The 4th Aquaphotomics International Conference

A closer look at preprocessing with focus on aquaphotomics

Federico Marini [1],*

[1] Department of Chemistry, University of Rome La Sapienza, P.le Aldo Moro 5, I-00185 Rome, Italy;
* Presenting author and Correspondence: Federico.marini@uniroma1.it

The workshop will illustrate the main chemometric strategies for data preprocessing with particular focus on predictive model building in the context of Aquaphotomics. First of all, the rationale for data preprocessing will be presented and the main family of techniques will be discussed. In detail, methods for smoothing, scatter correction, spectral differentiation, normalization and baselining/detrending will be presented. Moreover, some recently proposed approaches, such as Variable Sorting for Normalization, which allow a preliminary weighting of the variables will also be introduced. Lastly, strategies for boosting model performances based on the fusion of different preprocessing approaches will also be presented.

Keywords: Pre-processing, derivation, scatter correction, data fusion

Introduction

Spectroscopic data (or, more in general, experimental data) may be affected by several sources of variability, not all of interest for the specific task the data are collected for. On the other hand, when chemometric tools are applied to the data, very often model building is based on extracting components accounting for a relevant share of the variance in the predictor space, so that all the sources of data variability (wanted or unwanted) will be included in the model: accordingly, if spurious/unwanted variance is still present in the data, it can have a detrimental effect on the resulting model. To, at least partially, reduce or eliminate the effect of such unwanted variability, chemometric model building usually includes one or more pre-processing steps [1]. However, the choice of the best pretreatment or combination of pretreatments to be applied to the data is not always obvious and, in general, a trial and error procedure is followed. Aim of the present workshop will be to illustrate the main chemometric strategies for data preprocessing, by presenting their theoretical background and discussing the impact of varying their metaparameters, if any. All the techniques will be illustrated by means of worked examples in Matlab and functions will be made available to the participants upon request.

Content of the workshop

As anticipated, data may be affected by many different sources of spurious variation and the aim of data pre-processing is to remove as much as possible the impact of such sources on the spectroscopic signal. Since the attention will be mainly focused on NIR data, the techniques which will be illustrated in greater detail are the ones which are most frequently used when dealing with such profiles. A relevant role in this context is played by scatter correction techniques, such as the standard normal variate transform (SNV) [2] and multiplicative scatter correction (MSC) [3], together with its extended version (EMSC) [4], which has proved to be particularly effective in Aquaphotomics studies. Together with this “well-established” approaches, a recently proposed algorithm, named Variable sorting for normalization (VSN) [5], which allows to weight the predictors in a hypothesis-free way prior to the calculation of the correction parameters will also be discussed.

Lastly, some recent trends in the pre-processing of spectroscopic data based on the use of data fusion approaches will also be presented. Indeed, one of the main problems when dealing with data pre-processing is to select what the best individual technique or combination of techniques could be and, in the latter case, to define the order according to which the different techniques should be applied. Recently, the possibility of exploiting the advantages of multi-block data analysis to overcome the problems related to the limitations illustrated above was presented in the literature. In particular, it was noted how, by applying different pre-processings to the same data, the resulting matrices constitute a multi-block set, which can be processed by specifically designed data fusion approaches. In this framework, the possibility of using sequential and orthogonal approaches for multi-block modeling may represent an advantage, as they allow to evaluate the relevance of the individual pre-processings, the possible redundancies and the incremental contribution to the model. Accordingly, the combination of sequential and orthogonalized partial least squares modeling (SO-PLS) with the use of multiple pre-processing techniques to build a multi-block set is exploited by the recently proposed strategy called SPORT (Sequential Preprocessing through ORThogonalization) [6]. The theory and the practical use of this technique will also be illustrated during the workshop.

References

1. Roger J-M, Boulet J-C, Zeaiter M, Rutledge DN (2020) Preprocessing methods. In: Brown, SD, Tauler R, Walczak B (eds) Comprehensive chemometrics 2nd ed. Elsevier, Oxford, UK, vol.3 pp. 1-73.
2. Barnes RJ, Dhanoa MS, Lister SJ (1989) Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Appl Spectrosc 43:772-777.
3. Geladi P, McDougall D, Martens H (1985) Linearization and scatter-correction for near-infrared reflectance spectra of meat. Appl Spectrosc 39:491-500.
4. Martens H, Stark E (1991) Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. J Pharm Biomed Anal 9:625–635.
5. Rabatel G, Marini F, Walczak B, Roger J-M (2020) Variable sorting for normalization. J Chemometr 34:e3164.
6. Roger J-M, Biancolillo A, Marini F (2020) Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. Chemometr Intell Lab Syst 199:103975.