SIMCA®

Application Note

Multivariate Calibration in SIMCA of Spectroscopic Data

10 June 2020

Introduction

Using techniques like NIR, IR, Raman, UV and Fluorescence Spectroscopy, hundreds of signals are recorded per sample or process time point. The use of SIMCA® to analyze these multivariate data provides an overview of the signals and translates them into information about sample and/or process quality.

The fast and non-invasive spectroscopic sensors make up an important part of implementation of Process Analytical Technology (PAT) and can in combination with Umetrics multivariate tools provide process understanding and in the end lay the basis for a control strategy to be applied in the manufacturing process.

NIR case study

Near-infrared (NIR) spectroscopy has become an accepted analytical technique in the food, agricultural, petrochemical, and pharmaceutical industries. NIR is often applied for analysis of bulk substances such as moisture, protein, fat, and alcohol. The advantage is that NIR is non-destructive, fast and can potentially be implemented on-line to facilitate real-time quality control.

Carrageenan NIR data

Carrageenan's are polysaccharides extracted from seaweed and used as thickening agents in foods, pharmaceuticals, and cosmetics. In commercial production, the raw material (seaweed) contains a variety of carrageenan types with different thickening properties. The properties of the final product are a function of the concentration and type of carrageenan and hence it is critical to know the composition to ensure product consistency and quality.

The following case study consists of 128 carrageenan powder samples with varying amounts of 5 different carrageenan types: Lambda, Kappa, Iota, Mu, and Nu. NIR spectra were acquired from 1100–2500 nm, resulting in a total of 699 variables.

Plot spectra

The first step in the data analysis is to import the data and plot the spectra. SIMCA supports the direct import of standard file formats such as JCAMP, xls and csv, as well as several instrument vendors formats. In addition, SIMCA utilizes ODBC (Open Database Connectivity), allowing it to easily access data from a variety of other database management systems.

All NIR spectra are plotted in Figure 1, providing the first overview of the raw data. The spectra are colored according to the Lambda content (Y), a critical quality parameter.

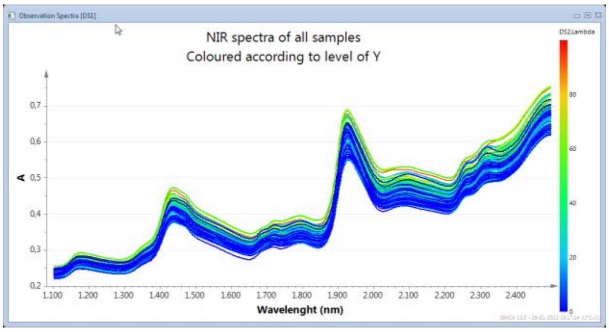

Figure 1: Raw NIR spectra of 128 carrageenan samples.

Spectral filtering

To remove irrelevant light scatter effect from the spectra or standardize your spectroscopic signal, SIMCA offers several spectral filters including Multiplicative Signal Correction (MSC), Standard Normal Variate (SNV) and 1st, 2nd, or 3rd derivatives. In figure 2 the SNV filtered NIR spectra for all 128 samples are plotted. Compared with the raw spectra in figure 1, the different baseline offset between the samples is now removed. The coloring of the spectra indicates several spectral regions that separate the samples according to Y-value and thus contain the relevant information about the critical quality parameter.

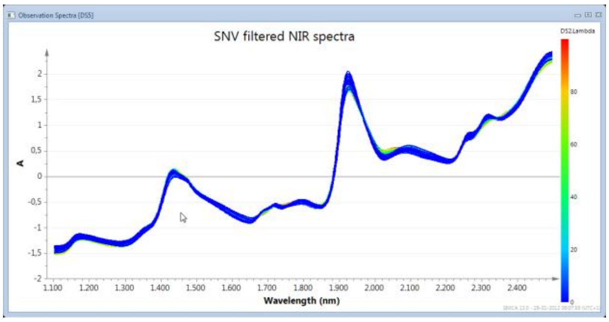

Figure 2: SNV filtered NIR spectra of 128 carrageenan samples.

Principal Component Analysis (PCA) for overview

The PCA score plot gives an overview of the distribution of all samples, as illustrated in figure 3. Again, coloring tools can be applied to explore patterns. The 128 carrageenan samples were collected over 5 days and the coloring scheme indicates no major systematic “day” effect in the main variation between the NIR spectra. The evenly distributed score plot with no outliers, as observed in figure 3, suggests that the data is a prime candidate for further multivariate calibration.

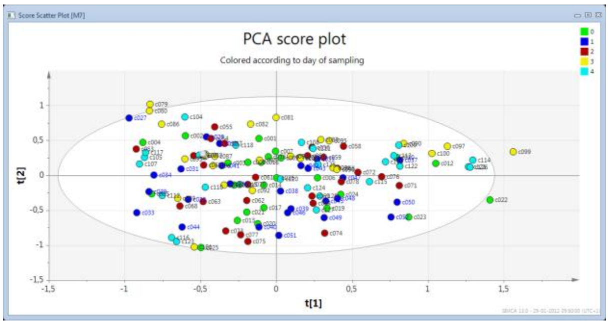

Figure 3: PCA score plot of all samples, based on SNV filtered NIR spectra.

Multivariate Calibration modelling

Regression models between spectra and quality parameters are built in SIMCA using Partial Least Squares (PLS) – or orthogonal PLS (OPLS) for improved interpretation. For model validation, cross-validation is applied to determine model dimensionality and permutation testing for further validation of the final model. For the carrageenan data, a 5 (1+4) component OPLS model between SNV filtered NIR data and the Lambda content is developed. Figure 4 summarizes the model in terms of a plot of observed versus predicted Lambda-values. The model has a predictive performance of Q2=0.99 and an average prediction error, RMSECV = 2,17 %.

Model interpretation is easily carried out with an OPLS model, where the first predictive loading vector, p1 indicates which wavelengths in the NIR spectra relates to the Y parameter, in this case the Lambda value. The loading in figure 4 shows a complex pattern meaning that several spectral regions are correlated with Y and contributing to the overall prediction model.

Perspectives

This small case study illustrates the common workflow in multivariate calibration in SIMCA including

- Easy import
- Flexible and user-friendly plotting tools
- Powerful multivariate models
- Sound model validation tools

The next step can be to combine spectral data with other data sources like process data in hierarchical modelling. Or use the models in real time monitoring. The multivariate calibration models built in SIMCA can be used directly in SIMCA-online for real time prediction of critical parameters or early fault detection.

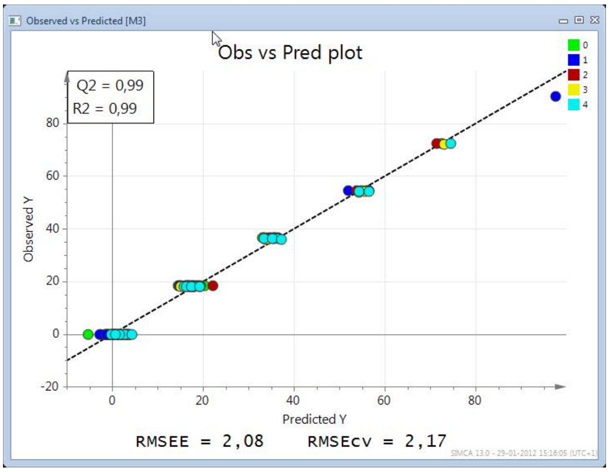

Figure 4: Observed vs. Predicted Lambda-values from 5 (1+4) component OPLS model based on SNV filtered NIR data.

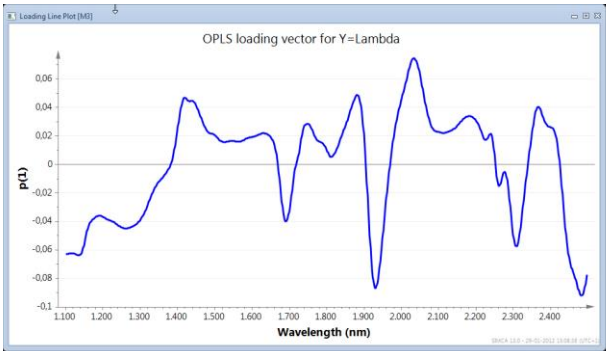

Figure 5: Line plot of first OPLS loading vectors indicating which wavelengths in the NIR spectra that carries information about the content of Y.

References

Dyrby, M. Carbohydrate Polymers 57, 337-348, 2004

FDA. 2004. Guidance for Industry, PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance.