The Analysis and Processing of Spectra with Splicing Abnormality in LAMOST Based on High Performance Computing Platform

Fanlong Meng, Jingchang Pan* and Jingjing Yu
School of Mechanical, Electrical & Information Engineering, Shandong University at Weihai, Weihai, China
Email: morphy8@163.com, *pjc@sdu.edu.cn, yujingjing2018@163.com

Abstract. Since the completion of LAMOST the observation had produced a large number of data, and officially released data. There are some poor quality spectra in the observation data and even the data released to the outside world. Splicing abnormal spectra is one kind of them. Splicing abnormality is poor continuity spectrum showed in the splicing wavelengths of the red and blue end. This problem can be caused by several factors, such as stability of instrument, observation condition, the response function, and so on. It has important effect on the research and analysis of spectra. It is of great significance to study the automatic identification of the splicing abnormal spectra. The method in this paper we first do the spectrum pretreatment, then the spectral was intercepted into red and blue in order to fit curves. For the two fitting curves, a series of wavelength points are selected to calculate the difference of these points. For the difference set, the statistical characteristics of the difference are analyzed, such as mean and standard deviation. Then we calculate the area of the two curves. Based on these values, we proposed an evaluation function. At the same time, we provided a method for detecting the splicing abnormal spectra based on the high performance computing platform. Through a lot of experiments, the efficiency was greatly improved compared to stand-alone.

1. Introduction
The Large Sky Area Multi-Object Fiber Spectroscopic Telescope [1][2][3] (LAMOST) had been established by Oct, 2008. Since its completion, LAMOST captured a large number of spectral data. In these data, there are still some low quality spectra such as splicing abnormal spectra. With the LAMOST survey going on, thousands of spectra had been captured at each observation night. In the face of such a large amount of data, there is no efficient tool for detecting abnormal splicing spectra and it leads to spectral studies cannot be carried out smoothly. Therefore, it is of great significance to find a method to detect the abnormal spectrum automatically.

In the process of LAMOST survey, the star is divided into red and blue end and captured separately. After observation, the red end data and blue end data are merge together. Because of stability of instrument, observation condition, the response function, some splicing spectra become abnormal. This abnormality in the spectrum is in the splicing region (5600Å -6000Å) spectrum shows flux jumping (Figure.1 (a)), fault (Figure.1 (b)), and trend anomaly (Figure.1 (c)). At present, there is little research on the automatic detection method of LAMOST spectral splicing anomaly. In this paper, an evaluation function to define the spectrum abnormal degree is proposed based on the characteristics of splicing abnormal spectrum. This study not only helps to enhance the detection of abnormal splicing speed spectrum, also contribute to the process of quantification and quality improvement, also it may
encourage further research, for example, based on the evaluation we can explore solutions to eliminate anomalies.

![Figure 1](spec-56233-EG042017S013757V01_sp16-001.fits)
(a)  ![Figure 1](spec-56233-EG01260689021203B03_sp07-174.fits)
(b)  ![Figure 1](spec-56256-VB006N29V2_sp15-109.fits)
(c)

**Figure 1** Abnormal splicing spectrum (a) this is the flux jumping. (b) this is the flux fault. (c) this is the trend anomaly

### 2. Method Detail
The splicing wavelength region is 5600Å-6000Å. After the capture, we mark the overlapped region as the baseline, keep one end motionless, make the other end enlarge or reduce a certain number of times to let the flux coincided in the overlapped wavelength region. Through this, we get the entire curve of full wavelength. Outside the overlapped region, the relative flux is same with origin spectrum and it becomes abnormal under the both effect of red and blue end flux.

In this paper, we propose a piecewise fitting method to analyze the spectrum. Firstly, we intercept the wavelength 5200 Å-6000Å as the blue end, 5600Å-6400Å as the red end. Then, fit the two end separately, so we get two curves. In figure.2, the subgraph upper is the fitting curve of origin spectrum. In this figure, the black line is the origin line, the blue solid line is the fitting curve of blue end and the red solid line is the fitting curve of red end. The two red dotted is the position of 5600Å and 6000Å. The subgraph below is the zoom in detail of the splicing region.

The method of this paper is based on the following principles. The two fitting curves are closely approximate if the spectrum is normal and the flux difference is small. The flux differences distribution is uniform.

In the splicing region, we select a series of points with equal wavelength intervals on the two fitting curves. Then calculate the flux difference of these points. For these differences, we get its mean value
and standard deviation and calculate the area surrounded by two lines. If the mean, standard deviation and the area are under the thresholds, we mark it as normal spectrum.

The thresholds of the above three statistical moments (characteristics value) can be got by the analysis of the template spectra. In this paper, the template spectra we use is proposed by Peng Wei in the paper “On the Construction of a New Stellar Classification Template Library for the LAMOST Spectral Analysis Pipeline” \(^{(4)}\) for LAMOST. Based on these thresholds, we put forward an evaluation function to define the abnormality degree for the splicing abnormal spectrum.

2.1 Get the Thresholds by Subsection Simulation

In this paper, we define the spectrum abnormal or not by subsection simulation method based on thresholds value, so the thresholds are fundamental. We got the thresholds by analyzing the template spectra because of the template spectra are good splicing spectra. The method to obtain the thresholds is:

The first step, we do the pretreatment, such as flux normalization, sodium line removal. Intercept the flux line of wavelength 5200 Å-6000Å as the blue end (mark the wavelength as bwave, the flux as bflux) , and 5600Å-6400Å as the red end (mark the wavelength as rwave, the flux as rflux).

The second step, we do six order polynomial fitting for the flux line of blue and red end and get red end fitting curve (rp) and the blue end fitting curve (bp).

The third step, for the two fitting curves rp and bp, calculate the flux difference by selecting a series of evenly spaced wavelength points and get flux difference array called df.

The fourth step, for the array df, we calculate its mean value (mark it as MEAN) and standard deviation (mark it as STD).

The fifth step, for the two fitting curves rp and bp, calculate the area surrounded by two lines in the splicing region (mark it as AREA).

![Subsection simulation curve](image)

**Table 1 Thresholds of eigenvalues (Note: unit: 10^-4)**

| TYPE | MEAN | STD  | AREA  |
|------|------|------|-------|
| O    | 0.86 | 1.18 | 325   |
| B    | 0.08 | 0.1  | 30    |
| A    | 0.15 | 0.2  | 80    |
| F    | 0.17 | 0.19 | 78    |
| G    | 0.25 | 0.3  | 100   |
| K    | 3.7  | 4.5  | 1400  |
| M    | 7.5  | 9    | 2900  |
In the data analysis process, we use 183 template spectra, it contains star type, such as O, B, A, F, G, K, M, double star and carbon star. We only analyze these kinds such as O, B, A, F, G, K, M and ignore other types. We deal with the rest 176 template spectra with the five steps above, so we get 176 groups of characteristic values (it can be expressed as three tuple (MEAN, STD, AREA)). With these statistics, the 176 groups’ data can be divided into 7 groups according to the star type. Each group takes the maximum value of the corresponding statistic as the threshold of this type. Through this, we get the thresholds of the MEAN value, standard deviation and the area. The thresholds in our paper is presented in table.1:

2.2 Remove Spectra of Zero Flux
In the splicing abnormal spectrum, there is a kind of abnormal spectrum of faults, as shown in Figure 1 (b), which the flux is all zero of continuous wavelength points. In this paper, firstly, we judge whether the spectrum exists this fault or not, if yes, we define it as splicing abnormal spectrum.

The specific methods are as follows.
(1) Define a window of “w” points, we set w 10 in this paper.
(2) In the splicing region, we traverse the flux values of all the wavelength. If the flux is zero at one point and all flux are zero in the window of w points from this point, we define this spectrum abnormal.

2.3 Anomaly classification
We use the same subsection simulation method to deal with the test spectrum as the template spectrum. Also, we get the mean value (α’), standard deviation (δ’) and the area surround by the fitting curve (ε’). The method to judge the spectrum is as follows.

The first step, get the star type.

The second step, find the characteristics values according to the star type. The mean value, standard deviation and area threshold are α, δ, ε.

The third step, we define an evaluation function. $p = \sum_i p_i \times \lambda_i$, where $\lambda_i = \frac{\alpha'}{\alpha}, \lambda_2 = \frac{\delta'}{\delta}, \lambda_3 = \frac{\varepsilon'}{\varepsilon}$.

$p_i$ is the weight of $\lambda_i$. In this paper, $p_i (i=1, 2, 3)$ are 0.4, 0.1 and 0.5 respectively. In the three statistics, the area is the most direct response to the abnormal situation, the weight is set to 0.5, the mean response to the flow of the overall situation, the degree of response to the abnormal situation is also large, set its weight is 0.4, the standard deviation of the difference shows the overall floating situation, the reference is slightly smaller, set the weight of 0.1. Considering these three factors, the weighted average value is used to quantify the degree of splicing abnormal spectra.

The fourth step, judge the degree of splicing abnormal spectra based on the p value calculated by the third step.

Experiments show that the design of the three weights and the evaluation function can be used to judge whether the spectrum is abnormal or not. The value of the evaluation function reflects the degree of abnormal spectrum, and the experimental results show that the evaluation function of the normal spectrum is below 0.9. Through lots of experiments, the abnormal degree is divided into three grades. For the abnormal spectrum, we define the evaluation function value of cut-off point was 5 and 10, namely: $p<=0.9$ is normal spectrum, $0.9<p<=5$ is defined as first level, $5<p<=10$ is defined as second level, $p>10$ is defined as third level. The two level anomaly and the boundary point of the three level anomaly can be adjusted manually according to different needs. The demarcation point between the second and third level can be adjusted manually. Some of the test spectral types are unknown (i.e. Non), and the threshold value is set as the mean of the remaining spectral types.

3. The Detecting Method of Splicing Abnormal Spectrum based on High Performance Computing Platform
The high performance computing cluster system of our school uses CPU+GPU+MIC architecture, which is composed of 34 computing nodes and the 1 fat nodes. In addition, there are two graphics
processor nodes and two Intel integrated core nodes. The cluster has total of 796 CPU computing core. With the powerful computing capabilities of the platform, we can process fits files rapidly.

In this paper, we use python to invoke mpi method, divide and conquer, a number of fits files are distributed to multiple processes. Each process is executed independently and does not affect each other. The root process gather the result when all processes finish and it will produce standard information stream and standard error stream.

4. Experimental Environment on High Performance Platform

High performance computing platform system uses Red Hat Enterprise Linux Server version of release 6.5. Due to the high performance computing platform does not configure parallel environment based on python, so users need to build their own run environment. The high performance platform has strict permissions, ordinary users can only operate in the current user directory. The software packages can only be manually installed and configured environment variable, in order to build the environment suitable for the requirements of this experiment, in the current user directory, we installed or upgraded some software such as python2.7/jdk1.7/astropy/numpy/scipy/mpi4py/glibc/gmp/mpfr/mpc/gfortran/gcc/jpype.

Python2.7. High performance computing platform comes with a version of Python is 2.6, in order to keep the same with the single experimental environment, we upgrade python from 2.6 to 2.7. Download python2.7 package, unzip and enter the compressed directory, installed by the python setup.py command, due to the current user for ordinary users, it cannot be installed in the default directory, through optional "order prefix" python will be installed below the current user’s directory. After installation, when we use python, the system will default to find the python executable file from the system directory and then run it, and did not run the newly installed python, so we need to configure the environment variables. There are two environmental information file ".Bash_profile" and ".Bashrc" under the root directory of current user. The .Bash_profile will be run when a user logs in, then it will call the.Bashrc. We modify the .Bashrc file, configure PYTHONPATH, and add it to the PATH variable. In the search for a python executable file, system will find it in PATH, we put PYTHONPATH before PATH to ensure the system calls the new version of python.

Numpy. Similar to the installation of python, we need to install it in the local directory. We can install the package like Python by downloading the compressed package and do manual installation. We can also use pip to install it. Pip is a python package management tool, it can be used to automatically install the software on PyPI. When using pip, we note that the default software source installation efficiency is low, we can specify the installation source through the "-i" command, such as Tsinghua University, University of Science & Technology China, and Douban. You can specify the installation package version. Pip greatly simplifies the installation process. We will install the package in the current python directory under the site-packages directory, where the installation of third packages put. Then it can be found and imported by the program.

For the software installation, we can configure the environment variables, also we can create a *.pth file in the site-packages directory, when python reads "sys.path", the PTH files will be read to load the configuration directory configured in this file.

There is no need to enumerate all the installation here, such as mpi4py, matplotlib, astropy and so on. We will install it in a separate directory, and configure the directory in the PTH file.

For the normal software, we can install it by the three steps, configure (configure), compile (make), install (make install) and then configure environment variable.

5. Analysis of Experimental Results

In this paper, the 4426 splicing abnormal spectra of LAMOST DR1 is used as the test spectra.

About 93% of all splicing abnormal spectra can be found out by this method. The zero flow spectra 1591, first level 1524, second level 608, third level 368. As shown in Figure 3, 4, 5. The scores calculated by the evaluation function are shown in table.2.
Figure 3 Spectrum judged to be normal splicing

Figure 4 Abnormal splicing spectrum in level I

Figure 5 Abnormal splicing spectrum in level II

Figure 6 Abnormal splicing spectrum in level III

Figure 7 The time change on HPC
### Table 2 The score of example spectra

| Figure score | Figure 3 | Figure 4 | Figure 5 | Figure 6 |
|--------------|----------|----------|----------|----------|
| score        | 0.78     | 2.47     | 6.24     | 37.86    |
| result       | normal   | first    | second   | third    |

The experiment analysis is based on the 4426 spectral data. We submit jobs of the same files to the high performance computing platform by specifying different number of CPU core. The run time of each experiment is as follows in Figure 7. The run time on stand-alone is 1471 seconds. It can be seen from the experimental results that the computational efficiency in the high performance environment is obviously higher than that in the stand-alone environment. With the number of CPU core increased, the processing time is shorter, and the efficiency is more and more high. The running time is not linear change. After the core number 80, the running time does not change, the reason is that too many core need a lot of inter process communication and file IO, limit the time shortened.

6. Conclusions
In this paper, a method for automatic detection and classification of splicing abnormal spectra in LAMOST spectra is presented. Firstly the test spectra were pretreated and divided into red and blue ends, then we fitted the flux of red and blue ends. For two fitting curves, we select a series of wavelength interval, calculate the flux difference between these points. After that, we analyze the differences set to get its mean value, standard deviation and the area. In this paper, we propose an evaluation function to deal with the mean, standard deviation and area to determine the degree of abnormal spectrum. A large number of experiments show that the method has a good effect on the detection of splicing abnormal spectra.

7. Acknowledgment
This work was supported by the National Natural Science Foundation of China (U1431102).

8. References
[1] Cui X-Q, Zhao Y-H, Chu Y-Q, et al. The large sky area multi-object fiber spectroscopic telescope (LAMOST)[J]. RAA, 2012, 12(9): 1197.
[2] Gang Zhao, et al. LAMOST spectral survey — An overview. Research in Astronomy and Astrophysics, 12(7):723-734, 2012.
[3] Luo, A., et al. Data release of the LAMOST pilot survey. Research in Astronomy and Astrophysics, 12(9): 1243-1246, 2012.
[4] Wei P, Luo A, Li Y, et al. On the Construction of a New Stellar Classification Template Library for the LAMOST Spectral Analysis Pipeline[J]. Journal of Telemedicine & Telecare, 2014,147(147):S129-32.
[5] Donald C. Wells, et al. FITS: A Flexible Images Transport System. Astronomy and Astrophysics Supplement, 44:363, 1981.
[6] Python. [https://www.python.org/](https://www.python.org/)
[7] OpenMPI. [https://www.open-mpi.org/](https://www.open-mpi.org/)
[8] Ari, Niyazi,Ustazhanov, Makhamadsulton. Matplotlib in python[C]//2014 11th International Conference on Electronics, Computer and Computation: 2014 11th International Conference on Electronics, Computer and Computation (ICECCO 14), Sept. 29 2014-Oct. 1, 2014, Abuja, Nigeria.2014:1-6.
[9] Thomas P. Robitaille,Erik J. Tollerud,Perry Greenfield et al. Astropy: A community Python package for astronomy[J].Astronomy and astrophysics,2013,558(Pt.1):A33-1-A33-1.
[10] LAMOST DR1. [http://dr1.lamost.org/](http://dr1.lamost.org/)