RAA 2015 Vol. **15** No. **8**, 1137–1153 doi: 10.1088/1674–4527/15/8/004
http://www.raa-journal.org *http://www.iop.org/journals/raa*

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

Spectral classification of stars based on LAMOST spectra

Chao Liu[1], Wen-Yuan Cui[2], Bo Zhang[1], Jun-Chen Wan[1], Li-Cai Deng[1], Yong-Hui Hou[3], Yue-Fei Wang[3], Ming Yang[1] and Yong Zhang[3]

[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; *liuchao@bao.ac.cn*
[2] Department of Physics, Hebei Normal University, Shijiazhuang 050024, China
[3] Nanjing Institute of Astronomical Optics & Technology, National Astronomical Observatories, Chinese Academy of Sciences, Nanjing 210042, China

Received 2015 April 1; accepted 2015 May 20

Abstract In this work, we select spectra of stars with high signal-to-noise ratio from LAMOST data and map their MK classes to the spectral features. The equivalent widths of prominent spectral lines, which play a similar role as multi-color photometry, form a clean stellar locus well ordered by MK classes. The advantage of the stellar locus in line indices is that it gives a natural and continuous classification of stars consistent with either broadly used MK classes or stellar astrophysical parameters. We also employ an SVM-based classification algorithm to assign MK classes to LAMOST stellar spectra. We find that the completenesses of the classifications are up to 90% for A and G type stars, but they are down to about 50% for OB and K type stars. About 40% of the OB and K type stars are mis-classified as A and G type stars, respectively. This is likely due to the difference in the spectral features between late B type and early A type stars or between late G and early K type stars being very weak. The relatively poor performance of the automatic MK classification with SVM suggests that the direct use of line indices to classify stars is likely a more preferable choice.

Key words: techniques: spectroscopic — stars: general — stars: fundamental parameters — stars: statistics — Galaxy: stellar contents

1 INTRODUCTION

The classification of normal stars plays important roles not only in the understanding of stellar physics, but also in the study of the overall structure and evolution of the Milky Way. MK classification (Morgan & Keenan 1973) is one of the most broadly used systems based on the spectral features of a small number of standard stars. Compared to segregating the stars directly using effective temperature, surface gravity and chemical abundance, MK classification is simple and effective. A usual procedure to process a spectrum in a spectroscopic survey is to first assign spectra to the MK classes and then estimate the stellar astrophysical parameters using the MK classes as the starting points (e.g. Luo et al. 2015). Spectral classifications are also very helpful in selecting targets for follow-up studies. For instance, in order to select the blue horizontal branch stars from the whole dataset, one might first select all A type stars to reduce the size of the sample; in order to study the

circumstellar environment of young massive stars, one needs to first select the OB type stars from the full sample; or in order to search for AGB stars, one has to first identify the M giant stars from the database.

Alternatively, stars can be classified based on their color indices. Nowadays, billions of stars have accurate multi-band photometry covering the region from ultraviolet to infrared bands, e.g. GALEX (Bianchi et al. 2011), SDSS (Ahn et al. 2014), PanSTARRS (Tonry et al. 2012), 2MASS (Skrutskie et al. 2006), WISE (Wright et al. 2010), etc., which provide abundant information on stellar astrophysical parameters. For instance, Covey et al. (2007) identified different types of stars in SDSS+2MASS multi-color parameter space and showed a clear continuous stellar locus, on which any reasonable stellar classification system can be set up, including the well known MK class system. The biggest advantage of the continuous stellar locus in color parameter space is that it naturally reflects how the spectral energy distribution varies with stellar astrophysical parameters, such as the effective temperature, surface gravity and metallicity. Therefore, this type of analysis is very important in research on both stellar evolution and the overall features of the Milky Way. In fact, in the context of a survey with millions to billions of stars, the color-based stellar locus may be more effective and straightforward in stellar classifications (see their applications in Yanny et al. 2000; Majewski et al. 2003; Yanny et al. 2009 etc.). However, when one examines the deep-sky objects, especially along the Galactic mid-plane, most of the photometric color indices of stars are reddened by absorption and scattering of the interstellar medium. Although some remarkable works have been done (Schlegel et al. 1998; Schlafly et al. 2014; Chen et al. 2014), knowledge of the three-dimensional reddening distribution of the Milky Way is still very limited, leading to certain systematics that vary with lines of sight and distances in multi-color index-based stellar classifications. Moreover, in most cases, the color index is an integration of the spectrum over a wide range of wavelength, with details that are shown in spectral lines being smoothed out. Therefore, in general, color index-based classifications of stars cannot completely take the place of spectral-based classifications.

Up to now, most MK type classification of stars are done manually by comparing the spectra with a small set of standard stars (e.g. the samples in Corbally et al. 1994), which is not efficient when the sample is huge and not always reliable. Although efforts have been made to automate the process of MK classification by developing automatic software (e.g., Gray & Corbally 2014), it is still a non-trivial task since the real stellar spectra are not only sensitive to the effective temperature and luminosity, but also dependent on the elemental abundances. Moreover, in a large spectroscopic survey, the spectra may have a wide range of signal-to-noise ratio, making the spectral features not always as clear as the small set of well observed high-quality standard spectra. For such a large spectroscopic survey, the mis-classification that arises from the template-matching technique based on some standard stars may be significant for data with low signal-to-noise ratio, and it may subsequently affect the effort of searching for peculiar and rare objects.

Other efforts at star classification based on automatic algorithms have been done in the past twenty years by various works. These algorithms include methods based on metric-distance (e.g. LaSala 1994), artificial neural networks (e.g. Bailer-Bailer-Jones 1997), fuzzy logic (e.g. Carricajo et al. 2004), etc. It can be noted that Bailer-Jones et al. (2008) and Saglia et al. (2012) described applications of a support vector machine (SVM) in star-galaxy-QSO classifications. In general, this new technique can also be used for the classification of stars.

Recently, the LAMOST survey (Cui et al. 2012; Zhao et al. 2012; Deng et al. 2012; Liu et al. 2015) has collected more than 4 million stellar spectra in its second internal data release (DR2). Unlike SDSS, the LAMOST survey does not include a photometric survey with its spectroscopic one, and targets are selected from several external photometric catalogs (Carlin et al. 2012; Yuan et al. 2015). This makes it difficult to establish star classification based on the photometric color indices since the multiple input catalogs are not well calibrated. With only the stellar spectra, it is not trivial to automatically classify stars into different MK types. The LAMOST pipeline (Luo et al. 2012; Luo et al. 2015) runs a cross-correlation based algorithm (correlation function initial; CFI)

to assign the MK types to each stellar spectrum. However, due to some technical issues (e.g., noise in the spectra, distortion of the continua due to interstellar extinction, limitations of the synthetic library used in CFI, etc.), this classification system, which has already appeared in the LAMOST catalog, is not very reliable, especially for O, B, A and M type stars. Therefore, a robust and reliable automatic classification method suitable for all spectral classes that can be applied to LAMOST spectra is anxiously required.

In this work, we map the MK classes to the parameter space defined by indices of prominent lines in stellar spectra. The line indices naturally form a stellar locus from the hottest to the coolest stars because there is a smooth transition in these spectral lines with the effective temperature and surface gravity of the stars. In principle, unlike the broadly used discrete MK classes, the line indices can automatically provide a continuous set of classes, although the elemental abundance may affect it. Meanwhile, the MK classes or other classification systems can be easily mapped to the parameter space defined by line index to find their counterparts. We also employ the SVM method to assign stars to MK classes and compare the results it produces with line index-based classification. We suggest that line index-based classification is one of the most robust ways to classify stars in the era of large data.

The paper is organized as follows. In Section 2, we give a brief introduction to the LAMOST survey and the selection of data to which classification analysis is applied. We also give a detailed definition of the indices for more than 20 spectral lines in the rest of that section. In Section 3, we show the features of the stellar locus in the parameter space defined by line indices and how the locus is associated with MK classes. In Section 4, we employ an SVM to classify the stars into MK types. We then compare the stellar locus-based classification with the SVM-based MK classification. We raise discussions in Section 5 and draw a short conclusion in Section 6.

2 DATA

2.1 LAMOST and SIMBAD Data

The LAMOST telescope, also called the Guo Shou Jing Telescope, is a 4-m reflecting Schmidt telescope with 4000 fibers configured on a 5-degree field of view (Cui et al. 2012; Zhao et al. 2012). The LAMOST Milky Way survey will finally target more than 5 million stellar spectra with a resolution of $R \sim 1800$ in observations that will be taken over 5 yr (Deng et al. 2012; Liu et al. 2014c). Early results indicate that it can obtain more spectra than originally planned after the LAMOST team released the DR2 catalog, which contains about 4 million stellar spectra, by the end of 2014.

We select about 1.52 million stellar spectra with signal-to-noise ratio larger than 20 (which means the averaged signal-to-noise ratio in both the g and i bands is larger than 20) to investigate how the spectral features vary with stellar classes. In order to identify their MK classes, we cross match them with the SIMBAD catalog (Wenger et al. 2000)[1] and obtain 3134 spectra of normal stars with MK classification flags in the SIMBAD catalog.

Table 1 shows the distribution of the MK classes for the sample. It demonstrates that the sample is nonuniformly distributed in MK classes. The stars between late B and early A, and the G and K type stars are prominent in the sample. In addition, there are many more main-sequence stars than giant stars, and supergiant stars are very rare.

2.2 Line Indices

In order to associate the stellar classes with spectral features, we measure the line indices of spectral lines instead of using the full spectra. In general, line indices do not require flux calibration, which is very hard to calibrate in the LAMOST pipeline due to having a complicated instrument, e.g.,

[1] *http://simbad.u-strasbg.fr/Simbad*

Table 1 The Number of MK Classes for the LAMOST-SIMBAD Sample

Type	Total	V	IV/III	II/I	Type	Total	V	IV/III	II/I
O5	1	1	0	0	F7	15	12	2	1
O7	2	2	0	0	F8	55	38	15	2
O8	1	1	0	0	F9	22	11	11	0
O9	4	3	1	0	G0	435	398	36	1
B0	14	9	5	0	G1	21	19	2	0
B1	15	8	7	0	G2	57	33	24	0
B2	19	11	8	0	G3	16	9	7	0
B3	9	7	0	2	G4	21	17	2	2
B4	9	7	2	0	G5	280	218	62	0
B5	34	18	15	1	G6	23	11	12	0
B6	3	2	0	1	G7	17	7	10	0
B7	23	13	10	0	G8	224	114	109	1
B8	75	65	10	0	G9	27	8	19	0
B9	175	157	18	0	K0	183	89	74	3
A0	420	386	30	4	K1	56	13	31	0
A1	67	63	4	0	K2	83	38	33	0
A2	186	175	10	1	K3	25	15	7	0
A3	61	60	1	0	K4	25	17	8	0
A4	11	10	0	1	K5	21	14	5	1
A5	43	39	2	2	K6	9	9	0	0
A6	3	2	1	0	K7	11	11	0	0
A7	27	21	6	0	K8	5	5	0	0
A8	12	10	0	2	K9	2	2	0	0
A9	1	0	1	0	M0	21	12	9	0
F0	53	34	16	3	M1	6	3	3	0
F1	4	1	3	0	M2	13	9	4	0
F2	36	27	8	1	M3	12	7	5	0
F3	14	10	4	0	M4	10	9	1	0
F4	7	5	1	1	M5	4	2	2	0
F5	67	54	12	1	M7	2	0	1	1
F6	36	21	14	1	M8	1	1	0	0

4000 fibers with different lengths, 16 spectrographs with slightly different performances, etc. The instruments are also very robust against random noise. Although it is very difficult to cleanly subtract the sky background from the spectra, the blue part of the spectra is less influenced. Fortunately, most of the well known line indices, e.g. the Lick indices (Worthey et al. 1994; Worthey & Ottaviani 1997), are in blue.

The principle behind the selection of spectral lines is two-fold. First, the lines should be strong enough that they can be effectively detected in low resolution spectra. Second, the lines should be sensitive to the effective temperature, surface gravity and metallicity so that they can be utilized in the classification.

Table 2 lists all 27 spectral lines used in this work. Most of them are adopted from Lick indices (Worthey et al. 1994; Worthey & Ottaviani 1997; Cohen et al. 1998). To better separate OB type stars, we add three helium lines. In addition, since the CaII K line may also be often used for classification, it is also included based on the definition by Beers et al. (1999).

We define the line index in terms of equivalent width (EW) with the following equation (Worthey et al. 1994)

$$\mathrm{EW} = \int \left[1 - \frac{f_{\mathrm{line}}(\lambda)}{f_{\mathrm{cont}}(\lambda)}\right] d\lambda \,, \tag{1}$$

where $f_{\mathrm{cont}}(\lambda)$ and $f_{\mathrm{line}}(\lambda)$ are the fluxes of the continuum and the spectral line, respectively, both of which are functions of wavelength λ. The continuum f_{cont} is estimated via linear interpolation of the fluxes located in the pseudocontinuum region on both sides of the bandpass for each index

Table 2 Line Index Definitions

Name	Index Bandpass (Å)	Pseudocontinua (Å)
CaII K[a]	3927.7–3939.7	3903–3923 4000–4020
Hδ[b]	4083.50–4122.25	4041.60–4079.75 4128.50–4161.00
CN[c]	4143.375–4178.375	4081.375–4118.875 4245.375–4285.375
Ca4227[c]	4223.500–4236.000	4212.250–4221.000 4242.250–4252.250
G4300[c]	4282.625–4317.625	4267.625–4283.875 4320.125–4336.375
Hγ[b]	4319.75–4363.50	4283.50–4319.75 4367.25–4419.75
Fe4383[c]	4370.375–4421.625	4360.375–4371.625 4444.125–4456.625
He4388	4381–4399	4365–4380 4398–4408
Ca4455[c]	4453.375–4475.875	4447.125–4455.875 4478.375–4493.375
He4471	4462–4475	4450–4463 4485-4495
Fe4531[c]	4515.500–4560.500	4505.500–4515.500 4561.750–4580.500
He4542	4536–4548	4526–4536 4548–4558
Fe4668[c]	4635.250–4721.500	4612.750–4631.500 4744.000–4757.750
Hβ[b]	4847.875–4876.625	4827.875–4847.875 4876.625–4891.625
Fe5015[c]	4977.750–5054.000	4946.500–4977.750 5054.000–5065.250
Mg_1[c]	5069.125–5134.125	4895.125–4957.625 5301.125–5366.125
Mg_2[c]	5154.125–5196.625	4895.125–4957.625 5301.125–5366.125
Mg_b[c]	5160.125–5192.625	5142.625–5161.375 5191.375–5206.375
Fe5270[c]	5245.650–5285.650	5233.150–5248.150 5285.650–5318.150
Fe5335[c]	5312.125–5352.125	5304.625–5315.875 5353.375–5363.375
Fe5406[c]	5387.500–5415.000	5376.250–5387.500 5415.000–5425.000
Fe5709[c]	5698.375–5722.125	5674.625–5698.375 5724.625–5738.375
Fe5782[c]	5778.375–5798.375	5767.125–5777.125 5799.625–5813.375
NaD[c]	5878.625–5911.125	5862.375–5877.375 5923.875-5949.875
TiO_1[c]	5938.375–5995.875	5818.375–5850.875 6040.375–6105.375
TiO2[c]	6191.375–6273.875	6068.375–6143.375 6374.375–6416.875
Hα[d]	6548.00–6578.00	6420.00–6455.00 6600.00–6640.00

Notes: [a] Beers et al. (1999); [b] Worthey & Ottaviani (1997); [c] Worthey et al. (1994); [d] Cohen et al. (1998).

(see Table 2). The line index under this definition is in Å. It is noted that the measurement of equivalent widths of the lines is based on the rest-frame spectra, in which the radial velocities have been corrected. The value of the radial velocity is adopted from the LAMOST catalog. For spectra with signal-to-noise ratio larger than 20, the median uncertainty of the equivalent widths of the lines is smaller than 0.1Å.

Figure 1 shows the median equivalent width of different spectral lines for each class of star. It is evident that the spectral lines are not equally sensitive to the stellar classes. All Balmer lines, i.e., Hα, Hβ, Hγ and Hδ, separate the classes well. The magnesium lines are also sensitive to the classes, particularly for late type stars. Although the change in the iron lines is not as significant as that in the Mg lines, they also show a clear trend in different classes. Finally, the TiO lines are very sensitive to the M type stars. It seems that many of the spectral lines are correlated. Hence, we do not need to use all of them for the classification. We select Hγ as the representative Balmer line, since it has the largest amplitude of variation among the Balmer lines. Then we average over the Mg_1, Mg_2 and Mg_b which represent the composite line index of Mg. We also average over all nine iron lines to form the composite line index of Fe. Finally, we select the G band (CH) and TiO2 to represent the molecular bands. In total, we give five (composite) line indices for all selected stars.

Although the CaII K line is frequently used in classifications and parameterizations, we decide not to use it because it does not provide extra information about spectral types and is located near

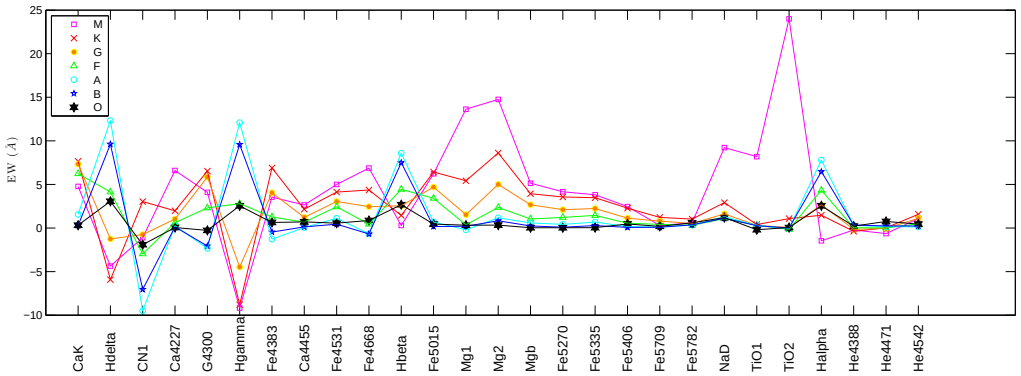

Fig. 1 The figure shows the line indices for different MK classes. Each grid in the x-axis corresponds to a spectral line and the y-axis indicates the median equivalent width for each line. The colors and symbols represent the O (*black hexagrams*), B (*blue pentagrams*), A (*large cyan circles*), F (*green triangles*), G (*small orange circles*), K (*red crosses*), and M (*magenta rectangles*) types.

the blue end, in which the wavelength calibration and efficiency of the instrument are not as good as other lines, making the line index of CaII K be not very stable.

3 LINE INDEX-BASED CLASSIFICATION

Figure 2 shows the stellar loci in the parameter space defined by five line indices, Hγ, Mg, Fe, G band and TiO2 for all 1.5 million selected stars (their distributions are represented as blue contours). The unit of the x- and y-axes is Å. The hollow circles with neighboring dark gray labels mark the median positions of the main-sequence MK classes from the SIMBAD catalog. For instance, a hollow circle with a label "G2V" is the median value of stars with types G0V, G1V, G2V and G3V. Also, a symbol with "G5V" is the median value of all stars with types G3V, G4V, G5V and G6V. The neighboring circles overlap each other by one decimal subtype in order to make the stellar locus smoother. Similarly, the red asterisks with the neighboring labels indicate the locus of MK classes for giant stars. The detailed positions of these circles (asterisks) for main-sequence (giant) stars are listed in Table 3 (Table 4). Figure 3 magnifies the smaller regions for better illustration of early type stars.

First, the stars from O to M type can be well separated and ordered in the Hγ vs. G4300 plane, shown in the top-right panel of Figure 2. In the Hγ vs. Fe plane (top-left panel), the stars from O to G type are well separated, but the M and K type stars overlap at the top of the stellar loci and are hard to disentangle. A similar trend is shown in the Hγ vs. Mg plane in the middle-left panel of Figure 2. However, the late type main-sequence stars are easily distinguished in the Mg vs. Fe, Fe vs. G4300, and Fe vs. TiO2 planes shown in the middle-right, bottom-left, and bottom-right panels, respectively. The early type stars in these planes are clumpy and hard to separate from each other. Combined with the five line indices, we are able to separate all types of main-sequence stars from O to M type.

Second, the separation of the luminosity type works well for K and M giant stars. Especially in the Mg vs. Fe (middle-right panel of Fig. 2) and Fe vs. TiO2 (bottom-right panel) planes, the cool star ends of the stellar loci for main-sequence and giant stars go to different directions. In the Mg vs. Fe plane, the locus of the main-sequence stars goes down toward smaller Fe and larger Mg indices at the coolest end, while the locus of the giant stars goes up toward larger Fe but smaller Mg indices. A similar trend can also be seen in the Fe vs. TiO2 plane. However, it is very hard to

Table 3 The Median Locus for Equivalent Width of the Spectral Lines for Main-sequence Stars

Type	EW_{G4300} (Å)	$EW_{H\gamma}$ (Å)	EW_{Mg} (Å)	EW_{Fe} (Å)	EW_{TiO2} (Å)	Number of stars
O6–9	–0.27±0.23	2.55±0.23	0.46±0.12	0.28±0.43	–0.00±0.28	6
B0–3	–1.07±0.45	4.20±1.72	0.22±0.12	0.10±0.35	0.03±0.49	35
B3–6	–1.60±0.59	6.76±1.62	0.12±0.11	0.35±0.28	–0.04±0.33	34
B6–9	–2.50±1.04	11.43±2.68	–0.01±0.28	0.49±1.96	–0.01±5.37	237
A0–3	–2.52±1.32	12.45±2.65	0.08±0.58	0.49±0.90	–0.05±0.43	684
A3–6	–1.29±1.01	11.07±2.09	0.51±0.33	0.62±0.29	–0.11±0.40	111
A6–9	–0.27±0.80	8.88±1.86	0.74±0.30	0.72±0.30	–0.27±83.98	33
F0–3	0.89±1.04	5.64±2.28	1.01±0.38	0.89±0.31	–0.21±7.70	72
F3–6	2.18±1.20	2.90±2.75	1.19±0.36	1.26±0.56	–0.17±7.61	90
F6–9	3.42±1.11	0.87±2.16	1.49±0.39	1.62±0.50	–0.15±3.48	82
G0–3	5.38±1.08	–3.15±2.29	1.98±0.61	2.59±1.93	–0.03±0.48	459
G3–6	5.86±0.98	–4.39±2.45	2.44±0.67	3.32±2.42	0.05±0.49	255
G6–9	6.21±0.72	–6.45±1.97	2.88±0.76	4.41±1.45	0.19±0.46	140
K0–3	6.28±1.35	–7.72±3.53	3.23±0.91	5.23±2.42	0.45±2.52	155
K3–6	5.92±0.84	–10.45±2.78	4.25±0.55	11.63±2.49	1.96±11.10	55
K6–9	5.12±0.94	–10.07±4.24	4.14±0.71	12.68±1.33	5.08±15.70	27
M0–3	3.52±1.31	–9.21±3.28	3.26±0.81	11.78±0.91	21.30±15.85	31
M3–6	2.72±1.02	–11.57±5.31	3.01±0.45	11.11±2.66	33.69±20.42	18

Table 4 The Median Locus for Equivalent Width of the Spectral Lines for Stars with Luminosity Types IV or III

Type	EW_{G4300} (Å)	$EW_{H\gamma}$ (Å)	EW_{Mg} (Å)	EW_{Fe} (Å)	EW_{TiO2} (Å)	Number of stars
B0–3	–0.51±0.34	2.78±1.78	0.34±0.21	0.09±0.33	0.07±13.83	20
B3–6	–1.55±0.59	6.70±1.33	0.17±0.21	0.16±0.32	0.01±0.34	17
B6–9	–1.85±1.04	8.36±2.35	0.04±0.35	0.50±0.34	–0.02±13.38	38
A0–3	–1.61±0.95	11.03±1.97	0.24±0.40	0.41±0.36	0.03±0.32	45
A3–6	–0.89±1.12	10.41±1.75	0.55±0.28	0.50±0.20	–0.35±0.08	4
A6–9	–0.83±1.16	10.02±2.17	0.63±0.32	0.63±0.23	–0.22±0.18	8
F0–3	1.09±1.10	5.50±2.58	1.14±0.21	1.04±0.36	–0.24±14.86	31
F3–6	2.46±0.91	2.41±1.80	1.34±0.31	1.30±0.45	–0.11±11.07	31
F6–9	3.62±1.03	0.61±2.01	1.61±0.33	1.56±0.50	–0.10±2.06	42
G0–3	5.05±1.28	–2.50±2.91	2.07±0.81	2.50±1.49	0.03±9.27	69
G3–6	6.27±1.17	–5.71±2.78	2.76±0.78	3.49±1.42	0.27±0.61	83
G6–9	6.93±0.91	–7.72±2.15	3.40±2.55	4.10±3.74	0.80±10.62	150
K0–3	6.90±0.88	–8.99±3.16	3.99±0.77	5.77±2.14	1.32±4.82	189
K3–6	6.67±0.93	–10.01±3.38	4.47±0.80	8.38±2.20	3.18±3.80	24
M0–3	5.79±1.42	–9.16±3.06	5.37±0.87	9.75±1.87	27.04±13.31	21
M3–6	3.59±0.99	–4.85±9.35	6.54±0.67	6.84±1.67	42.62±6.16	8

disentangle the early type giant stars, e.g. B, A and F type giant stars. These types of giant stars are located at almost exactly the same position as the same type main-sequence stars. According to Gray & Corbally (2009), some weaker lines, such as OII (at 4070, 4076, 4348 and 4416Å), SIIV at 4116Å, etc., may be helpful for discriminating the luminosity types of B type stars. However, they are very weak in the low-resolution LAMOST spectra and may be significantly affected by noise.

It is worth pointing out that variation in the Fe index for late type stars is mostly but not exactly related to the Fe lines, but is significantly affected by prominent molecular bands, e.g. TiO, happens to overlap at the same wavelength. Also, the response of the Mg index to cool stars is actually dominated by the MgH band.

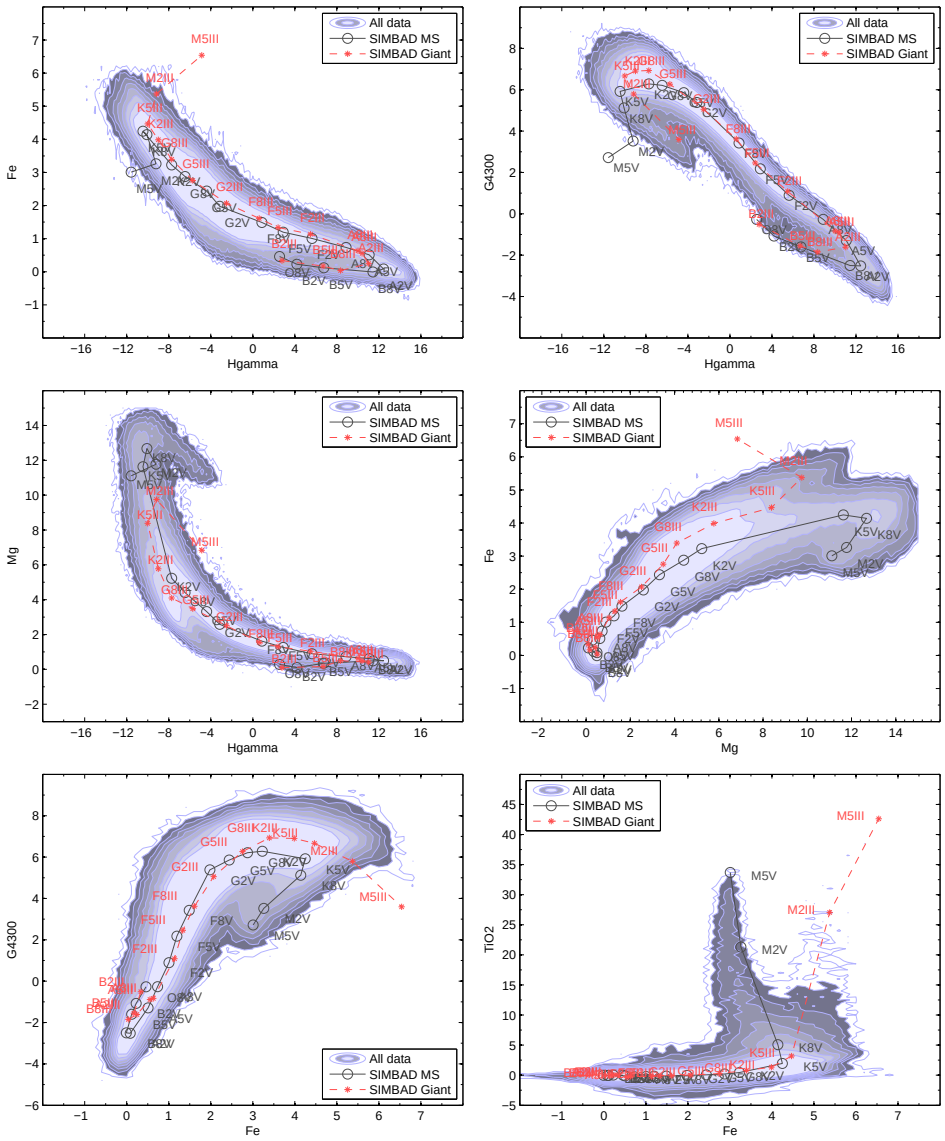

Fig. 2 The contours show the distribution of LAMOST stars in the parameter space defined by line indices. The top-left, top-right, middle-left, middle-right, bottom-left and bottom-right panels are for Hγ vs. Fe, Hγ vs. G band, Hγ vs. Mg, Mg vs. Fe, Fe vs. G band, and Fe vs. TiO2 planes respectively. The black lines with circles indicate the stellar loci of the main-sequence stars with their MK designations from the SIMBAD database, while the red dashed lines with asterisks indicate the stellar loci of the giant stars (type IV/III) with their MK designations from the SIMBAD database.

It can also be noted that the dispersions shown in Figure 2 are not only contributed by uncertainties in the line indices, which are only about 0.1Å. These dispersions may be intrinsic and related to the broad diversity in metallicity. In this paper, we mainly focus on the effective temperature, which corresponds to the spectral types, and the surface gravity, which is related to the luminosity. The

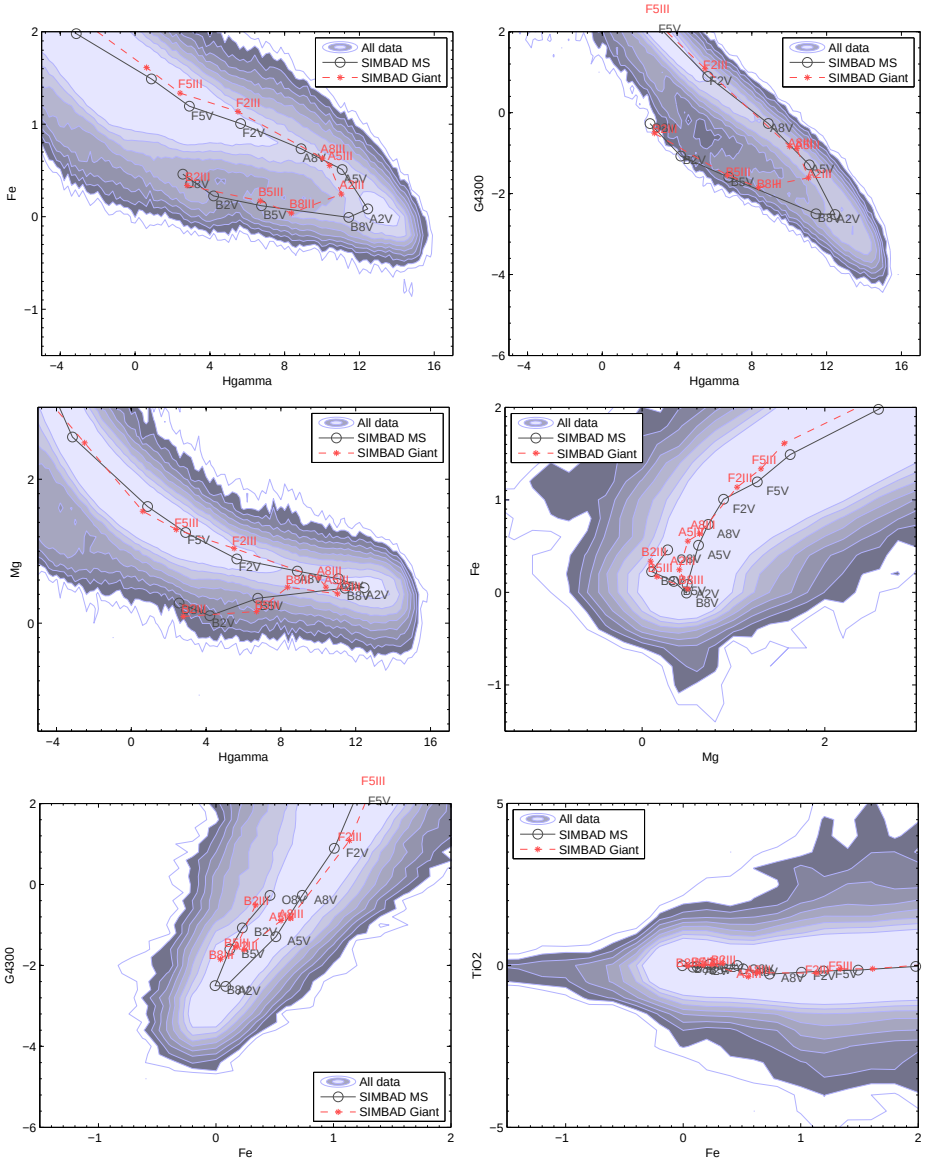

Fig. 3 The panels are same as their counterparts in Fig. 2, but are enlarged to show details around the locus of early type stars.

effect of metallicity in spectral classification may be more complicated, because it also reflects the evolution of different stellar populations. We would like to leave this topic for future works.

The classification of stars based on their stellar loci can be done by looking up the line indices in Tables 3 and 4. For any statistical study of the Milky Way, one can conveniently select stars located in a segment of the stellar loci in Figures 2 and 3 according to the listed MK classes. Compared with the classical MK classes, the stellar loci in line indices, acting just like the color indices in a

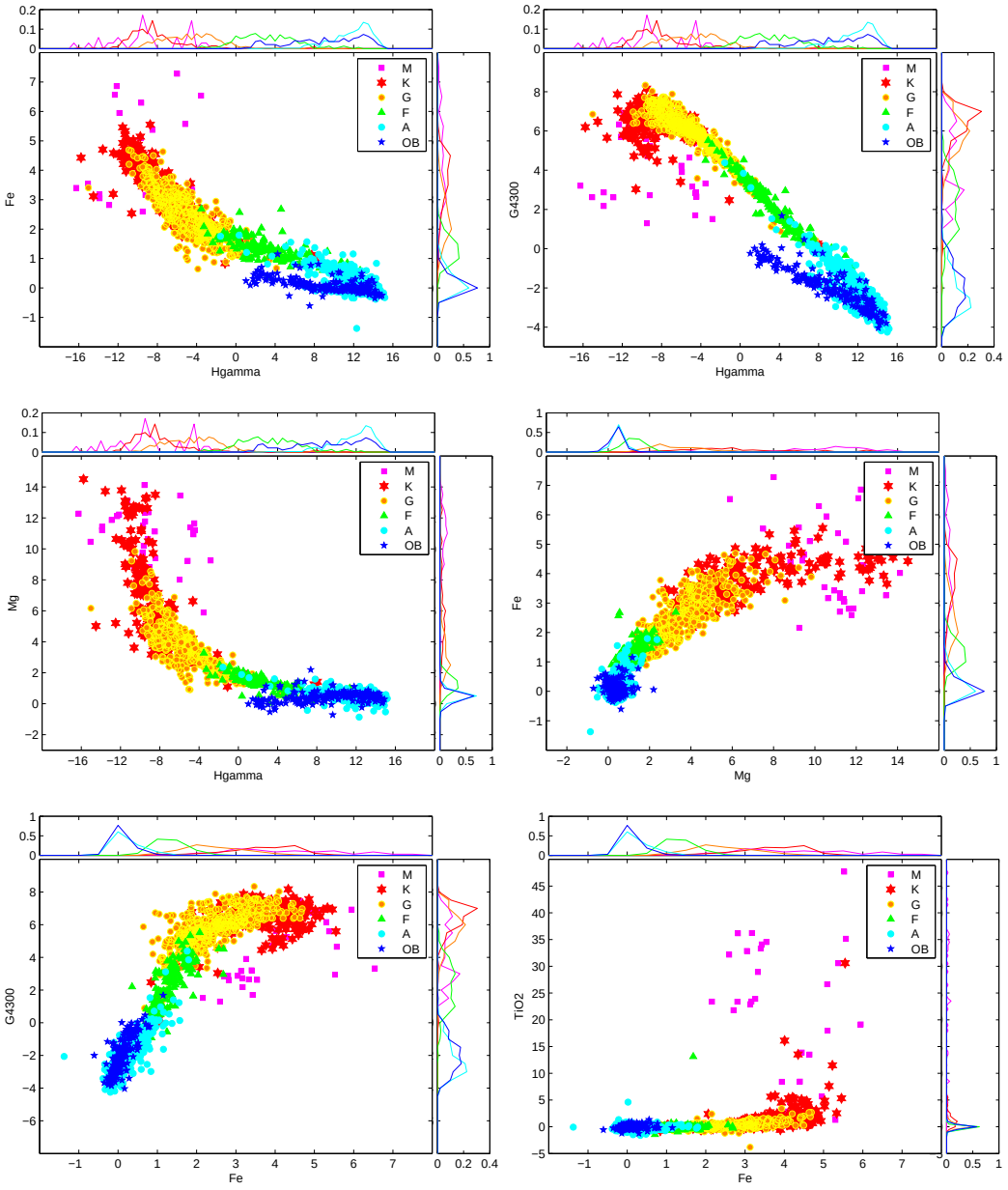

Fig. 4 The distribution of the SIMBAD MK classes of the test data resulting from an SVM applied to the parameter space defined by line indices. The top-left, top-right, middle-left, middle-right, bottom-left and bottom-right panels show the distributions in Hγ vs. Fe, Hγ vs. *G* band, Hγ vs. Mg, Mg vs. Fe, Fe vs. *G* band, and Fe vs. TiO2 planes respectively. The marginalized distributions of one single line index for different spectral types are shown at the right and top edges of each panel. The colors and symbols denote the OB (*blue pentagrams*), A (*large cyan circles*), F (*green triangles*), G (*small orange circles*), K (*red hexagrams*), and M (*magenta rectangles*) types.

multi-band photometric system, provide natural and continuous sequences of stars, which are easier to analyze using quantitative statistics. More discussions can be found in Section 5.

4 SVM-BASED CLASSIFICATION

Alternatively, we can also translate the line index-based stellar loci into the MK class system for individual stars. To do this, we employ an SVM algorithm to automatically assign a stellar spectrum to the proper MK class.

SVM is a supervised machine learning algorithm for classification and regression (Cortes & Vapnik 1995). In general, a supervised algorithm uses a small sample with multi-dimensional input variables and known labels of classes as the training dataset. The SVM classification is built in two steps. First, with the training dataset, the optimized non-linear boundaries among different classes for the input parameters are determined and defined by a subset of the training dataset, which is called the support vectors, located around the boundaries. Second, for a given input datum, the trained SVM model gives a prediction of the class depending on where the input datum is located with respect to the support vectors. A typical sample of SVM classification can be found in Liu et al. (2014a) and a sample of SVM regression can be found in Liu et al. (2012) and Liu et al. (2014b).

Chang & Lin (2011) provide a package in multiple programming languages, LIBSVM[2], that implements the SVM algorithm. Here, we use LIBSVM to classify the stars into MK types based on their line indices. We arbitrarily separate the 3134 stars with both high signal-to-noise ratio LAMOST spectra and SIMBAD MK types into two equal-size groups. One group is selected as the training dataset to train the SVM, and the other is used as the test dataset to assess the performance. We use all 27 line indices listed in Table 2 as the input vector. We only adopt six classes, which are OB, A, F, G, K and M, and ignore the decimal subtypes and the luminosity types in the SVM classification. O and B types are merged into one class since there are only very few O type stars in the sample.

Figure 4 shows the stellar loci composed of the ~ 1500 test dataset with color coded SIMBAD class labels from the parameter space defined by line indices Hγ, Fe, Mg, G band and TiO2. Because we use the SIMBAD MK classes as the training dataset, it implies that we assume the SIMBAD MK classes to be the "standard" classes with which to compare results. Figure 5 shows a similar set of stellar loci with exactly the same test dataset as in Figure 4, but with colors representing the MK classes derived from SVM.

Comparing Figure 4 and Figure 5 can give a qualitative impression of the performance of the SVM classification. It is obviously seen that some OB type stars (blue pentagrams), located in the bottom-right corner in the Hγ vs. Fe and Hγ vs. G4300 planes and which are shown in the two top panels in Figure 4, are mistakenly classified as A type stars (cyan circles) by the SVM method, as shown in the corresponding panels in Figure 5. Moreover, although the SVM classification works quite well for stars from M to F type, it can still be relatively hard when applying artificial boundaries to F, G, and K type stars in Figure 5.

A quantitative assessment of the performance of the SVM classification is based on the so called confusion matrix shown in Table 5, in which the columns stand for the "true" class labels and rows stand for the SVM derived class labels. The intersections give the percentage of stars which belong to the class in a column but are assigned to the class in a row by the SVM. The diagonal items show the completeness of the classification, i.e., the percentage of stars in class X being correctly assigned to the same class. The last column in Table 5 gives the contamination, which is the percentage of stars in the derived class X being contaminated by other classes.

Table 5 shows that A and G type stars have the highest completeness larger than 90%. This means that more than 90% of A or G type stars are correctly classified by the SVM algorithm. The completenesses of F and M type stars are about 72% and 68%, respectively, which are still

[2] *http://www.csie.ntu.edu.tw/∼cjlin/libsvm/*

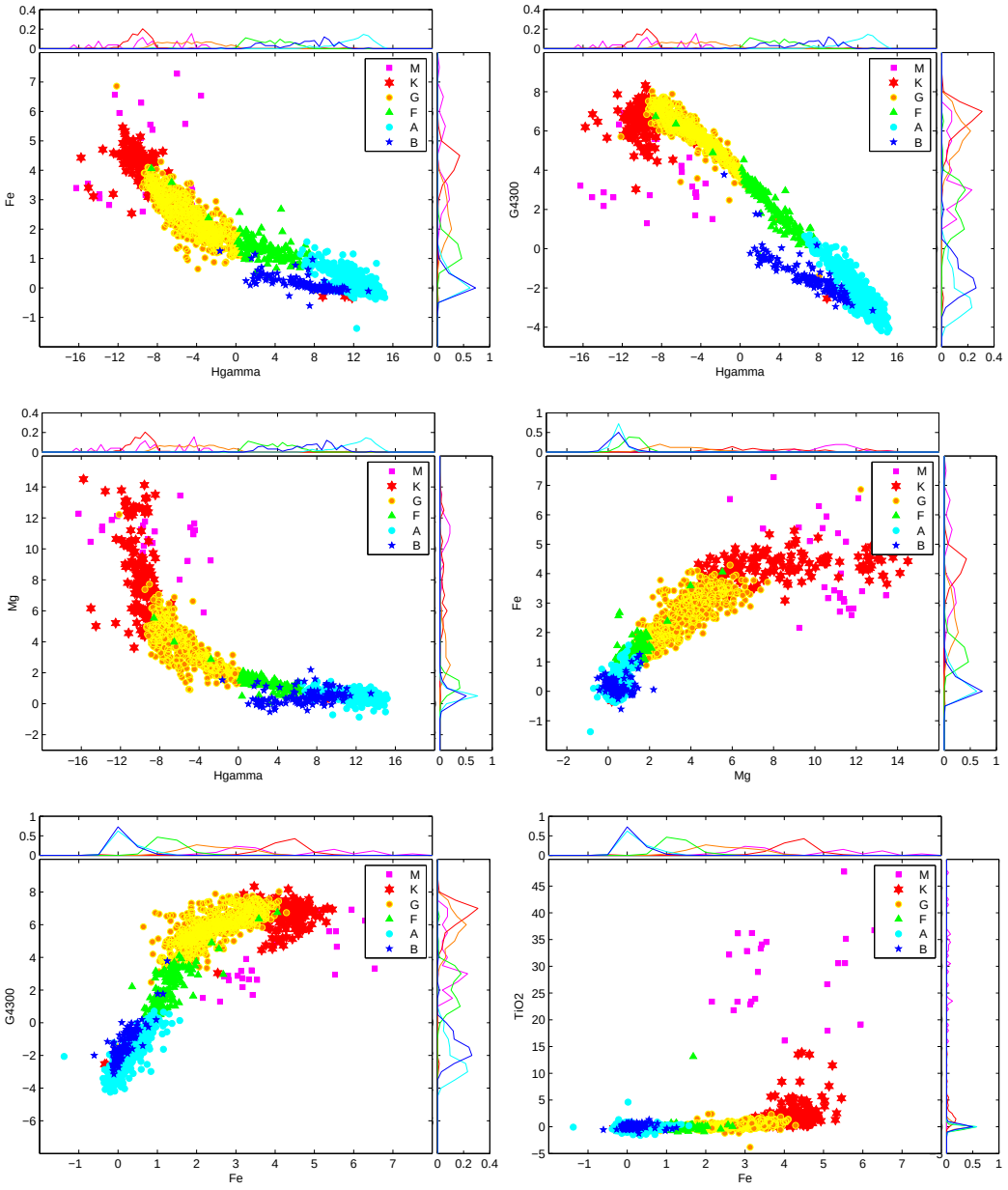

Fig. 5 The distribution of the MK classes of the test data resulting from an SVM applied to the parameter space defined by line indices. The top-left, top-right, middle-left, middle-right, bottom-left, and bottom-right panels show the distributions in Hγ vs. Fe, Hγ vs. *G* band, Hγ vs. Mg, Mg vs. Fe, Fe vs. *G* band, and Fe vs. TiO2 planes, respectively. The marginalized distributions of one single line index for different spectral types are shown at the right and top edges of each panel. The colors and symbols denote the OB (*blue pentagrams*), A (*large cyan circles*), F (*green triangles*), G (*small orange circles*), K (*red hexagrams*), and M (*magenta rectangles*) types.

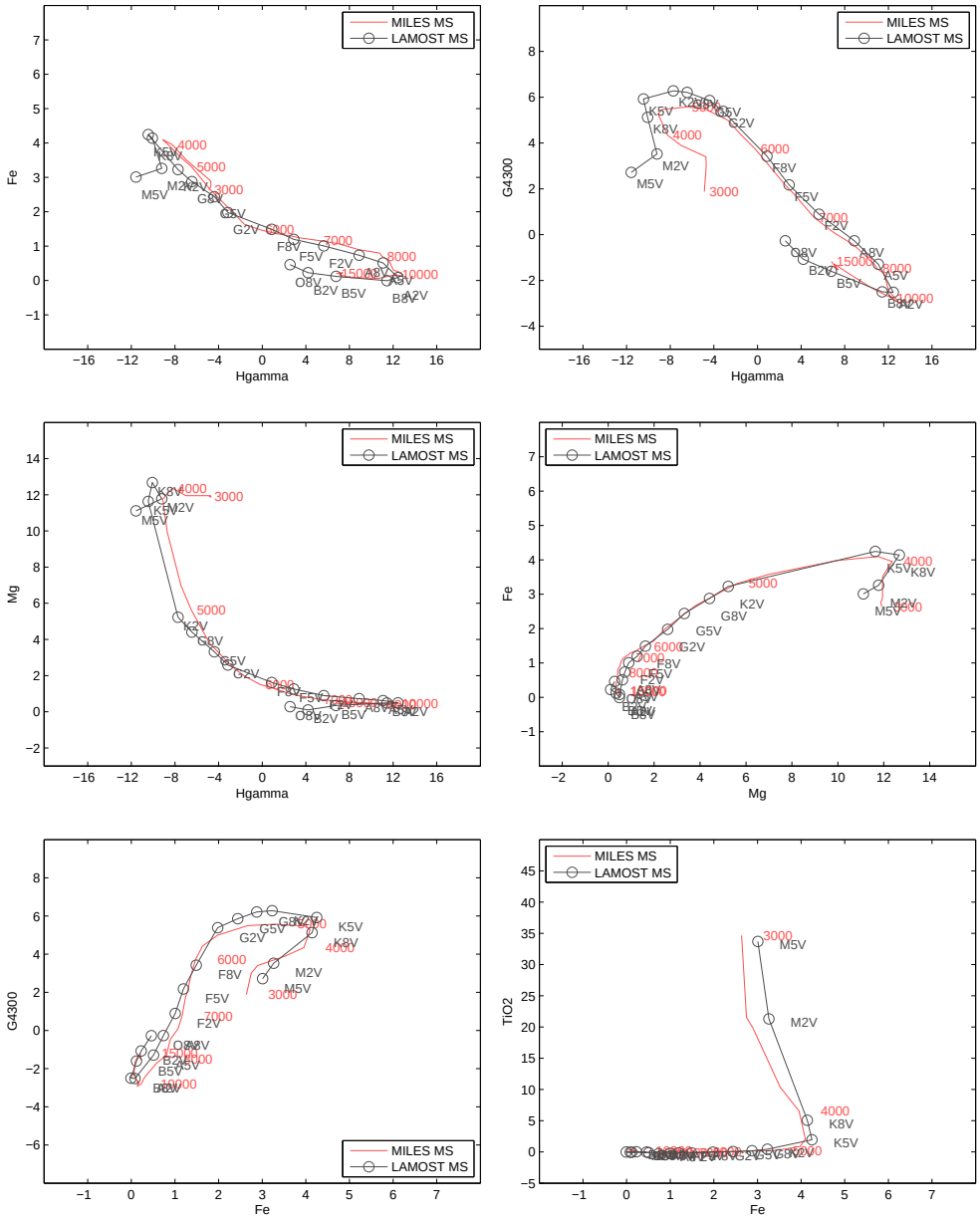

Fig. 6 The stellar loci for main-sequence (luminosity type V) stars calculated from the median location of each subtype resulting from the parameter space defined by line indices. The top-left, top-right, middle-left, middle-right, bottom-left, and bottom-right panels show the loci in Hγ vs. Fe, Hγ vs. *G* band, Hγ vs. Mg, Mg vs. Fe, Fe vs. *G* band, and Fe vs. TiO2 planes respectively. The black lines with circles indicate the stellar loci of main-sequence stars from LAMOST spectra, while the red lines with the effective temperatures labeled show the main-sequence locus of the MILES library.

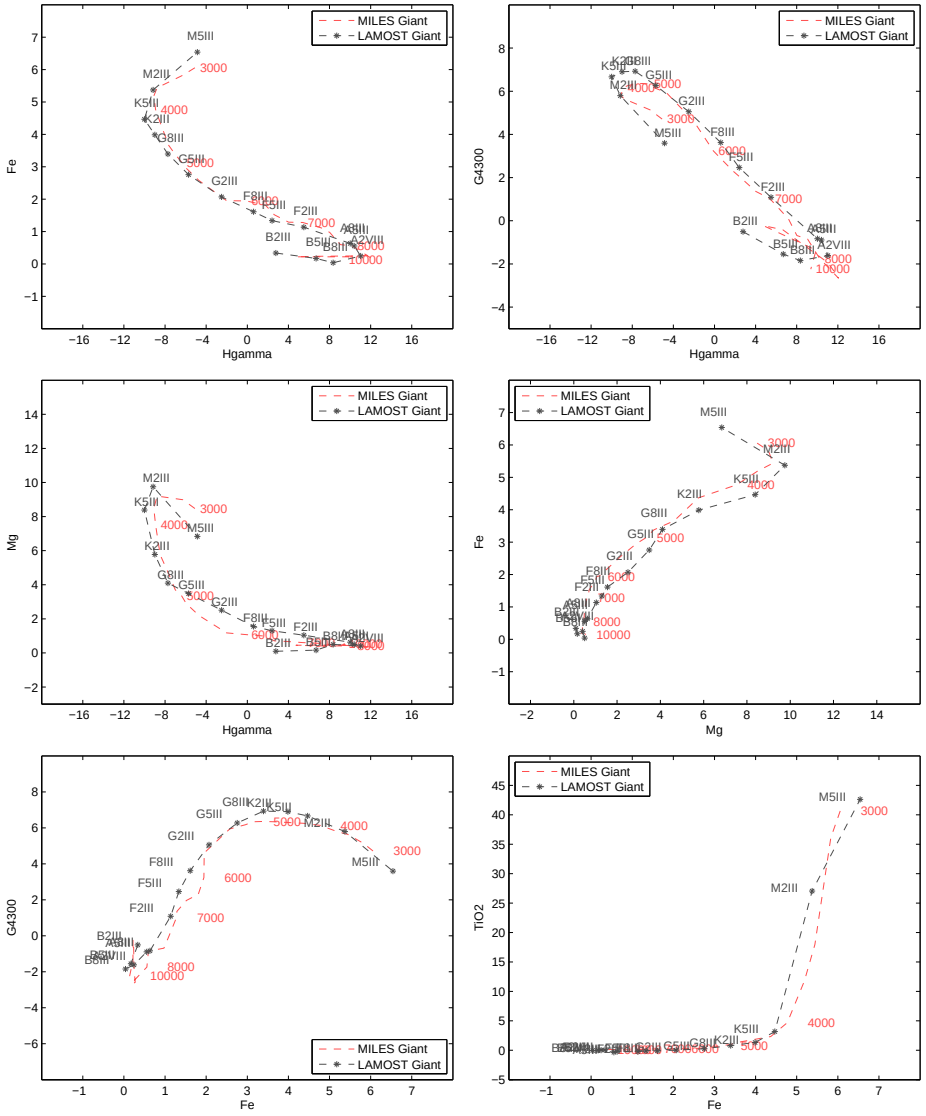

Fig. 7 The stellar loci for giant (luminosity type IV/III) stars calculated from the median location of each subtype resulting from the parameter space defined by line indices. The top-left, top-right, middle-left, middle-right, bottom-left, and bottom-right panels show the loci in Hγ vs. Fe, Hγ vs. G band, Hγ vs. Mg, Mg vs. Fe, Fe vs. G band, and Fe vs. TiO2 planes respectively. The black dashed lines with asterisks indicate the stellar loci of giant stars from LAMOST spectra, while the red dashed lines with the effective temperatures labeled show the loci of giants from the MILES library.

acceptable. However, the completeness for OB and K stars is only about 52%, implying that almost half of these two types of stars are mis-classified in the SVM classifier. Indeed, about 44% of "true" OB type stars are mis-classified as A type. A similar percent of "true" K type stars are mis-classified as G type. This is probably because the spectral features of late B (early K) type stars are very similar as those of early A (late G) type stars and thus they are very difficult for an SVM to disentangle. It

Table 5 The Confusion Matrix in terms of Percentage of the SVM-based MK Classification

		SIMBAD						
		OB	A	F	G	K	M	Contamination
SVM	OB	52.60%	6.97%	1.94%	0.00%	0.00%	0.00%	24.06%
	A	44.79%	90.38%	8.39%	0.53%	1.43%	0.00%	21.83%
	F	1.56%	1.68%	72.26%	3.57%	0.95%	0.00%	22.22%
	G	0.52%	0.48%	17.42%	90.91%	43.81%	2.86%	19.43%
	K	0.52%	0.48%	0.00%	4.99%	52.86%	28.57%	26.97%
	M	0.00%	0.00%	0.00%	0.00%	0.95%	68.57%	7.69%

may also be that the adopted "true" classes from the SIMBAD database are compiled from various literatures and classified by eye, and hence are not well calibrated with each other. Therefore, the large dispersions in the manually assigned MK classes may affect the performance of classification using an SVM.

5 DISCUSSION

5.1 The Discrepancy between MILES and LAMOST Spectra

In order to provide an external comparison of the stellar locus arising from the parameter space defined by the line indices, we calculate the same line indices for a sample prepared with the MILES library (Sánchez-Blázquez et al. 2006), which contains 985 bright stellar spectra with wide extensions in stellar parameters. We plot the stellar loci of MILES data with red lines in Figures 6 and 7 for main-sequence and giant stars, respectively. To be convenient, we also label the averaged effective temperatures along the stellar loci as a reference. They show that the stellar loci derived from LAMOST and MILES do not completely overlap each other, especially for late type dwarf stars and all giant stars. However, according to Tables 3 and 4, these differences are mostly within the uncertainty of 1- or 2-σ, and the overall shifts in the loci of the MILES library in most panels of Figures 6 and 7 are likely systematic. Looking back to the bottom-right panel of Figure 2, it can be seen that the SIMBAD stellar loci for M dwarf stars show a similar systematic bias compared to the full sample of LAMOST data (the contours). Therefore, it is likely that the M dwarf stars in the SIMBAD database are a biased sample and cannot provide an accurate representation for the majority of the LAMOST M dwarf samples. This also indicates that the stellar loci based on line indices derived from one survey should not be directly extended to another survey. Calibrations in the line indices and in the sample selection function are necessary before applying the extension.

5.2 How to Make the Decision, the MK Class or the Line Index-based Stellar Locus?

In the previous sections we show two kinds of classifications. The line index-based stellar locus orders the different types of stars as a simple sequence, along which the effective temperature monotonically changes from coolest to hottest. No hard boundary has to be set in the stellar locus to artificially separate the stars into discrete classes. The users who want to select specific stars for their statistical studies on the Milky Way can simply cut the data from any segment of the stellar locus.

On the other hand, the SVM based classification assigns discrete MK type labels to stars based on prior knowledge—the SIMBAD MK class labels. The compiled MK classes in the SIMBAD database are from many related literatures, and most of them are done by comparing the spectra with a small sample of standard stars by eye. This may raise significant inconsistency between the literatures. Calibrations among different literatures seem to be very difficult, since the MK classes are not continuous but rather discrete.

A realistic issue in large spectroscopic surveys, such as the LAMOST survey, is that millions of stars are observed and it is impossible to inspect each spectrum by eye. As shown in the exercise with SVM classification in Section 4, state-of-the-art machine learning techniques may not be very helpful because they need to be trained by prior knowledge which should be accurate and self-consistent.

Based on this analysis, we therefore suggest that LAMOST users rely on the stellar locus that is based on line indices, rather than directly using the derived MK classes from the catalog, to select the proper types of stars to satisfy their specific requirements. If these users want to compare their sample with literatures, which may use MK classes, they can quantitatively calculate the percentage of completeness and contaminations via a comparison of the stellar loci defined by SIMBAD and the SVM MK classes.

6 CONCLUSIONS

In this paper, we revisit the fundamental issue of stellar classification using 3000 high signal-to-noise ratio LAMOST spectra with known MK classes obtained from cross-identification with the SIMBAD database. Although the MK classes have been widely used for more than 70 yr and have become a standard, it seems not easy to adapt the large amount of data from current spectroscopic surveys. The MK classes are constructed based on a very small sample of standard stars, which are mostly very bright and located in a local volume near the Sun. New spectroscopic surveys, e.g. SDSS and LAMOST, can detect deep sky targets as far away as 100 kpc and hence contain millions of stars from very different populations compared with the solar neighborhood. The current standard star library then becomes incomplete compared with a survey that is a few orders of magnitude larger. Another issue is that almost all stars with known MK classes are classified by eye. This is unfortunately impossible in the era of large data. A third issue is that the MK classes are discrete, which makes them difficult to calibrate.

We map the MK classes to the parameter space defined by line indices and find that the resulting stellar loci can describe the MK classes well. Moreover, these loci naturally follow the change in effective temperature. For late type stars, the different luminosity types can also be disentangled by using the stellar loci.

We then investigate the performance of an automatic MK classification based on the SVM technique. We find that although A, F, G, and M type stars can be accurately classified, almost half of the B or K type stars are mis-classified.

We therefore suggest that the classification of stars should be based on continuous stellar loci representing line indices. The advantages of the stellar loci are that (1) they are continuous and one can cut a group of data at any point on the loci; (2) the stellar loci are consistent with effective temperatures; and (3) after selecting a group of stars from the stellar loci, one can easily estimate the completeness and contamination of the sample in terms of MK classes.

Acknowledgements This work is supported by the Strategic Priority Research Program "The Emergence of Cosmological Structures" of the Chinese Academy of Sciences (Grant No. XDB09000000) and the National Key Basic Research Program of China (2014CB845700). CL acknowledges the National Natural Science Foundation of China (NSFC, Grant Nos. 11373032, 11333003 and U1231119). The Guo Shou Jing 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 National Astronomical Observatories, Chinese Academy of Sciences.

References

Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17
Bailer-Jones, C. A. L. 1997, PASP, 109, 932
Bailer-Jones, C. A. L., Smith, K. W., Tiede, C., Sordo, R., & Vallenari, A. 2008, MNRAS, 391, 1838
Beers, T. C., Rossi, S., Norris, J. E., Ryan, S. G., & Shefler, T. 1999, AJ, 117, 981
Bianchi, L., Herald, J., Efremova, B., et al. 2011, Ap&SS, 335, 161
Carlin, J. L., Lépine, S., Newberg, H. J., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 755
Carricajo, I., Manteiga, M., Rodríguez, A., & C. 2004, Lecture Notes and Essays in Astrophysics, 1, 153
Chang, C.-C., & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology (TIST), 2, 27
Chen, B.-Q., Liu, X.-W., Yuan, H.-B., et al. 2014, MNRAS, 443, 1192
Cohen, J. G., Blakeslee, J. P., & Ryzhov, A. 1998, ApJ, 496, 808
Corbally, C. J., Gray, R. O., & Garrison, R. F. 1994, Astronomical Society of the Pacific Conference Series, 60, The MK Process at 50 Years. A Powerful Tool for Astrophysical Insight
Cortes, C., & Vapnik, V. 1995, Machine learning, 20, 273
Covey, K. R., Ivezić, Ž., Schlegel, D., et al. 2007, AJ, 134, 2398
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1197
Deng, L.-C., Newberg, H. J., Liu, C., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 735
Gray, R. O., & Corbally, J., C. 2009, Stellar Spectral Classification (Princeton: Princeton Univ. Press)
Gray, R. O., & Corbally, C. J. 2014, AJ, 147, 80
LaSala, J. 1994, in Astronomical Society of the Pacific Conference Series, 60, The MK Process at 50 Years: A Powerful Tool for Astrophysical Insight, eds. C. J. Corbally, R. O. Gray, & R. F. Garrison, 312
Liu, C., Bailer-Jones, C. A. L., Sordo, R., et al. 2012, MNRAS, 426, 2463
Liu, C., Deng, L.-C., Carlin, J. L., et al. 2014a, ApJ, 790, 110
Liu, C., Fang, M., Wu, Y., et al. 2014b, ApJ in press, arXiv:1411.0235
Liu, X.-W., Yuan, H.-B., Huo, Z.-Y., et al. 2014c, in IAU Symposium, 298, eds. S. Feltzing, G. Zhao, N. A. Walton, & P. Whitelock, 310
Liu, X. W., Zhao, G., & Hou, J. L. 2015, RAA (Research in Astronomy and Astrophysics), 15, 1089
Luo, A.-L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1243
Luo, A. L., Zhao, Y. H., Zhao, G., et al. 2015, RAA (Research in Astronomy and Astrophysics), 15, 1095
Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., & Ostheimer, J. C. 2003, ApJ, 599, 1082
Morgan, W. W., & Keenan, P. C. 1973, ARA&A, 11, 29
Saglia, R. P., Tonry, J. L., Bender, R., et al. 2012, ApJ, 746, 128
Sánchez-Blázquez, P., Peletier, R. F., Jiménez-Vicente, J., et al. 2006, MNRAS, 371, 703
Schlafly, E. F., Green, G., Finkbeiner, D. P., et al. 2014, ApJ, 789, 15
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Tonry, J. L., Stubbs, C. W., Lykke, K. R., et al. 2012, ApJ, 750, 99
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Worthey, G., Faber, S. M., Gonzalez, J. J., & Burstein, D. 1994, ApJS, 94, 687
Worthey, G., & Ottaviani, D. L. 1997, ApJS, 111, 377
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Yanny, B., Newberg, H. J., Kent, S., et al. 2000, ApJ, 540, 825
Yanny, B., Newberg, H. J., Johnson, J. A., et al. 2009, ApJ, 700, 1282
Yuan, H.-B., Liu, X.-W., Huo, Z.-Y., et al. 2015, MNRAS, 448, 855
Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, RAA (Research in Astronomy and Astrophysics), 12, 723