Preprocessing revisited

Jean Michel Roger1, Alessandra Biancolillo2, Federico Marini3,\*

1ITAP, INRAE, 361 Rue Jean François Breton, 34196 Montpellier, France, jean-michel.roger@inrae.fr

2Dept. of Physical and Chemical Sciences, University of L’Aquila, via Vetoio, 67100 Coppito, Italy, alessandra.biancolillo@univaq.it

3Dept. of Chemistry, University of Rome La Sapienza, P.le Aldo Moro 5, 00185 Rome, Italy, federico.marini@uniroma1.it   
\*Corresponding author

Spectroscopic data (or, more in general, experimental data) may be affected by severl 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. 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. In the present communication, a recently proposed strategy, called Sequential Preprocessing through ORThogonalization (SPORT) and based on the idea that the same set of spectra, differently preprocessed could result in a multi-block data, and, accordingly, be processed through dedicated multi-block strategies, will be presented (Roger et al., 2020). It relies on the use sequential and orthogonalized partial least squares regression (SO-PLS; Biancolillo & Næs, 2019), due to the possibility of including/excluding blocks, evaluating their incremental contribution and identifying which matrices carry common and distinctive information). With the occasion, a recently proposed alternative to data normalization called Variable Sorting for Normalization (VSN; Rabatel et al., 2020) will also be introduced.

**Keywords:** Data preprocessing, Sequential and Orthogonalized Partial Least Squares regression (SO-PLS), Sequential Preprocessing through ORThogonalization (SPORT)

REFERENCES

Biancolillo, A., Næs, T., 2019. The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions. In: Cocchi, M. (Ed.), Data fusion methodology and applications, Elsevier, Oxford, 157-177. https://doi.org/10.1016/B978-0-444-63984-4.00006-5.

Rabatel, G., Marini, F., Walczak, B., Roger, J.-M., 2020. VSN: Variable sorting for normalization. J. Chemom. 34, e3164. https://doi.org/10.1002/cem.3164.

Roger, J.-M., Biancolillo, A., Marini, F., 2020. Sequential Preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. Chemom. Intell. Lab. Syst. 199, 103975. https://doi.org/10.1016/j.chemolab.2020.103975.