Automated classification of stellar spectra

Author: Cristina Jiménez Palau
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Josep Maria Solanes Majúa

Abstract: We have analyzed the performance of a PCA-based automated classifier of stellar spectra into the MK system, using as benchmark the dataset of optical spectra listed in the SDSS-DR15. We have found that it is possible to account for 99% of the total variance arising from good-quality A, F, G, K, and M stellar spectra with only three principal components, though we have ended up using four to further increase the discriminant power of our methodology. The projections of a subset of $50,000$ good-quality spectra on this 4D space have been used to determine the most probable spectral type of test stars in samples of spectra of increasing quality, with which we have evaluated the goodness of our classification procedure. Within a general scenario of excellent results, we found that a Gaussian Kernel performs somewhat better than a Top-Hat Kernel when calculating membership probabilities, that the efficiency of our classification method improves with the S/N of the spectra, and that the classification of G-type stars is the less reliable and that of F-type stars the most incomplete.

I. INTRODUCTION

Stellar spectra contain a huge amount of information on the photospheres of stars such as their ionization state, which in turn gives an objective measure of the photosphere's temperature. Most stars are currently classified under the system proposed by W.W. Morgan and P.C. Keenan at their first photographic spectral atlas *An Atlas of Stellar Spectra* [1]. In the MK system stars are grouped according to their spectral characteristics using the letters O, B, A, F, G, K, and M, optionally followed with a finer subdivision by Arabic numerals (0-9). Physically, the spectral classes indicate the effective temperature of the star's atmosphere and are normally listed from the hottest (O type) to the coolest (M type). The MK classification is a purely taxonomic system; the different categories of spectra are defined by a set of standard stars based on distinguishing features.

Not so very long ago, this classification was performed matching the overall appearance of a spectrum to the closest MK standard one [2]. Not only are these manual techniques very time-consuming and involve a great amount of human resources, they also constitute a subjective process, since a given spectrum may be classified differently by different people. Besides, recently, the advent of multifiber spectroscopy has increased by orders of magnitude the number of stars with available spectra, which has created the necessity of developing new and powerful classifiers. These problems are currently alleviated through the use of automated techniques that are fast, objective and repeatable [3].

The aim of this work is precisely to statistically evaluate the performance of an automated method of classification for stellar spectra based on the Principal Component Analysis (PCA), an unsupervised, non-parametric statistical tool primarily used for dimensionality reduction in machine learning. We will be supported in this task by the large dataset of optical spectra of Milky Way stars obtained within the Sloan Extension for Galactic Understanding and Exploration program of the Sloan Digital Sky Survey (SDSS), which we will use as a benchmark. Besides, we modified some of the tools developed by [4] for the analysis of galaxy spectra, to fulfill the specific requirements of the present study.

This manuscript is organized as it follows. In Section II, we discuss the data selection and processing. In Section III, we briefly explain the mathematical principles of the PCA when it is applied to spectra and define our training dataset of reference spectra. We also determine the minimum number of dimensions that are needed for an adequate description of most of the MK system, as well as the specific regions (CR) occupied by each spectral type in this low-dimensional subspace. In Section IV we quantify and discuss the performance of our PCA-based method as a predictor of stellar spectra, where we use two different kernel density estimators to calculate the probability the membership probability. Finally, in Section V we summarize our main results.

II. DATA TREATMENT

A. Selection of the spectra

The main sample of stellar spectra used in the present analysis has been selected from the set of optical spectra classified by the SDSS spectral pipeline as being of survey quality. These spectra are retrieved uploading a query to the SQL Search form of the fifteenth Data Release of the SDSS (SDSS-DR15). In this query we require the SDSS spectra: 1) to be of the 'STAR' class; 2) to have a subtype that coincides with one of the principal spectral types of the MK system; and 3) to have the flag 'ZWARNING' set to zero, which eliminates those spectra with potential problems on their classification or on their estimated redshift. In addition, we require the mean value of the signal-to-noise ratio (S/N) of the spectra to be greater or equal than 20 to guarantee a minimum level of measurement accuracy of the main spectral features, especially the absorption ones.

This query has been also used to evaluate the general quality of the stellar spectra accessible from the SDSS. We have carried out a simple statistical test consisting in determining how many stellar spectra from a random subset of 100,000 of them match both the requirements specified in the SQL Query defined above and the additional restrictions we imposed in the processing of spectra (see Sec. II B). The results of this test are shown in Table I. One can notice, for instance, the acute shortage of *O*- and *B*-type stars, something that was to be expected given that these are massive and very luminous objects, so they are very short lived. In fact, there are only 366 *O*-type stars and 1239 *B*-type stars in the entire SDSS catalog, which has forced us to remove these two types from the present study. For their part, *M*-type stars show the opposite behavior. They are by far the most common, but their intrinsic faintness and the presence of strong molecular absorption bands on the shortest wavelength part of the optical window make it difficult to obtain high-quality spectra. In fact, there are only 9992 *M*-type stars in the entire SDSS-DR15 catalog that satisfy all our quality constraints.

Spectral type	**Fractional Abundance (%)**	**Quality Spectra (%)**
O	1.35	45.9
B	3.35	37.0
A	12.1	51.7
F	23.4	68.6
G	5.35	71.3
K	18.0	43.7
M	33.5	1.20
Other	7.30	–

TABLE I: Fractional abundances and percentage of spectra of a given type in the SDSS-DR15 catalog that met our quality criteria. Results are based on a sample of 100,000 random spectra.

B. Processing of the spectra

The SDSS-DR15 Sky Server provides spectra in `.fits` format that were processed in two stages. The first stage was aimed to normalize each individual spectrum. Previously, we removed spectral bins affected by sky lines or bad data by blacking out those pixels whose errors were set to infinity, those that had the mask bit `BRIGHTSKY` activated[1], as well as the ones with a negative flux, since they have no physical meaning. Any spectrum containing more than 10% of troublesome pixels according to these criteria was fully discarded. All the accepted spectra were then normalized following the expression [4]

$$f_{ij}^{\rm nor} = f_{ij}^{\rm obs} \cdot \left(\sum_k \frac{f_{ik}^{\rm con}}{N_i} \right)^{-1} , \qquad (1)$$

where $f_{ij}^{\rm obs}$ is the original flux of the jth pixel of spectrum i, and $f_{ik}^{\rm con}$ is the continuum flux at the unmasked kth pixdel and N_i the total number of unmasked pixels of this same spectrum.

The removal of problematic pixels introduces null values in the array of normalized fluxes, thus preventing the proper performance of the PCA. To solve this problem, we employed the gap-correction formalism described in [5] and implemented in the `astroML`[2] module.

III. PRINCIPAL COMPONENT ANALYSIS

The Principal Component Analysis (PCA) is a well-know feature extraction method that provides an optimal representation of the data in terms of a few mutually-orthogonal linear variables, called Principal Components (PC). When applied to a set of spectra, this objective and a priori-free technique enables to reduce the highly multidimensional data of each spectrum — the number p of pixels per SDSS spectrum is about 3800 — to a low-dimensional space that discriminates most effectively among the spectra. This is because it relies on the most important projections along a few new orthogonal axes, the PC or eigenvectors of the data covariance matrix (also known in this case as eigenspectra, ES), that maximize the variance and, hence, minimize information loss. Consider a set of N p-dimensional data vectors representing individual spectra $\vec{x}_i$. Mathematically, the standard PCA decomposes each spectrum as follows [6],

$$\vec{x}_i = \vec{\mu} + \sum_{j=1}^{n} a_{ij} \vec{v}_j \qquad (i = 1, \ldots, N) \ , \qquad (2)$$

where $\vec{\mu}$ is the mean spectrum of the set, and $\vec{v}_j$ and a_{ij} are, respectively, the first M PC and their corresponding eigencoefficients, i.e. the projections of $\vec{x}_i$ on $\vec{v}_j$.

Next, we explain the strategies adopted both to determine the minimal optimal subspace for the classification of the spectral types and to identify the regions that these spectra occupy in such low-dimensional space.

A. Determining the optimal subspace

To determine the optimal classification subspace of the stellar spectra, we started with a training sample made up of $10,000$ random spectra with $S/N \geq 20$ of each one of the types from A to M. We extracted from this dataset a subset of 5000 spectra, 1000 per type, consisting of the spectra with the highest quality. This led us to select spectra with $S/N > 50$ for F-, G-, and K-type stars and with $S/N > 40$ for stars of the A and

[1] According to the SDSS spectral mask criteria at `http://www.sdss.org/dr12/algorithms/bitmasks/\#SPPIXMASK`.

[2] Machine Learning and Data Mining for Astronomy, `http://www.astroml.org`

M types (there were not enough spectra with $S/N > 50$ for these two types). These more stringent values of the S/N were adopted to guarantee that the different dimensions inferred from the PCA decomposition were entirely physically motivated and not affected by noisy data. Besides, the higher the S/N of the spectra the larger the amount of variance explained by the different PC, i.e. the more optimal the dimensionality reduction, allowing us to obtain the smallest possible subspace (as long as we are still dealing with a subset representative of the whole population).

Although the amount of explained variance with three PC already achieved what is usually considered an ideal target for this sort of analysis (99%) we decided to work with the subspace of the first four PC since we realized that this extra component facilitated the differentiation between F- and G-type stars, the two spectral types that show the most important overlap in their PC values. The medium spectrum and first four ES (PC) can be seen in Fig. 1.

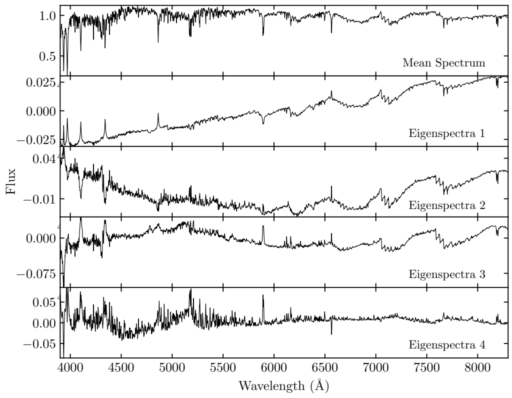

FIG. 1: Medium spectrum and first four eigenspectrum inferred from our highest-S/N sample.

B. Defining the classification regions

We used the whole training sample to define the regions occupied by the different spectral types in the 4D classification subspace just inferred. The $50,000$ spectra of that sample were projected along the directions delineated by the first four ES. The coefficients of these projections are shown in Fig. 2 separately for each spectral type. We can see that each spectral type essentially populates a different region of the classification space, although adjacent types tend to show a certain degree of overlap (somewhat more pronounced for the types F and G). It can also be seen that M-type stars are the ones encompassing the largest and most disperse region of the subspace. As stated before, this is a consequence of the fact that the bluest part of their optical spectrum is crowded with deep absorption bands, features that are responsible for a great deal of the fractional variance. Both circumstances will play a significant role when evaluating the performance of our classification method (see next Section).

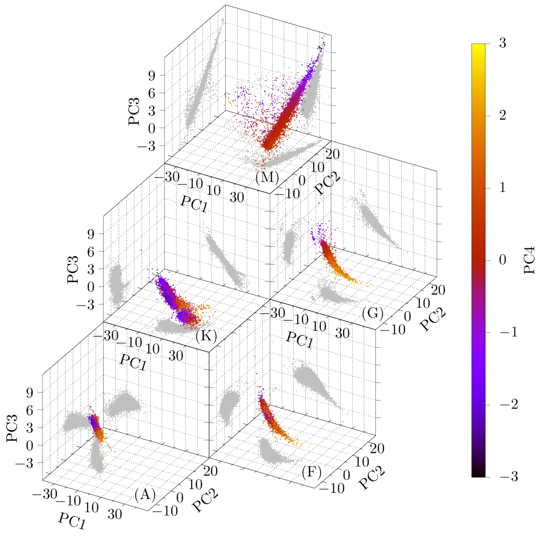

FIG. 2: 4D CR for the A, F, G, K and M spectral types defined by the PC. The first three PC are shown as a 3D point plot, while the fourth PC is shown as a color gradient.

IV. AUTOMATED CLASSIFICATION OF SPECTRA

We now proceed to automatically classify the spectra of three different test samples of stars with increasing S/N ($> 20, 30$ and 40, respectively) randomly chosen from the SDSS-DR15. The classification technique relies on the CR defined in the previous section via the PCA. The tests samples contain 5,000 stars of each spectral type (according to the SDSS spectral pipeline), except in the case of M-type stars where we simply take all the objects (always $< 5,000$) that satisfy our quality constraints.

The classification procedure is as follows. We first project each individual test spectrum on the first four PC that define the optimal classification subspace; this reduces each test spectrum to a 4D vector in that subspace $\vec{x}_i = (a_{i1}, a_{i2}, a_{i3}, a_{i4})$. Next, we define a (small) spherical 4D region of radius R surrounding the $\vec{x}_i$ and use the relative abundances of points in the training sample that fall within this sphere to determine the most probable spectral type of the test point. The probability that a test spectrum is of a certain type k, with $k = A, F, G, K,$ and M, is given by $P_k = W_k/W_{\rm tot}$, where W_k is the *weighted* number of reference spectra of type k in the region around the evaluation point and $W_{\rm tot}$ is the *weighted* total number of such spectra in that region. This non-parametric way of estimating probabilities around a n-dimensional point using the spatial distribution of a finite set of datapoints of reference in a surrounding region is based on the concept of *kernel density estimation.*

A. Top-Hat versus Gaussian kernels

Mathematically, a kernel is a positive function $K(\vec{y}; R)$ which is controlled by the bandwidth parameter R. The density estimate at an arbitrary point $\vec{x}_i$ within a group of M points y_j is given by:

$$W_k(\vec{x}_i) = \sum_{j=1}^{M} K(\vec{x}_i - \vec{y}_j; R) \qquad i = 1, \ldots, N \; . \tag{3}$$

The bandwidth acts as a smoothing parameter, controlling the tradeoff between bias and variance: a large bandwidth leads to a very smooth (i.e. high-bias) density estimation, while a small bandwidth leads to a highly non-uniform (i.e. high-variance) density estimation.

In the present work, we compare the outcomes that result form using the Top-Hat kernel (THK), whose form is

$$K(\vec{x}; R) \propto 1 \;\text{ if } x < R \text{ and } 0 \text{ otherwise} \; , \tag{4}$$

and the Gaussian kernel (GSK) :

$$K(\vec{x}; R) \propto \exp\left(-\frac{x^2}{2R^2}\right) \; , \tag{5}$$

with x the modulus of $\vec{x}$.

After exploring a broad range of bandwidth values, we have found that the optimal value for the THK is $R = 0.5$. The main limitation of this kernel is the equal weighting it assigns to all the points inside the region where one attempts to calculate the probabilities. This is not optimal: the points nearest to a test point are more likely to have a similar spectrum than the farthest ones. An additional weakness of this kernel is the sharp reduction of its performance that we have detected for radial distances $R < 0.2$, a logical consequence of the fact that for small integration volumes the number of classification datapoints can easily go to zero. This became specially evident when attempting to classify stars with an M-type spectrum, because these objects occupy a rather disperse region in the subspace of classification. For its part, the best bandwidth for the GSK is $R = 0.25$. In this case, the integration volume is extended up to 10σ to ensure that the calculations can be performed even for tiny values of the bandwidth.

B. Discussion of the results

The performance of our automated classifier has been evaluated by taking the SDSS classification as reference. This means that we have neglected errors arising from possible misclassifications made by the SDSS pipeline.

To determine the goodness of the classification we have considered three different types of results: 'right', 'wrong', and 'unclassified'. A classified spectrum is labeled as 'right' if our method returns for that spectrum a $P_k > 0.5$ that it has a spectral type k consistent with that of the SDSS. Conversely, if we obtain a $P_k > 0.5$ but for a spectral type k that does not match the SDSS classification, we flag the spectrum as 'wrong'. Finally, if none of the inferred P_k values reach the fifty percent mark or if there are no classification spectra surrounding the spectrum, we consider it as 'unclassified'. The numbers of stars that fall into each category for the three test datasets are presented in Table II segregated by SDSS spectral type. We have included two parameters that provide a better indication of the performance of our procedure. On one hand, we have the reliability, defined as $R = N_k^*/N_k$, which compares N_k, the number of the spectra classified as of type k, with N_k^*, the amount of them that are well classified. On the other hand, we have the completeness, defined as $C = N_k^*/N_{k0}$, which provides the fraction of spectra of type k in the sample that have been correctly classified. Although the differences are small, we find that for both kernels the completeness improves as the quality of the test samples increases, with a slight advantage, as expected, for the GSK. However, the reliability presents a more erratic behavior and does not seem much affected either by the S/N of the sample or the type of kernel employed. This means that an increase in the number of classified spectra does not necessarily lead to an improvement in the goodness of the classification.

The analysis of the results in terms of the spectral types shows that the reliability surpasses 0.9 for all types but G, whereas the completeness is very high (always $\gtrsim 0.95$) for the A, G, and K types, quite good for M-type stars, especially with the GSK, and somewhat less satisfactory ($C \lesssim 0.8$) for stars of the F type. The reason for this reduced performance seems to be the significant overlap shown by the F and G CR (see Fig. 2), which results in a significant fraction of F-type stars being erroneously identified as type G objects, giving rise to the low reliability inferred for this last spectral type. For their part, M-type stars have the largest and most dispersed classification region, which explains the superior performance of the GSK for all subsets regarding the completeness. Besides, they also exhibit the highest reliability due to the fact that they occupy a region in the classification space that is well separated from the others, preventing non-M-type stars from being classified into such category.

V. CONCLUSIONS

We have analyzed the performance of a PCA-based automated method for the classification of stellar spectra with good signal-to-noise ratio into the principal MK system types. We used as a benchmark for this task the dataset of optical spectra of Milky Way stars classified through the SDSS-DR15 spectral pipeline. After a preliminary evaluation of the abundances and general quality of the SDSS spectra that have led us to remove the O and B types from the present study, we implemented for the rest of the MK system a classification

S/N	**Type**	**Right**	**Wrong**	**Unclassified**	**Reliability**	**Completeness**	**Right**	**Wrong**	**Unclassified**	**Reliability**	**Completeness**
		Top-Hat Kernel					**Gaussian Kernel**				
> 20	**A**	4697	110	193	0.993	0.939	4862	130	8	0.919	0.972
	F	3825	1073	102	0.918	0.765	3827	1163	10	0.925	0.765
	G	4768	153	79	0.800	0.954	4816	181	3	0.880	0.963
	K	4672	85	243	0.982	0.934	4889	96	15	0.979	0.978
	M	4317	8	675	0.999	0.863	4830	26	144	0.999	0.966
> 30	**A**	4784	75	141	0.920	0.957	4904	90	6	0.920	0.981
	F	3854	1087	59	0.933	0.771	3846	1146	8	0.925	0.769
	G	4825	117	58	0.892	0.965	4862	137	1	0.885	0.972
	K	4869	84	47	0.982	0.974	4912	86	2	0.975	0.982
	M	2094	2	238	1.000	0.897	2281	9	54	0.999	0.977
> 40	**A**	4856	42	102	0.946	0.971	4950	47	3	0.946	0.990
	F	3933	1031	36	0.939	0.787	3937	1054	9	0.943	0.787
	G	4819	143	38	0.880	0.964	4873	126	1	0.879	0.975
	K	4908	71	21	0.979	0.982	4933	67	0	0.979	0.987
	M	903	0	89	1.000	0.910	976	1	24	0.999	0.975

TABLE II: Results, segregated per MK spectral type, of our PCA-based spectral classification method using Top-Hat (left) and Gaussian (right) kernels for subsets of test spectra with $S/N > 20, 30,$ and 40 (see the text for more details).

procedure that is based on the use of the PCA. We have found that it is possible to account for 99% of the total variance arising from good-quality A, F, G, K, and M stellar spectra with only three principal components. Nevertheless, we decided to increase the number of PC to four as we have found that this extra component helps to discriminate between the F and G types (see below), raising the explained variance to 99.3%. The projections of the highest-quality spectra on this 4D space have been used to determine the CR that are associated with the different spectral types.

To evaluate the efficiency of our automated classifier, we have performed a statistical study using a series of test samples containing spectra with different thresholds of S/N. Specifically, we defined three samples with $\sim 25,000$ spectra each and $S/N > 20$, 30, and 40, respectively. Next, we have determined the most probable spectral type of each test spectrum from the relative densities of reference spectra of different type that surround its projection on the 4D classification space of PC defined previously. Two kernels have been used for density estimation: a basic Top-Hat Kernel and a Gaussian kernel. The quality of the results for the different levels of S/N and the different spectral types has been quantified by means of two statistical parameters, completeness and reliability, which we have calculated taking the SDSS spectral types as reference. Our findings indicate that both kernels produce similarly good results when implemented with their optimal bandwidths, though, as expected, the Gaussian density estimator performs somewhat better. For the two kernels the completeness of the outcomes increases as the S/N of the test samples increases, but the reliability does not seem to depend much on the quality of the sample or the kernel type. This leads us to conclude that an increase in the number of classified spectra does not necessarily lead to an improvement in the goodness of the classification. As regards to the spectral types, we generally obtain values of both completeness and reliability well above 90%. The exception are F-type stars whose completeness is always less than 0.8. We attribute this lower performance to the significant overlap shown by the F and G CR, which results in a large fraction of F-type stars being erroneously classified as type G, which in turn reduces the reliability for this last spectral type.

Further work should be done by testing our methodology on other catalogs of optical spectra to confirm the consistency of our results.

Acknowledgments

I would like to thank J.M. Solanes and J.L. Tous for the excellent guidance and support during this exceptional pandemic situation. Special thanks to A. Romero for constructive criticism and help on plotting issues and his unconditional support.

[1] Morgan W.W.; Keenan P.C.; Kellman E. 1943, *An Atlas of Stellar Spectra, with an Outline of Spectral Classification*, University of Chicago Press, Chicago

[2] Bailer-Jones Coryn, A.L.; Irwin, M.; von Hippel, T. 1998, MNRAS, **298**, 361

[3] Sánchez Almeida, J.; Allende Prieto, C. 2013, AJ, **763**, 50

[4] Tous, J.L.; Solanes, J.M.; Perea, J.D. 2020, MNRAS, in press, arXiv:2005.09016

[5] Yip, C.W., et al. 2004, AJ, **128**, 585

[6] Ivezić, Z.; Connolly, A.J.; VanderPlas, J. T.; Gray, A. 2014, *Statistics, Data Mining, and Machine Learning in Astronomy*, Princeton, New Jersey