VERIFICATION OF STATISTICAL PROPERTIES FOR HYPERSPECTRAL IMAGES: HETEROSCEDASTICITY AND NON-STATIONARITY

Jihao Yin, Chao Gao, and Jianying Sun
School of Astronautics, Beihang University
37 Xueyuan Road, Haidian District, Beijing 100191, China

E-mail: yjh@buaa.edu.cn, chao-gao@sa.buaa.edu.cn, bhsunjy@sa.buaa.edu.cn

ABSTRACT

This paper investigates the heteroscedasticity and non-stationarity, two statistical properties, of hyperspectral remote sensing data. In the field of mathematical sciences, a collection of variables is heteroscedastic if there are sub-populations that have different variances or volatilities than others, while a non-stationary process refers to a stochastic process whose joint probability distribution are changing when shifted in time or space. To be treat as sequences, hyperspectral data are investigated via Bartlett Test and Wald-Wolfowitz Runs Test to verify the heteroscedasticity and non-stationarity, respectively. Most experimental results fail to pass Bartlett Test and Wald-Wolfowitz Runs Rest statistically significant, indicating that both heteroscedasticity and non-stationarity are intrinsic properties of spectral response sequence.

KEYWORDS

Hyperspectral data, Spectral response sequence, Heteroscedasticity, Non-stationarity, P-value

1 INTRODUCTION

Imaging spectrometer is one of fast developing remote sensing tools for the ground observation in the last three decades. Hyperspectral images are widely used in civil and military fields, such as mineral exploration and camouflage detection, indicating a significant value in theoretical research as well as a promising future in real applications. A hyperspectral image cube contains hundreds of bands with fine spectral resolution as well as spatial information. Compared with other traditional remote sensing techniques, the imaging spectrometer is capable to identify indiscernible targets that have a similar appearance but a different composition of substance [1].

The large data contained in the hyperspectral image cubes, however, causes difficulties in data record, storage, transmission and processing. Not all bands effectively contribute to specific tasks, as the contiguous bands are strongly correlated and highly redundant. Therefore, researchers have explored numerous techniques of hyperspectral image processing, focusing on dimensionality reduction, land-covering recognition, spectral unmixing, and multisource fusion [2]. Many well-developed theories, such as chaotic dynamics, wavelet packet analysis, and cointegration theory have been introduced in order to improve the performance of hyperspectral data processing [3-5]. It should be noted that these mentioned methods all fall in the framework of time series analysis.

Time series analysis is a dynamic statistical data processing scheme, which applies probability theory and mathematical statistics to study statistical property in given stochastic data sequences [6]. There is no doubt that time series analysis could have huge potential to deal with hyperspectral image cube. In fact, time series analysis can be implemented to any series or curves, DNA series as example. On this note, it's reasonable to apply relevant methods for spectral response sequences although hyperspectral data are not real time series.

In the field of nonlinear time series analysis, two of the most important statistical properties are heteroscedasticity and non-stationarity [7, 8]. A collection of variables is heteroscedastic if there are sub-populations that have different variances or volatilities than others, while a non-stationary process refers to a stochastic process whose joint probability distribution are changing when shifted in time or space. But limited literatures involve in verifying these two properties of hyperspectral data.

This paper employs Bartlett Test and Wald-Wolfowitz Runs Test to perform the test and verification for the heteroscedasticity and non-stationarity, respectively [9, 10]. The remainder of this paper is organized as follows. Detailed procedures are given in Section II. In Section III, experiments and results are presented, followed by conclusions in Section IV.

2 STATISTICAL HYPOTHESIS TEST

2.1 Bartlett Test

In statistics, Bartlett Test is used to test if samples are from populations with equal variances. The situation of equal variances across samples is called homoscedasticity or homogeneity of variances [9].

The null hypothesis H_0 is that all population variances are equal against the alternative that at least two are different. If there are k samples with size $n_i, i=1,\cdots,k$ and sample variances $S_i^2, i=1,\cdots,k$, then the statistic χ^2 of Bartlett Test is:

$$\chi^2 = \left((N-k)\ln S^2 - \sum_{i=1}^{k}(n_i-1)\ln S_i^2\right)\Big/C, \quad (1)$$

where:

$$\begin{cases} N = \sum_{i=1}^{k} n_i \\ S^2 = \sum_{i=1}^{k}(n_i-1)S_i^2 \big/ (N-k) \end{cases}, \quad (2)$$

and C is correction value:

$$C = 1 + \left(\sum_{i=1}^{k} 1\big/(n_i-1) - 1\big/(N-k)\right)\Big/(3k-3). \quad (3)$$

The statistic χ^2 has approximately a χ^2_{k-1} distribution with degrees of freedom $k-1$. Thus the null hypothesis H_0 is rejected if $\chi^2 > \chi^2_{k-1,\alpha}$, where α, often set as 0.1 or 0.05, is significance level, and $\chi^2_{k-1,\alpha}$ is the upper tail critical value.

Since α need be preset, accepting or rejecting the null hypothesis H_0 may be influenced by subjective factors. This paper uses the p-value instead of statistic χ^2 to measure test result. One rejects the null hypothesis H_0 when the p-value is less than α. In Bartlett Test, the p-value is the probability of χ^2_{k-1} distribution between χ^2 and ∞.

2.2 Wald-Wolfowitz Runs Test

Wald-Wolfowitz Runs Test is a non-parametric statistical test that checks a randomness hypothesis for a two-valued data sequence. A "run" of a sequence is a maximal non-empty segment of the sequence consisting of adjacent equal elements. For example, the sequence "1111000011100011111000" consists of six runs, three of which consist of '1' and the others of '0'. If the number of runs is significantly lower than expected, the hypothesis of statistical stationarity should be rejected [10].

Under the null hypothesis H_0, the number of runs R in a sequence of length N is a random variable, whose conditional distribution is asymptotically normal with mean μ and variance σ^2:

$$\begin{cases} \mu = 1 + 2N^+N^- \big/ N \\ \sigma^2 = (\mu-1)(\mu-2)\big/(N-1) \end{cases}, \quad (4)$$

given the observation of N^+ values larger than median and N^- values less than median and ($N = N^+ + N^-$). The statistic Z of Runs Test is:

$$Z = (R-\mu)\big/\sigma, \quad (5)$$

which has an asymptotically standard norm distribution $N(0,1)$. Thus the null hypothesis H_0

is rejected if $Z < z_{\alpha}$, where z_{α} is the α -percentile of $N(0,1)$.

Again, Wald-Wolfowitz Runs Test also uses the p-value, which is the probability of $N(0,1)$ distribution within range $(-\infty, Z]$.

3 EXPERIMENTAL RESULTS

3.1 Date Sets

In this paper, two real data sets are applied to verification the heteroscedasticity and non-stationarity of hyperspectral image. They are the urban area of Washington D. C. Mall and the mixed forest/agricultural field of Indian Pine [3].

Figure 1. False Color Image of Washington D. C. Mall with Ground Truth Labels.

The data set of the Washington D. C. Mall was collected by the HYDICE on August 23, 1995. The original data set includes 210 bands in the range of 0.4-2.4μm. Those bands from infrared spectrums have more noise than bands from visible wavelengths due to water absorption, which leads to the removal of 19 noisy bands. The candidate region is the subset including 542 rows by 307 columns by 191 bands from the original image. Its false color image with ground truth labels is depicted in Figure 1. In terms of this selected image, there are six types of objects. They are Roof, Street, Path, Grass, Tree and Water.

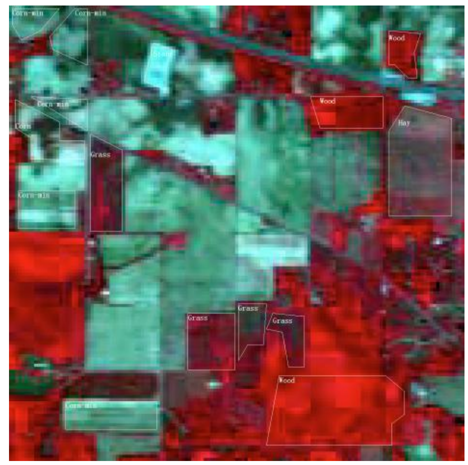

Figure 2. False Color Image of Indian Pine with Ground Truth Labels.

Gathered by the AVIRIS, the Indian Pine hyperspectral image (145 by 145 pixels) has 220 spectral bands with spatial resolution of 20m. Corn-min, Corn, Grass, Hay, Wood and Stone are selected for investigation. Its false color image with ground truth labels is depicted in Figure 2.

3.2 Verification for Heteroscedasticity

For given class, compute the variance of each band. The null hypothesis H_0 of Bartlett Test is that three adjacent successive bands have the equal variances. All possible combinations of three adjacent successive bands are tested by Bartlett Test to get corresponding p-values, and Histograms of p-values for different classes in two datasets are depicted in Figure 3 and Figure 4, respectively.

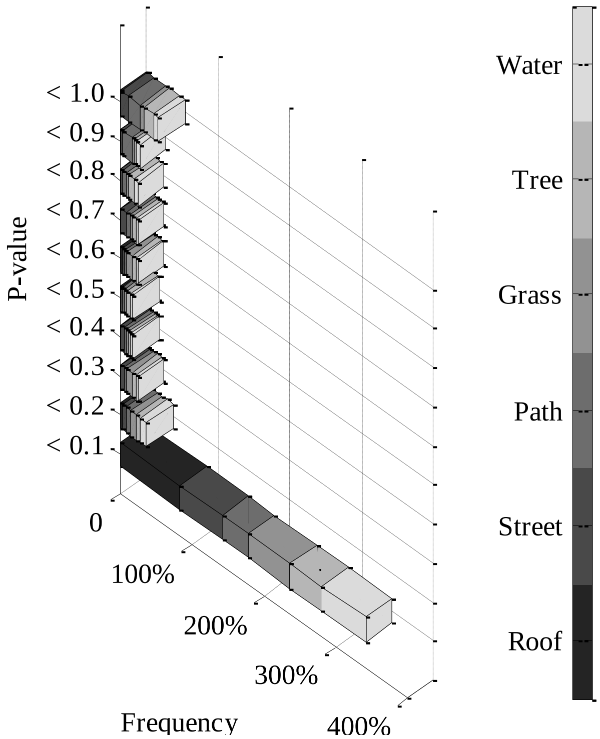

Figure 3. Histograms of P-values for Different Classes under Bartlett Test in Washington D. C. Mall Data Set.

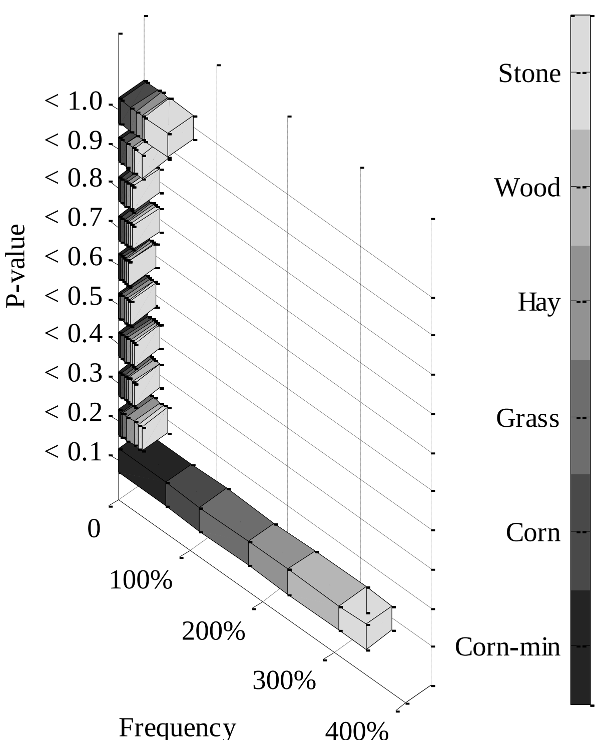

Figure 4. Histograms of P-values for Different Classes under Bartlett Test in Indian Pine Data Set.

For the data set of Washington D. C. Mall, it is clear that the majority of p-values are less than 0.1 for all classes. Most bands fail to keep equal variances with their neighbors in terms of statistics, though hyperspectral images have strong relationship and heavy redundancy within adjacent bands. For the data set of Indian Pine, the results are consistent with the previous one. These two experiments verify that heteroscedasticity is an intrinsic property of hyperspectral data.

3.3 Verification for Non-stationarity

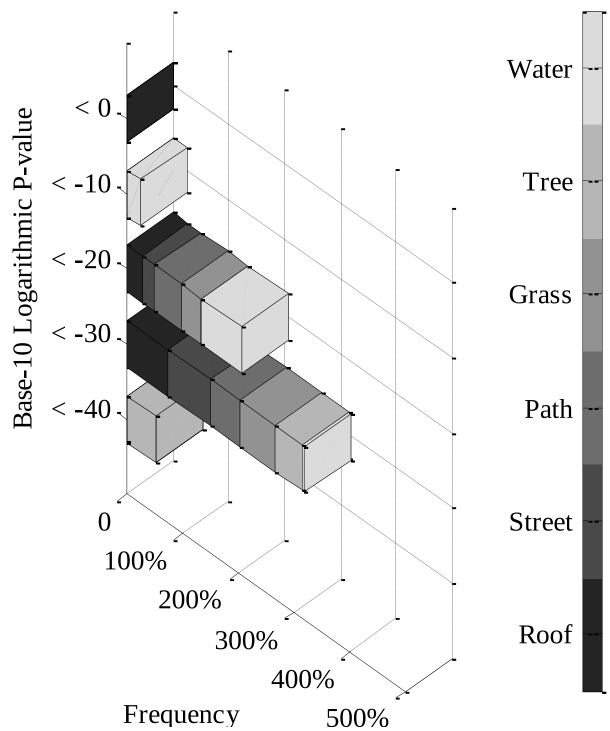

Figure 5. Histograms of Base-10 Logarithmic P-values for Different Classes under Wald-Wolfowitz Runs Test in Washington D. C. Mall Data Set.

Treat hyperspectral response sequence of given pixel as time series, and each hyperspectral response sequence is tested by Wald-Wolfowitz Runs Test. The null hypothesis H_0 is that hyperspectral response sequence are stationary. Test results are grouped by classes of corresponding pixels. Histograms of base-10 logarithmic p-values for different classes in two datasets are depicted in Figure 5 and Figure 6, respectively. There two experiments show that all p-values are too low to receive the null hypothesis H_0, which demonstrates that hyperspectral data are non-stationary statistically significant.

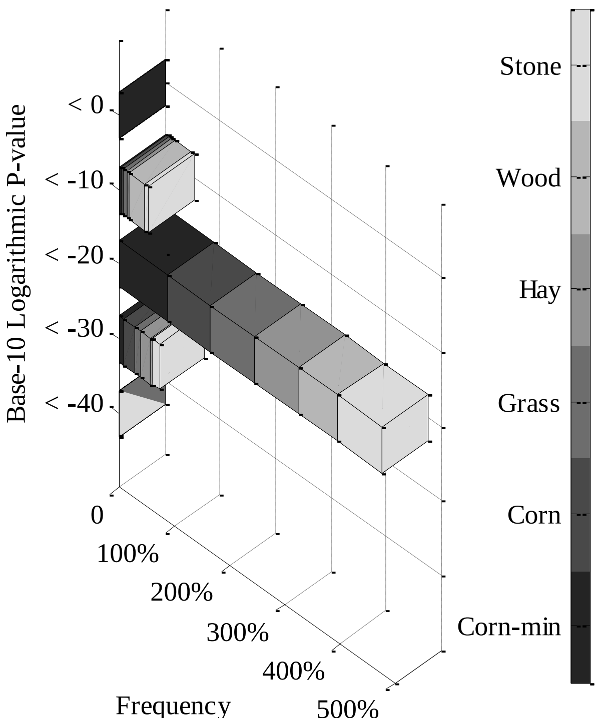

Figure 6. Histograms of Base-10 Logarithmic P-values for Different Classes under Wald-Wolfowitz Runs Test in Indian Pine Data Set.

4 CONCLUSION

This paper addresses the issue of the heteroscedasticity and non-stationarity for hyperspectral image by introducing Bartlett Test and Wald-Wolfowitz Runs Test, respectively. Heteroscedastic series have different variances in different components, while non-stationary processes often include a trend-like behavior. Quantitative experiments show that both heteroscedasticity and non-stationarity are significant statistical properties of hyperspectral data cube. Therefore, future work should consider these two intrinsic properties when designing algorithms of hyperspectral image processing.

5 ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China and Chinese Academy of Sciences Joint Fund of Astronomy under Grant 11078007, by the Research Fund for the Doctoral Program of Higher Education of China under Grant 20101102120030.

6 REFERENCES

1. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. John Wiley & Sons Publication, Hoboken, New Jersy (2003)
2. Richards, J.A., Jia, X.: Remote Sensing Digital Image Analysis, 4th. Springer-Verlag, Berlin, (2006)
3. Yin, J., Gao, C., Jia, X.: Using Hurst and Lyapunov Exponent for Hyperspectral Image Feature Extraction. IEEE Geoscience and Remote Sensing Letters 9, 705--709 (2012)
4. Yin, J., Gao, C., Jia, X.: Wavelet Packet Analysis and Gray Model for Feature Extraction of Hyperspectral Data. IEEE Geoscience and Remote Sensing Letters 10, 682--686 (2013)
5. Yin, J., Gao, C., Jia, X.: Cointegration Theory for Adaptive Target Detection in Hyperspectral Images. In 2012 IEEE Conference on 32nd Geoscience and Remote Sensing Symposium, pp. 4150--4153, IEEE Press, New York (2012)
6. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control ,4th. John Wiley & Sons Publication, Hoboken, New Jersy (2003)
7. C. W. J. Granger, “Some properties of time series data and their use in econometric model specification,” Journal of Econometrics, vol. 16, pp. 121-130, 1981.
8. R. F. Engle and C. W. J. Granger, “Co-integration and error-correction: representation, estimation and testing,” Econometrica, vol. 55, pp. 251-276, 1987.
9. M. S. Bartlett, “Properties of sufficiency and statistical tests,” Proceedings of the Royal Society of London Series A, vol. 160, no. 901, pp. 268-282, 1937.
10. A. Wald and J. Wolfowitz, “On a test whether two samples are from the same population,” The Annals of Mathematical Statistics, vol. 11, no. 2, pp. 147-162, 1940.