P-12

Classification and clustering multivariate statistical methods for hyperspectral datasets in R Environment

K. Banas[1]*, A. Banas[1], E. Jasek-Gajda[2], M. Gajda[2], W. M. Kwiatek[3], B. Pawlicki[4] and M. Breese[1]

[1]Singapore Synchrotron Light Source, National University of Singapore, 5 Research Link, Singapore 117603, Singapore
[2]Department of Histology, Jagiellonian University Medical College, Kopernika 7, 31-034 Krakow, Poland
[3]Institute of Nuclear Physics PAN, Radzikowskiego 152, 31-342 Krakow, Poland
[4]Gabriel Narutowicz Hospital, Pradnicka 37, 31-202 Krakow, Poland

Keywords: synchrotron radiation, x-ray fluorescence, multivariate statistical analysis, classification techniques

*e-mail: slskb@nus.edu.sg

The experiments performed at synchrotron light sources very often provide as the result big datasets. This is especially true with 2D spectroscopy. Hyperspectral datasets (spectra with additional information for example about the position of the place where spectrum was recorded) should be treated in a special way. They are highly correlated in two-fold way: each spectrum is a superposition of the number of peaks with additional baseline function, but also spectra from adjecent regions are usually very similar due to local homogenity of the sample. Additionaly, very often these datasets represent so-called wide data case where the number of variables is bigger than the number of observations.

While there is a number of software solution for evalution of the experimental results in the image format (for example for imaging and tomography experimental results) hyperspectral data evaluation standardised approach is still missing.

In this contribution discussion and comparison of two methods for X-ray fluorescence (XRF) spectral datasets evaluation is presented.

Traditionally each spectrum is deconvoluted by fitting the model in order to obtain elemental concentration values, subsequently these concentrations are used as the variables in building classification models by using linear discriminant analysis (LDA) or partial least-square discriminant analysis (PLSDA).

Proposed alternative approach is using directly complete spectral datasets. By using multivariate statistical techniques reduction of the dimension is performed, then new variables (principal or latent components) are included for constructing classification functions.

Comparison of the performance of models constructed with both methods and LDA or PLSDA will be shown.

Cross-validation of the models is done by leave-one-out (LOOCV) method. Alternative approach for unsupervised classification based on hierarchical cluster analysis allows for additional independent validation.

XRF spectral data were recorded at beamline L of Hasylab synchrotron source. Samples were 15 microns thin sections of biological material stretched on Mylar foil. Policapillary was used to focus X-rays into small spot size allowing spatially resolved study of heterogenous material.

Complete data preprocession, evaluation and visualization (except deconvolution of XRF spectra model fitting) was performed with R environment [1] for statistical analysis and RStudio Graphic User Interface [2].

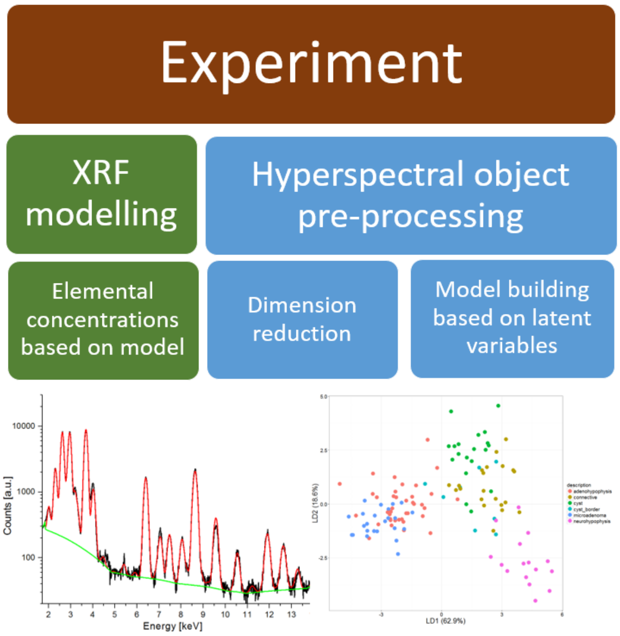

Figure 1. Diagram showing two possible approaches for hyper-spectral data analysis.

Acknowledgments: This work was partially supported by NUS Core Support C-380-003-003-00 and the European Community under the Contract RII3-CT-2004-506008 (IA-SFS) (HASYLAB project No II-20052050 EC)

[1] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2016.

[2] RStudio: Integrated development environment for R. 2016; http://www. rstudio.com