A new chemometrics preprocessing based on effective information truncation to handle matrix rank deficiencies as well as the effects of noise and light scattering in 3D excitation emission fluorescence matrices

M. Haouchine[1*], C.Biache[1], C. Lorgeoux[2], P.Faure-Catteloin[1], M. Offroy[1]
[1] Université de Lorraine, CNRS, LIEC, F-54000 Nancy, France
[2] Université de Lorraine, CNRS, GeoRessources, F-54000 Nancy, France

Keywords: 3D fluorescence spectroscopy, Excitation Emission Matrix (EEM), Polycyclic Aromatic Compounds (PACs), chemometrics preprocessing, rank deficiencies, truncated SVD, image analysis.

1 Introduction

Fluorescence spectroscopy exploits the phenomenon of natural or induced fluorescence emission, from intrinsic fluorophores or fluorescent chemical derivatives after addition of extrinsic fluorophores, respectively [1]. Polycyclic aromatic compounds (PACs) constitute a large family of mainly anthropogenic chemical contaminants. They have at least two aromatic rings which give them intrinsic fluorescence [2]. Their characterization by eco-friendly 3D fluorescence spectroscopy coupled with chemometrics algorithms constitutes a powerful alternative to the separative techniques conventionally used. However, the systematic presence of Rayleigh and Raman scattering signals in the Excitation Emission Matrices (EEMs) makes spectral decomposition via PARAllel FACtor analysis (PARAFAC) difficult due to the non-trilinear structure of these signals and the matrix rank deficiencies that they generate. There are several strategies to overcome these light scattering effects but weakness remain [3]. Thus, a new chemometrics approach to push back matrix rank deficiencies and to handle these interferences and noise in the data is suggested in this work. It is based on advanced truncation strategy in singular value decomposition (SVD) [4].

2 Theory

The home-made algorithm is structured in three main steps, the step #1 is about data formatting. It allows to prepare data for processing. In the case of EEMs, the reshape operation allows to toggle from 3D space to 2D space thanks to the row-wise or column-wise matrix augmentation. Step #2 exploits an advanced SVD truncation strategy. The challenge with this method is determining the truncation threshold which is the number of singular values to retain and which low rank to choose. The proposed approach attempts to overcome this difficulty because the optimal low-rank is not chosen according to singular values curve versus their numbers, but is deduced through image analysis. Step #3 is a reconstruction of clean data matrix deduced from selected singular values representing all the chemical compound information.

3 Material and methods

In order to implement the home-made algorithm, EEMs of 47 samples were acquired to build a database. Four PACs were selected: Naphthalene (NPH); Benz[a]Anthracene (BaA); Anthracene (ANT) and Pyrene (PYR). The database was distributed as three datasets, where dataset 1 was for individual PAC in dichloromethane at six different concentrations (20, 10, 1, 0.25, 0.1 and 0.05 mg. L-1), while dataset 2 was for mixtures of NPH and BaA at varying concentrations in the same solvent.

Dataset 3 was for mixtures of the four species at varying concentrations. An Aqualog® fluorescence spectrometer was used to acquire EEMs. It is equipped with a charge coupled device detector (CCD) set to medium gain and time integration equal to 1 second. The samples were excited using a range of excitation wavelengths between 239 and 800 nm with a pitch of 3 nm. The fluorescence emission was collected in a wavelength range between 248.27 and 829.32 nm with a resolution of 4 pixels (i.e. 2.33 nm). A Quartz SUPRASIL® cell with a light path equal to 10 mm was used.

4 Results and discussion

In the example shown in the Figure 1, the chemical information is kept intact while the scattering signals have been removed by the preprocessing. Furthermore, the optimal low-rank is found through image analysis and the percentage of information, explained at the truncated matrix level coupled with the analysis of residual information. Finally, a region-based segmentation algorithm enabled automatic cropping of the cleaned map.

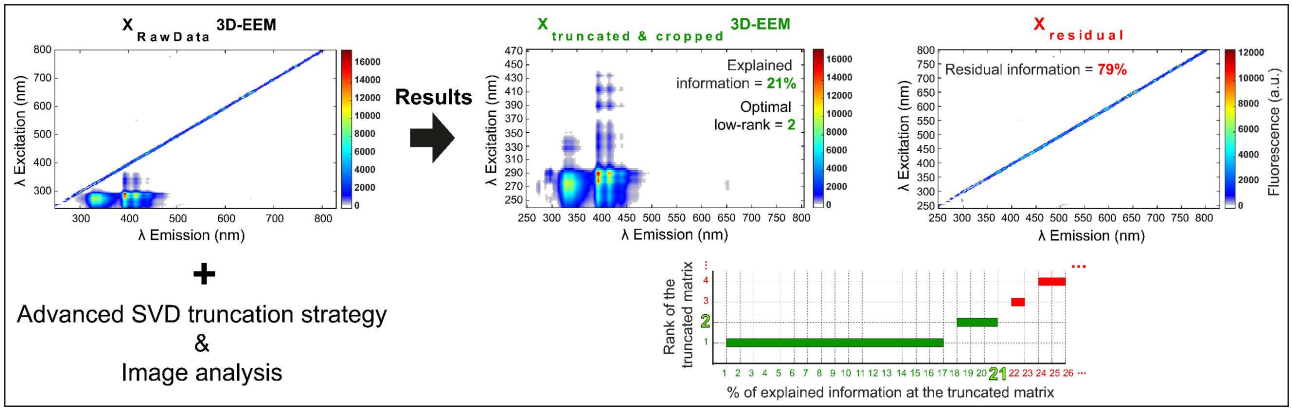

Figure 1 : Fluorescence landscape of a sample from dataset 2 (NPH at 0.5 mg. L-1 and BaA at 0.025 mg. L-1) before and after preprocessing to which are added the different information obtained by our chemometrics approach.

5 Conclusion

The method proposed in this work is based on one of the most common algorithms in linear algebra (i.e. SVD) with an original imaging approach to its application with EEM or EEMs data. Its advantages are that it does not require any information concerning the scattering signals and effectively handle these interferences and noise. Moreover, it provides the percentage of chemical information and noise in the raw data. Finally, it fends off matrix rank deficiencies and generates an estimation of the number of factors to choose for spectral decomposition like PARAFAC.

6 References

[1] J. R. Lakowicz, *Principles of fluorescence spectroscopy*. Springer, 2006.

[2] N. Locquet, A. Aït-Kaddour, and C. B. Y. Cordella, *3D Fluorescence Spectroscopy and Its Applications*. 2018.

[3] M. Bahram, R. Bro, C. Stedmon, and A. Afkhami, "Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation," *J. Chemom.*, vol. 20, no. 3–4, pp. 99–105, 2006, doi: 10.1002/cem.978.

[4] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press, 2019.