Using functional representations in spectrophotoscopic variables selection and regression

F. Corona[1,2], E. Liitiäinen[2], A. Lendasse[2] and R. Baratti[1]
[1] Department of Chemical Engineering and Materials, University of Cagliari, Italy
[2] Laboratory of Computer and Information Science, Helsinki University of Technology, Finland

An intrinsic characteristic of high-resolution measurements from a spectrophotometer is that the spectra can be regarded as continuous functions observed at discretized arguments in the instrument's range of wavelengths. Because of such a distinctive feature, the problem of estimating the output variable is defined from very high-dimensional, inherently collinear and, thus, redundant inputs.

To address the problem, two approaches are commonly used. The first solution is to rely on full-spectrum methods for linear dimension reduction coupled with regression: PCR and PLSR are reference methods. The natural refinement of such an approach is to perform a preliminary selection of relevant wavelength ranges. Unfortunately, the most classical methods are intrinsically limited by their linear structure and, because based on combinations of the original variables, are not trivial to interpret. Moreover, the insight might be further reduced with nonlinear generalizations of the methods. The second solution consists of selecting only individual inputs that truly contribute to a correct estimation of the output and, that are as much as possible not collinear. Typically, the input selection approach is based either on first-principle considerations or data-driven methods like dependency measures and stepwise techniques. Thus, variable selection is the limit extension of range selection where the chemical interpretability of the spectral inputs is explicitly retained and used in the regression model.

This study focuses on the problem of variable selection starting from a functional point of view and, specifically, it addresses the case where no *a priori* chemical knowledge on the problem is available. The recognition that spectra are discretized smooth functions over the wavelengths' domain is regarded as an important source of information for selecting only the spectral inputs that are maximally relevant for the output and also minimally redundant. Such variables emerge in the correspondence of the functional features that characterize the shape of the spectral curves: that is, at wavelengths where not only the functions' values but also the slope and curvature are significant for estimating the output.

The methodology we suggest is general and model independent. Generality arises from the fact that it is suitable for any continuous measures of relevance: for instance linear correlation and mutual information can be used. By model independence we mean that the methodology does not require any assumptions on the regression technique to be used afterwards: in fact, both linear and nonlinear models can be used without compromising the understandability of the problem.

To demonstrate such properties and to support the presentation, the methodology is validated and discussed on a number of study cases ranging from laboratory to full-scale applications where the prediction accuracy as well as the complexity of the resulting regression models is compared with the standard methods used in spectroscopy.