Statistical Hypothesis Testing of the Number of Chemical Components in Spectrum Image Data

Shiga, M.[1,2], Fukaya, M.[1] and Muto, S.[3]

[1] Gifu University, Japan, [2] Japan Science and Technology Agency, Japan, [3] IMaSS, Nagoya University, Japan

Recently spectrum imaging (SI) measurements such as STEM-EELS, STEM-EDX and Raman imaging have been generating a massive size of datasets. Then, extensive manual analysis of such datasets has become much more difficult. To reduce this analysis cost, an automatic method based on machine learning techniques is necessary. Our group has tackled this problem based on nonnegative matrix factorization (NMF) and demonstrated the effectiveness for STEM-EELS/EDX datasets [1]. It also provided an optimization of the number of chemical components based on a sparse modeling approach but it still remains parameters to be set by the data analyst. To avoid this issue, we proposes a statistical hypothesis testing to choose the number of components hidden in SI.

Such hypothesis testing is still a hot topic on statistics. For example, Harsanyi, Farrand and Chang's test (HFC test) [2], Kac-Rice test [3] and conditional singular value test (CSV test) [3] were developed using the distribution of singular values of a given data matrix. These methods assume that observation noise is statistically independent both among spectrum channels and spatial points, though the assumption does not always hold in real datasets, resulting the wrong number of components. We propose a new procedure of hypothesis testing using the Henze and Zirker's multivariate normality test (HZ test) [4]. Our procedure first implements singular value decomposition (SVD) and then applies HZ-test to validate the normality of subspaces generated by SVD. Then our procedure concludes that there exist components in a subspace if the HZ-test rejected the null hypothesis of normality. On the other hand, if the HZ-test could not reject the hypothesis, then it concluded that the subspace consisted of Gaussian noise alone. Because it does not require the statistical independence assumption of observation noise, our procedure is more suitable to analyze real SI datasets than existing methods.

We evaluated our proposed method using synthetic and real data. Fig. 1 shows the synthetic data with three components. Fig 1 (a) shows each true component spectrum consisting of a Gauss function with the different center and Fig. 1 (b) shows the component spatial distribution, where the observed data was generated by mixing three spectra with Gaussian noise overlaid. Fig. 2 (a) and (b) show the *p*-values for independent noise data and that for correlated noise data, respectively. These results with significance level of 5% show that all methods chose the correct number for independent noise data but only our procedure chose the correct number (three) for the correlated noise data. We further evaluated them using a real STEM-EELS of Mn_3O_4, in which there exist three chemical components [1]. From the *p*-values shown in Fig. 3(a), only our procedure chose the reasonable number. Using the number chosen by our method, we applied automatic chemical component analysis by vertex component analysis (VCA) [5]. Fig. 3 (b) and (c) show that the estimated component spectra and spatial distributions are identical to theoretical analysis, confirming the correctness of our method.

References:

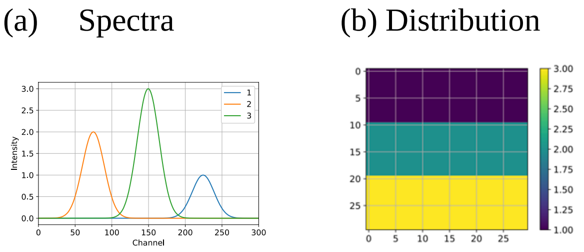
[1] M. Shiga, *et al.*, *Ultramicroscopy*, 170, 43-59, 2016.

[2] C.-I. Chang and Q. Du, *IEEE Trans. on Geoscience and Remote Sensing*, 42(3), 608-619, 2004.

[3] Y. Choi, *et al.*, *The Annuals of Statistics*, 45(6), 2590-2617, 2017.

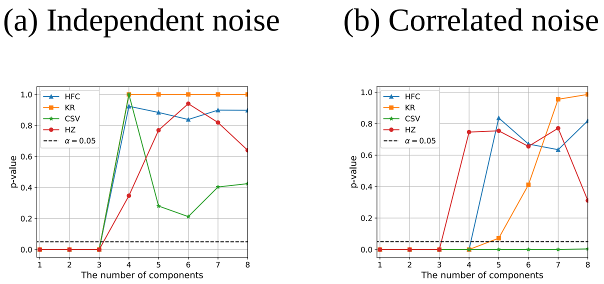
[4] N. Henze and B. Zirkler, *Communications in Statistics-Theory and Methods*, 19(10), 3595-3617, 1990.

[5] J.M. Nascimento and J. M. Dias, *IEEE Trans. on Geoscience and Remote Sensing*, 43(4), 898-910, 2005.



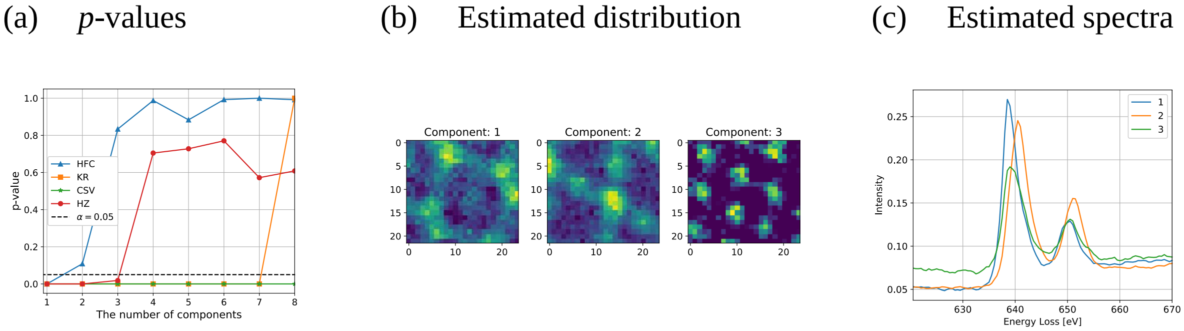
(a) Spectra (b) Distribution

Fig. 1 Synthetic dataset with three components.



(a) Independent noise (b) Correlated noise

Fig. 2 *p*-values by statistical hypothesis testing of the number of components.



(a) *p*-values (b) Estimated distribution (c) Estimated spectra

Fig. 3 Analysis result of real STEM-EELS dataset [1]. (a) Only HZ test chose the reasonable number of components but the chosen number by other methods are only one or much larger. In (b) and (c), components 1, 2 and 3 estimated by VCA indicate Mn^{2+}, Mn^{3+} and O (oxygen), respectively.

Acknowledgement:

This work was partially supported by JST PRESTO #JPMJPR16N6 and JPSP KAKENHI #16H00736, #16H02866 and #25106004.