2
196	A7V	647	222	K3	77164	247	O	79
197	A9	4	223	K5	73045	248	OB	16
198	A9V	2170	224	K6	1	249	T2	11
199	B	15	225	K7	45839	250	WD	535
200	B0	1	226	K9	4	251	WDmagnetic	14
201	B6	492						

Consequently, there are 251 groups in total. The group with less number are not neglected to avoid excluding some relatively rare types of stars.

3.3 Used Methods

3.3.1 LOcal Outlier Probabilities(LoOP)

Kriegel et al. (2009) proposed a local outlier factor(LOF, see Breunig et al. (2000)) based outlier detection method. LoOP is a local density based method that uses statistical concepts to output the final score. The LoOP score represents the probability that a particular point is a local density outlier. The LoOP is calculated as follows(Kriegel et al. 2009):

1. (k-$distance$ of an object p) For any positive integer k, the k-distance of object p, denoted as k-$distance(p)$, is defined as the distance $d(p, o)$ between p and an object $o \in D$ such that: (i) For at least k objects $o' \in D \setminus p$, it holds that $d(p, o') \leq d(p, o)$. (ii)For at most k-1 objects $o' \in D \setminus p$, it holds that $d(p, o')<d(p, o)$.
2. (k-$distance$ neighborhood of an object p) Given the k-$distance$ of p, the k-$distance$ neighborhood of p contains every object whose distance from p is not greater than the k-$distance$, i.e. $N_{k-distance(p)}(p) =$

$\{q \in D \setminus p \mid d(p,q) \leq k\text{-}distance(p)\}$. These objects q are called the k-$nearest$ neighbors of p. Simplify the notation to use $N_k(p)$ as a shorthand for $N_{k-distance}(p)$.

3. The standard distance. $\sigma(p, N_k(p))$ is defined as the standard deviation of the distance around p:

$$\sigma(p, N_k(p)) = \sqrt{\frac{\sum_{s \in N_k(p)} d(p,s)^2}{|S|}} \tag{6}$$

4. The probabilistic set distance. $pdist(\lambda, p, N_k p)$ is defined as follows:

$$pdist(\lambda, p, N_k p) = \lambda * \sigma(p, N_k(p)) \tag{7}$$

5. The Probabilistic Local Outlier Factor PLOF. $PLOF_{\lambda,S(p)}(p)$ represents the ratio of the density estimation:

$$PLOF_{\lambda,S(p)}(p) = \frac{pdist(\lambda, p, N_k(p)) * |N_k(p)|}{\sum_{s \in N_k(p)} pdist(\lambda, s, N_k(s))} - 1 \tag{8}$$

6. The aggregate Probabilistic Local Outlier Factor nPLOF. This is the scaling factor that makes the score independent from any distribution:

$$nPLOF = \lambda * \sqrt{\frac{\sum_{p \in D} PLOF_{\lambda,S(p)}(p)^2}{|D|}} \tag{9}$$

7. The Local Outlier Probability:

$$LoOP_{N_k p}(p) = max(0, erf\frac{PLOF_{\lambda,S}(p)}{nPLOF * \sqrt{2}}) \tag{10}$$

Although the spectra are divided into different groups by g*-r* or by the classification, there are also some outliers in each group. We use the LoOP method to exclude these outliers in each group. The distance measurement function used in our work is the cosine distance $d = 1 - \frac{A*B}{|A|*|B|}$

3.3.2 *Principal Component Analysis(PCA)*

The Principal Component Analysis (PCA) (Jolliffe 2002) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. The detailed description of the PCA method is discussed in Whitney (1983); Jolliffe (2002).

As a viable tool, Principal Component Analysis(PCA) has been applied in the classification of spectra(Whitney 1983; Bailer-Jones et al. 1998; Yip et al. 2004; Almeida & Prieto 2013). In addition, Tu et al. (2009, 2010); Jiang et al. (2013); Wei et al. (2013) have used PCA to do the data dimension reduction in finding relatively rare objects. McGurk et al. (2010) applied PCA to 200,000 stellar spectra obtained by the Sloan Digital Sky Survey (SDSS). They discussed correlations of eigen-coefficients with metallicity and gravity estimated by the Sloan Extension for Galactic Understanding and Exploration Stellar Parameters Pipeline(SSPP)(Lee et al. 2008).

3.4 The Construction Steps

Combing the above methods and group dividing method, the construction steps is carried out as follows:

1. For the groups with more than 5,000 spectra, only first 5,000 spectra with the largest SNR are selected.
2. De-redshift the spectra and unify the wavelength to 3800Å-9000Å with fixed step 1Å(the amount of all sampling points is N=5201) and then get the unified flux F for each spectrum.
3. Exclude these spectra existing $F \leq 0$ and normalize the remaining spectra F with:

$$F_i = \frac{F_i}{\sqrt{\sum\limits_{j=1}^{N} F_i^2}} \tag{11}$$

4. Calculate LoOP in each group
5. These spectra with $LoOP \geq 0.4$ are excluded.
6. Do PCA of all selected spectra in each group to get a feature matrix T and the respective eigen values λ.
7. Select the first k-th principal components(eigen spectra) while the variance contribution rate μ :

$$\mu = \frac{\sum\limits_{i=1}^{k} \lambda_i}{\sum\limits_{i=1}^{N} \lambda_i} > \theta \tag{12}$$

   where θ is a fixed given threshold (0.99 is used in our work). k is set to 2 when $k = 1$.
8. Reconstruct each selected spectra using obtained first k principal components
9. Calculate LoOP of remaining reconstructed spectra in each group again and exclude these spectra with $LoOP \geq 0.2$
10. Get the SNR weighted average spectrum as the template spectrum .

Following the above steps, the template spectra are successfully constructed in 216 groups (nearly 86%). Other 35 groups fail mainly because of lacking enough high quality spectra.

3.5 MK Class Labeling

Each spectrum should be labeled a subclass for latter usage in classification. We compare these spectra with three libraries and label each spectrum a subclass.

Danks & Dennefeld (1994) presented spectra for MK standards in the wavelength range 5800Å-10200Å. The stars cover the normal spectral types O to M and luminosity types I, III, and V. The projected slit width along the dispersion is about 4Å and the resolution R is about 1200. Two wavelength ranges [7500Å,7700Å] and [6800Å,7000Å] are masked to get rid of the strong telluric lines left in the spectra. We decrease the resolution of our templates to R 1200 by convolving a gaussian function. All template spectra and standard spectra are unified into the wavelength range [6100Å,9000Å] with a fixed step 4Å. For each template spectrum, the first four closest are chosen and the corresponding spectra are drawn together with the template spectra. Those figures are used for following visual inspection.

Bolton et al. (2012) describe the detail of the pipeline for SDSS III and publish the template used on the web page[1]. For stellar spectral classification, 123 templates are provided which are created from the

[1] http://www.sdss3.org/svn/repo/idlspec2d/tags/v5_4_45/templates/

Table 2 The main information of groups 59,180, 207 and 237

Group ID	All spectra	used spectra	Subclass	Subclass1	Subclass2	Subclass3
59	18534	3048	F4V	F3V/F5V	F8III-IV	F5
180	653	381	A1IV	A4V/A1V	A2V	A1V
207	27	13	CV	-	-	-
237	17231	325	M2	M1/M0	M1.5V/M3V	M2/M1

Notes: Subclass is the final MK class. Subclass1 is the best fit Mk class with Bolton et al. (2012). Subclass2 is the best fit Mk class with Danks & Dennefeld (1994). Subclass3 is the best fit Mk class with Luo et al. (2013).

full database of Indo-U.S. spectra. Each spectrum are labeled a MK class by matching with POLLUX database. The resolution R of these 123 spectra is about 2000 and the wavelength coverage is from 3500Å to 11200Å. These spectra are unified into the wavelength range [3800Å,9000Å] with a fixed step 1Å similar with the spectra in the library . Similarly, for each template spectrum, the first four closest are chosen and the corresponding spectra are drawn together with the template spectra. And those figures are used for following visual inspection.

As introduced above, the current library used for stellar classification in LAMOST contains 36 classes plus 20 subclasses specially for A-type star. The resolution R of these 56 spectra is about 2000 and the wavelength coverage is from 3800Å to 9200Å. These spectra are unified into the wavelength range [3800Å,9000Å] with a fixed step 1Å similar with the spectra in the library . Similarly, for each template spectrum, the first four closest are chosen and the corresponding spectra are drawn together with the template spectra. And those figures are used for following visual inspection.

Each template spectrum is visually inspected by checking the three figures drawn above. And then each spectrum is labeled a MK class. Meanwhile, those template spectra with bad data or low S/N R are excluded. Finally, there are 164 spectra and 65 different MK classes are left in the template library.

4 RESULTS AND DISCUSSIONS

A updated library is formed after adding some spectra of new subclasses and replacing some spectra. These spectra of these types not existing in our library are left. The library of the current version(V1.0) is publicly available on the web site[2]. The current library has been used in the new new version of LAMOST 1D pipeline for spectra after data release one.

4.1 Examples

Here we choose four typical groups (Group 59,180 ,207 and 237) to discuss in detail. The main information of these groups is shown in Table.2. The MK classes are F6V, A1IV, CV and M2 respectively. The finally constructed template spectra of these groups are as shown in Fig.5.

Group:59 This group contains the spectra in the color $g^* - r^*$ range [-0.34,-0.32]. There are totally 18,534 spectra and 3,048 spectra are selected from the first 5,000 spectra with the highest SNR. As shown in Fig 6, the spectrum is very close to F3V/F5V in Bolton et al. (2012). Consequently, the template spectrum is labeled as ’F4V’.

[2] http://sciwiki.lamost.org/lamost_sctl/v1

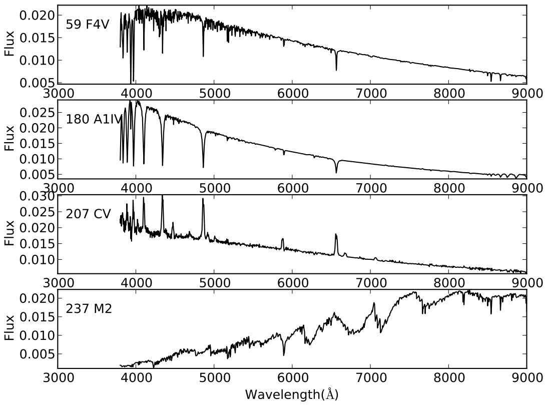

Fig. 5 The template spectra of groups 59,180, 207 and 237.

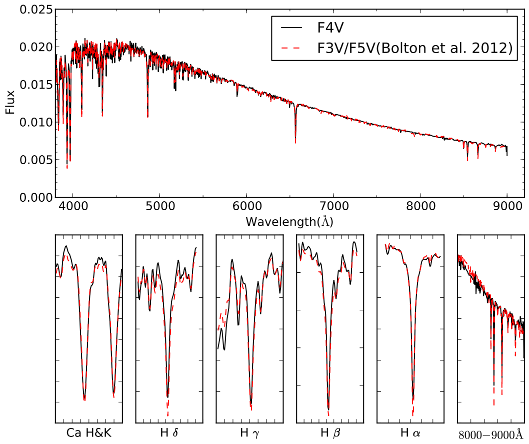

Fig. 6 The comparison of the template spectrum in group 59 with F3V/F5V in Bolton et al. (2012). The black line is the spectrum constructed in our work. The red one is the closest spectrum in Bolton et al. (2012).

As shown in Fig 7, the variance of the first principal component exceeds more than 99% of the total variance of the original data. That is due to the high similarity of the spectra in the group. Consequently, the reconstructed spectra using first two principal components are nearly similar to the origin spectra(see Fig 8).

Group:180 This group contains the spectra classified as 'A1IV' by pipeline. There are totally 653 spectra and 381 spectra are selected. The spectrum is labeled as 'A1IV' following the group selection criteria. As shown in Fig 9, the SNR of the template is a little larger than the template in Luo et al. (2013) while these two spectrum are nearly close to each other.

Similar with group 59, the variance of the first principal component also exceeds more than 99% of the total variance of the original data(see Fig 10). However, there are not as many spectra as in group 59.

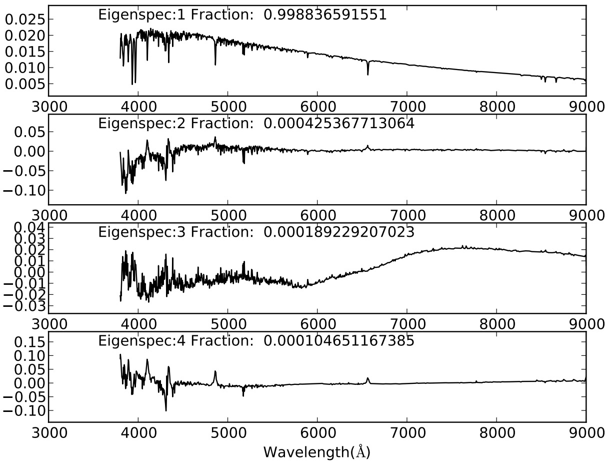

Fig. 7 The first four eigen spectra(principal components) of group 59.

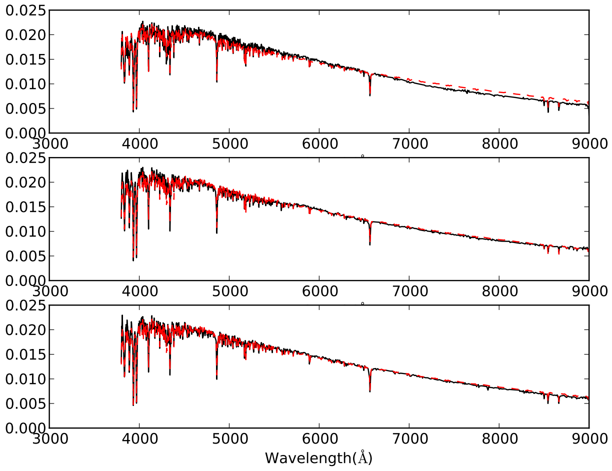

Fig. 8 Three examples of reconstructed spectra in group 59

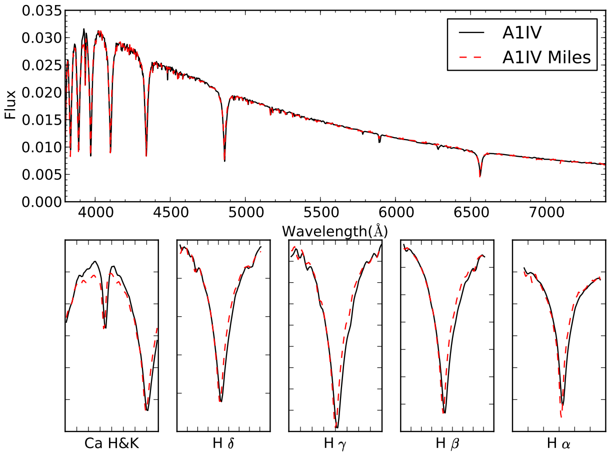

Fig. 9 The comparison of the template spectrum in group 59 with A1IV in Luo et al. (2013). The black line is the spectrum constructed in our work. The red one is the closest spectrum in Luo et al. (2013).

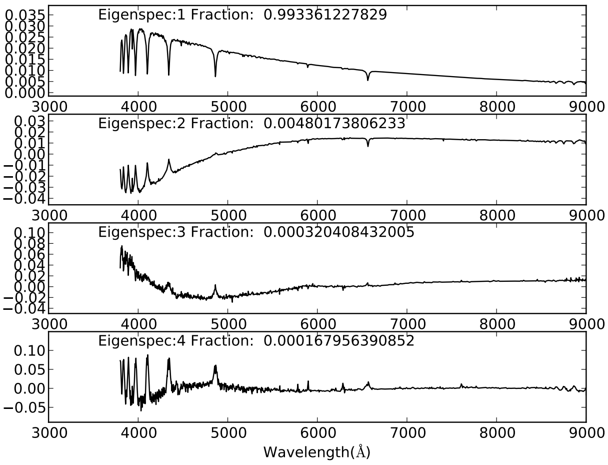

Fig. 10 The first four eigen spectra(principal components) of group 180.

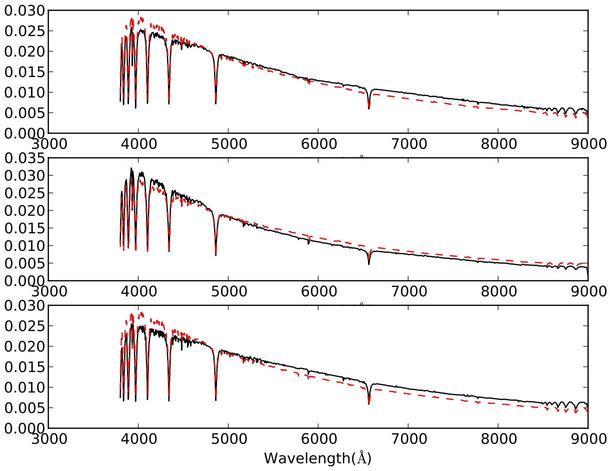

Fig. 11 Three examples of reconstructed spectra in group 180

Consequently, some spectra are not well reconstructed(as shown in Fig 11). In spite of this, the template spectrum is well constructed after excluding these badly reconstructed spectra.

Group:207 This group contains the spectra classified as 'CV'. There are totally 27 spectra and 13 spectra are selected. The spectrum is labeled as 'CV' following the group selection criteria.

Compared with normal stars, the spectra of CV stars are these with strong hydrogen Balmer and helium emission lines that typically signify ongoing accretion. As shown in Fig 12, the first two principal components show obvious and strong emission lines and the sum of the variances of these two principal components exceeds more than 99% of the total variance of the original data. Compared to normal stars misclassified as 'CV', the spectra of CV stars are almost faultlessly reconstructed(see Fig 13). And then these misclassified spectra are excluded in the next following steps.

Group:237

This group contains the spectra classified as 'M2'. There are totally 17,231 spectra and 325 spectra are selected from first 5,000 spectra with the highest SNR. Due to the existence of wavelength points with

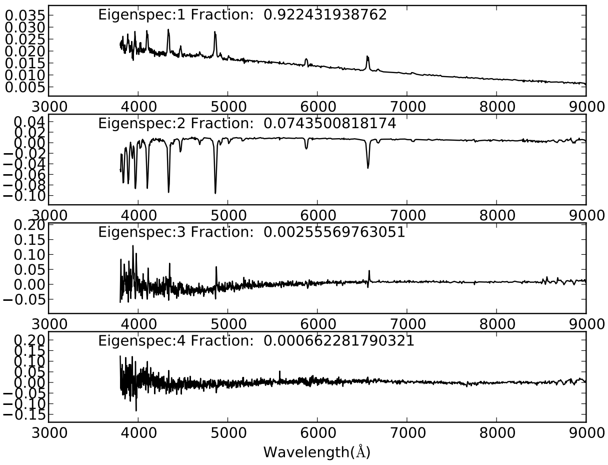

Fig. 12 The first four eigen spectra(principal components) of group 207. Note that the strong lines in eigen spectra 2 are emission lines not absorption lines.

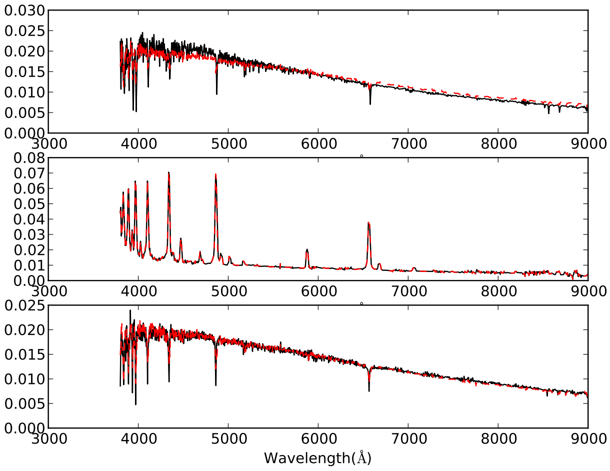

Fig. 13 Three examples of reconstructed spectra in group 207.

$flux \leq 0$, a large amount of spectra are excluded in this group. The spectrum is labeled as 'M2' following the group selection criteria.

As shown in Fig 14, the sum of the variances of the first two principal components exceeds more than 99% of the total variance of the original data. The selected spectra are not well reconstructed in the blue arm(as shown in Fig 15). In spite of this, the template spectrum is also well constructed.

To check the quality, we choose the M2-type template spectrum in the current template library (Wang et al. 2010) and compare it with the template spectrum of group 237(see Fig 16 upper panel).. Bochanski et al. (2007) presented template spectra of low-mass (M0-L0) dwarfs derived from over 4000 Sloan Digital Sky Survey spectra. We choose the M2-type template spectrum and alsoe compare it with the template spectrum of group 237(see Fig 16 bottom panel). As shown in Fig 16, we can infer that our constructed M2-type spectrum is a little better than these two spectra.

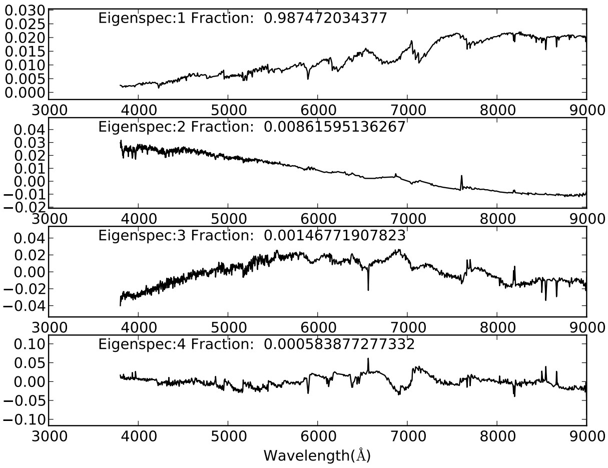

Fig. 14 The first four eigen spectra(principal components) of group 237.

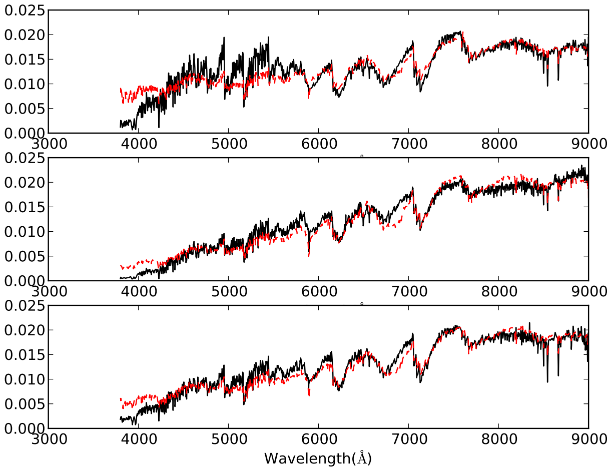

Fig. 15 Three examples of reconstructed spectra in group 237.

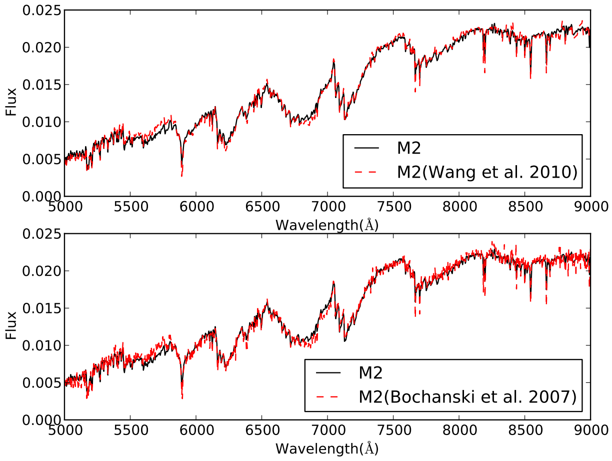

Fig. 16 The comparison of the template spectrum in group 59 with M2 in Bochanski et al. (2007). The black line is the spectrum constructed in our work. The red one is the closest spectrum in Bochanski et al. (2007).

Table 3 The comparison with McGurk et al. (2010)

Group ID	ESID1	ESID2	ESID3	Group ID	ESID1	ESID2	ESID3	Group ID	ESID1	ESID2	ESID3
38	1	2	3	59	19	20	18	80	39	38	40
39	1	2	3	60	19	20	21	81	40	41	39
40	2	3	1	61	20	23	22	82	41	40	42
41	3	2	4	62	24	23	25	83	42	41	43
42	4	5	3	63	25	24	26	84	43	42	44
43	5	6	7	64	26	25	27	85	44	43	45
44	6	7	5	65	27	26	28	86	45	44	46
45	7	8	6	66	28	27	29	87	45	46	44
46	8	9	7	67	28	29	27	88	46	47	45
47	9	10	11	68	29	30	28	89	47	48	46
48	11	10	12	69	30	29	31	90	48	47	49
49	12	11	10	70	30	31	29	91	49	48	50
50	12	13	11	71	-	-	-	92	50	49	51
51	13	14	12	72	32	33	31	93	51	50	52
52	14	15	13	73	33	32	34	94	52	51	53
53	15	14	16	74	34	35	33	95	53	52	54
54	16	15	17	75	35	34	36	96	53	54	52
55	16	17	15	76	35	36	37	97	54	55	53
56	17	18	16	77	36	37	35	98	55	54	53
57	18	17	19	78	37	38	36	99	55	54	53
58	18	19	20	79	38	37	39				

Notes: The ESID1, ESID2 and ESID3 are bin ID of the first three closest eigen spectra in McGurk et al. (2010) respectively.

4.2 Discussions

4.2.1 Comparison with McGurk et al. (2010)

McGurk et al. (2010) applied PCA to about 100,000 SEGUE spectra by dividing all spectra into 55 different bins. For each bin, the first four eigenspectra are published and the first one is a high SNR mean spectra. For the template spectra in all groups, to check the difference, we use similar method as MK class labeling to find three closest mean spectra in McGurk et al. (2010). As our color range is wider than McGurk et al. (2010), not all template spectra are similar with these in McGurk et al. (2010). As shown in Table 3, the groups with groups id from 38 to 99(totally 60 groups, not including group 71) cover nearly all mean spectra constructed by McGurk et al. (2010). The finally subclasses are from A3 to K3, which coincides with sayings in McGurk et al. (2010).

4.2.2 Comparison with current templates of LAMOST spectra analysis pipeline and Bolton et al. (2012)

As supplements to current templates, our constructed templates replace most spectra in the current library including the A-type stellar spectra. From the comparisons discussed in section 4.1, we can infer that newly constructed template spectra are a little better than current ones.

There are 123 stellar subclasses in Bolton et al. (2012). These template spectra are individual spectra in Indo-U.S. database. We notice that there are 80 subclasses which contain less than 500 spectra in SDSS

DR9. There are totally 12,897 spectra(about 1.66% in all 773,275 spectra) and 1,062 spectra(about 0.22% in all 475694 spectra) with SNR$>$ 10. Meanwhile, the average SNR of these spectra is about 4.93 which is much less the one of all spectra. In other words, there are mainly 43 stellar subclasses containing most spectra especially these spectra with high SNR.

Considering the differences between LAMOST spectra with spectra in other survey, these newly constructed spectra are more reliable and more similar to the spectra observed in LAMOST survey. That is because these template spectra are constructed from a healthy sum of spectra from LAMOST DR1.

4.2.3 Remaining problems

The result show that our constructed template spectra can be used in the classification of stellar spectra. There are also some problems needing to solve.

1. We notice that most of our template spectra are these of main sequence stars. To construct the template spectra of some other rare types such as K-type giants, DC and DZ white dwarfs, some spectra need to be labeled manually or picked out by other methods.
2. Three libraries are used to label each template. However, how to label the template spectra better is also a remaining problem to solve.

In addition, there are some outliers excluded in each group while constructing the templates. It is also worth studying these objects and finding rare types even new types of star.

5 SUMMARY

To improve the performance of LAMOST 1D Pipeline, a new stellar spectral classification template library is constructed. About 750,000 spectra are chosen from the first data release of LAMOST. All spectra are divided into different groups by our proposed pseudo color $g^* - r^*$. Additional groups are formed by different subclasses classified the current pipeline. In addition, some special types of psectra are manually picked out to add into the groups. And then these spectra are reduced by our proposed method, which use PCA to reconstruct the original spectra to get spectra of high quality and LoOP to exclude outliers in each group. The weighted average spectra in each group are treated as the constructed template spectra. By automatically matching with two known MK class standard spectra library and visual inspection, each spectrum is labeled a MK class. Also some incorrect spectra are excluded.

Acknowledgements The authors would like to thank the anonymous referees for their constructive comments, and are grateful to Prof. Yang Haifeng,Mrs Re Juanjuan for useful suggestions on methods and analysis of the results. This work is supported by the National Natural Science Foundation of China (Grant No. 61202315, 11078013 and 11203045,11303036,11303061).

Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

References

Aihara, H., Prieto, C. A., An, D., et al. 2011, ApJS, 193, 29 2

Almeida, J. S., & Prieto, C. A. 2013, ApJ, 763, 50 7

Bai, Z. 2012, Proceedings of the International Astronomical Union, 8, 189 2

Bailer-Jones, C. A., Irwin, M., & Hippel, T. V. 1998, MNRAS, 298, 361 7

Bochanski, J. J., West, A. A., Hawley, S. L., & Covey, K. R. 2007, AJ, 133, 531 13, 14

Bolton, A. S., Schlegel, D. J., Aubourg, É., et al. 2012, AJ, 144, 144 8, 9, 10, 15

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. 2000, in ACM Sigmod Record, vol. 29, 93–104 (ACM) 6

Carlin, J. L., Lépine, S., Newberg, H. J., et al. 2012, RAA, 12, 755 2

Chen, L., Hou, J.-L., Yu, J.-C., et al. 2012, RAA, 12, 805 2

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA, 12, 1197 2

Danks, A. C., & Dennefeld, M. 1994, PASP, 382–396 8, 9

Deng, L.-C., Newberg, H. J., Liu, C., et al. 2012, RAA, 12, 735 2

Falcón-Barroso, J., Sánchez-Blázquez, P., Vazdekis, A., et al. 2011, A&A, 532, A95 5

Glazebrook, K., Offer, A. R., & Deeley, K. 1998, ApJ, 492, 98 2

Jiang, B., Luo, A., Zhao, Y., & Wei, P. 2013, MNRAS, 430, 986 5, 7

Jolliffe, I. T. 2002, Principal component analysis (Springer verlag) 7

Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. 2009, in Proceedings of the 18th ACM conference on Information and knowledge management, 1649–1652 (ACM) 6

Lee, Y. S., Beers, T. C., Sivarani, T., et al. 2008, AJ, 136, 2022 7

Liu, X.-W., Yuan, H.-B., Huo, Z.-Y., et al. 2013, arXiv preprint arXiv:1306.5376 2

Luo, A.-L., Wu, Y., Zhao, J., & Zhao, G. 2008, in Proc. of SPIE Vol, vol. 7019, 701935–1 2

Luo, A.-L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, RAA, 12, 1243 2

Luo, A.-L., Zhang, Y.-X., & Zhao, Y.-H. 2004, in Astronomical Telescopes and Instrumentation, 756–764 (International Society for Optics and Photonics) 2

Luo, A.-L., & Zhao, Y.-H. 2001, Chinese Journal of Astronomy and Astrophysics, 1, 563 2

Luo, A.-L., et al. 2013, in preparation 2, 9, 10, 11

McGurk, R. C., Kimball, A. E., & Željko Ivezić 2010, AJ, 139, 1261 7, 15

Song, Y.-H., Luo, A.-L., Comte, G., et al. 2012, RAA, 12, 453 3

Tu, L., Luo, A., Wu, F., & Zhao, Y. 2010, Science China Physics, Mechanics and Astronomy, 53, 1928 7

Tu, L.-P., Luo, A.-L., Wu, F.-C., Wu, C., & Zhao, Y.-H. 2009, RAA, 9, 635 7

Željko Ivezić, Sesar, B., Jurić, M., et al. 2008, ApJ, 684, 287 3

Wang, F., Luo, A., & Zhao, Y. 2010, in SPIE Astronomical Telescopes and Instrumentation: Observational Frontiers of Astronomy for the New Decade, 774031–774031 (International Society for Optics and Photonics) 2, 13

Wang, F., Zhang, H., Luo, A.-L., et al. 2013, arXiv preprint arXiv:1306.1600 2

Wei, P., Luo, A., Li, Y., et al. 2013, MNRAS, 431, 1800 7

Whitney, C. 1983, Astronomy and Astrophysics Supplement Series, 51, 443 7

Wu, Y., Luo, A.-L., Li, H.-N., et al. 2011, RAA, 11, 924 3, 5

Yang, F., Carlin, J. L., Liu, C., et al. 2012, RAA, 12, 781 2

Yi, Z., Luo, A., Song, Y., et al. 2013, arXiv preprint arXiv:1306.4540 5

Yip, C., Connolly, A., Berk, D. V., et al. 2004, AJ, 128, 2603 7

Zhang, Y.-Y., Carlin, J. L., Yang, F., et al. 2012, RAA, 12, 792 2

Zhang, Y.-Y., Deng, L.-C., Liu, C., et al. 2013, AJ, 146, 34 5

Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, RAA, 12, 723 2

Zhao, J., Luo, A., Oswalt, T., & Zhao, G. 2013, AJ, 145, 169 5