Application of R to LAMOST Spectral Analysis: A Preliminary Study (Postprint)

Authors: Chen Shuxin, Luo Ali, Sun Weimin

Date: 2017-09-26T00:00:00+00:00

Abstract

Employing the highly extensible open-source R programming language as the tool and leveraging its powerful data analysis capabilities in statistics and data mining, this study focuses on investigating the primary functions and characteristics of the RFITSIO package for reading and writing FITS format files in R, and provides a detailed introduction to FITS files collected by LOMAST. The massive LOMAST survey spectroscopic DR2 data is utilized to extract stellar spectra via RFITSIO, and R' s principal component analysis tools are employed to extract characteristic quantities—namely principal components—from various types of spectral data. By extracting principal components that represent stellar spectral characteristics from spectra containing substantial redundant information, the spectral features obtained through principal component analysis can effectively mitigate the impact of noise on original spectral data after reconstruction, thereby providing a research foundation for subsequent data mining endeavors.

Full Text

Preamble

Astronomical Research and Technology, Vol. 14 No. 3, July 2017
ISSN 1672-7673

A Preliminary Study on Applying R Language to LAMOST Spectral Analysis

Chen Shuxin[1], Luo Ali[3], Sun Weimin[2]
[1]School of Mechanical and Electrical Engineering, Qiqihar University, Qiqihar, Heilongjiang
[2]Key Laboratory of In-Fiber Integrated Optics, Ministry of Education, College of Science, Harbin Engineering University, Harbin, Heilongjiang
[3]Key Laboratory of Optical Astronomy, Chinese Academy of Sciences, Beijing

Abstract

This study explores the application of the highly extensible, open-source R programming language as a tool for astronomical data analysis, leveraging its powerful capabilities in statistics and data mining. We focus on investigating the primary functions and features of the RFITSIO package for handling LAMOST stellar spectra, utilizing R' s principal component analysis (PCA) tools to extract characteristic features from various types of spectral data. By extracting principal components that represent the essential features of stellar spectra from data containing substantial redundant information, PCA reconstruction effectively reduces noise impact on original spectral data while preserving key spectral characteristics. This approach provides a solid foundation for subsequent data mining research. The RFITSIO package enables detailed processing of massive FITS format files, offering a simple and efficient solution for reading and writing astronomical data.

Keywords: R language; RFITSIO; Principal Component Analysis; LAMOST; Spectroscopic Survey

1. Introduction

R is an open-source software programming language that integrates numerous data analysis and visualization methods. It has become an essential tool for big data analysis in the information age, offering exceptional extensibility and powerful statistical computing capabilities that can effectively simplify data analysis workflows. To apply R' s advantages to astronomical data analysis, one must first understand the standard formats used in astronomy.

The Flexible Image Transport System (FITS) has become the most widely used data format in astronomy for describing both the definition and encoding of astronomical data [**?**]. First proposed in 1981, this standard enables storage of images, data tables, and metadata in a single file structure. The FITSIO software library, developed by NASA' s High Energy Astrophysics Science Archive Research Center, provides convenient and powerful functions for reading and writing FITS files that can be called directly from Fortran, IDL, Python, and other programming languages [**?**]. For the R environment, Andrew Harris developed the RFITSIO package, which can be downloaded from the Comprehensive R Archive Network (CRAN) mirror sites.

2. RFITSIO Package Overview

The RFITSIO package provides R users with straightforward functions to read and write FITS format files, which differ from common image formats like GIF and JPG. FITS serves as a standard data format that simultaneously stores image data and tabular data. The package includes functions such as `readFITS()` and `readFrameFromFITS()` that automatically identify and process image files, ASCII tables, and binary table extensions.

The `readFITS()` function primarily accepts file parameters and returns a list containing header and data unit information. Key return values include: `header` (header file parameters), `imDat` (image data), `axDat` (axis scaling and labels), `colNames` (column name vectors), `colUnits` (column unit vectors), `TNULLn` (undefined value definitions), `TSCALn` (scaling factors), and `TDISPn` (format information vectors). These functions enable seamless integration of FITS data into R' s data frame structures for subsequent statistical analysis.

3. R Language for Spectral Data Analysis

R' s powerful statistical functionality supports various analytical methods from regression and classification to clustering and exploratory data analysis. Astronomical statistical methods have evolved from traditional approaches to modern data mining techniques. Our experimental analysis is based on the R programming platform, using LAMOST DR2 data classified according to the released stellar classifications.

The LAMOST (Large Sky Area Multi-Object Fiber Spectroscopic Telescope) survey has released massive spectral datasets. The file naming convention follows the format: `spec-PLANID-YYYY_spXX-FFF.fits`, where PLANID represents the plan identifier, YYYY indicates the Modified Julian Date, spXX denotes the spectrograph number, and FFF represents the fiber ID. Each FITS file contains a primary header unit with mandatory keywords, telescope parameters, data reduction parameters, and spectral analysis results.

4. Principal Component Analysis of LAMOST Spectra

4.1 Methodology

When processing high-dimensional spectral data, dimensionality reduction techniques are essential for projecting high-dimensional feature information into a lower-dimensional space while preserving the most significant characteristics. Principal Component Analysis (PCA) is a multivariate linear method that effectively handles large-sample, multi-parameter quantitative data analysis in an unsupervised manner.

By selecting principal components with the largest variances (typically achieving a cumulative variance contribution rate of 80%-90%), PCA reduces problem complexity while minimizing information loss. The method eliminates correlations between flux features through orthogonal transformation, achieving both dimensionality reduction and noise removal.

4.2 PCA Implementation in R

The `prcomp()` function in R' s `stats` package efficiently performs PCA on spectral data. To analyze LAMOST spectra, we first load the RFITSIO package and read the FITS files:

```
require(FITSio)
M_STAR <- readFITS("spec-55938-GAC_070N40_V1_sp04-161.fits")
```

The `summary(prcomp())` function provides essential information about each principal component, including standard deviation, proportion of variance, and cumulative proportion. For an A1IV type stellar spectrum, the first seven principal components achieve a total variance contribution of 99.796%, demonstrating that most spectral information can be preserved in a dramatically reduced dimensionality space.

4.3 Spectral Reconstruction

Using only the first principal component for reconstruction yields a spectrum where noise amplitude is significantly reduced while major spectral features remain intact. Figure 4 shows the reconstructed A1IV type spectrum compared to the original, confirming that PCA-based feature extraction is highly effective for preserving spectral characteristics while suppressing noise.

5. Conclusion and Outlook

With the growing collaboration between astronomy and cloud computing in big data research [**?**], powerful analytical tools have become essential for exploring the universe. The R language platform and its extensive statistical packages demonstrate significant performance advantages in massive data processing and mining.

This preliminary study demonstrates that applying R to astronomical data mining—particularly through RFITSIO for data I/O and PCA for feature extraction—efficiently captures the primary characteristics of high-dimensional spectral data. R' s big data analytics capabilities enable more effective extraction of information from astronomical spectra. This work represents a valuable attempt to apply statistical computing tools to astrophysics. Future research will focus on combining these methods with domain-specific astronomical knowledge to identify multivariate relationships in spectral data and develop more sophisticated data mining approaches for next-generation astronomical surveys.

References

1. Wells, D. C., Greisen, E. W., & Harten, R. H. 1981, *Astronomy and Astrophysics Supplement*, 44, 363
2. Ke, D. R., & Zhao, Y. H. 2004, *New Technology of Library and Information Service*, 25-26
3. Li, H. N., Xiao, Q. B., & Shao, Z. Y. 2005, *Annals of Shanghai Observatory, Academia Sinica*, 119-124
4. Cui, C. Z., Li, W., Yu, C., et al. 2008, *Astronomical Research & Technology*, 116-123

5. Zhong, S. B., Han, B., Zhang, Y. X., et al. 2015, *Astronomical Research & Technology*, 511-515
6. Zhao, Y. H. 2010, *Scientia Sinica: Physica, Mechanica & Astronomica*, 1041-1048
7. Luo, A. L., Zhao, Y. H., et al. 2015, *Research in Astronomy and Astrophysics*, 1095-1124
8. Tu, Y., Zhang, Y. X., Zhao, Y. H., et al. 2012, *Astronomical Research & Technology*, 124-132
9. Pence, W. D. 2016, *Astronomical Data Analysis Software and Systems XXV*, 487
10. Babu, G. J., & Feigelson, E. D. 1996, *Astrostatistics*, Chapman and Hall
11. Cui, C. Z., Yu, C., Xiao, J., et al. 2016, *Chinese Science Bulletin*, 445-449

Note: Figure translations are in progress. See original paper for figures.