Du Tc H Chemometrics Society

30th Annual Chemometrics Symposium

October 13th, 2016
Ravenstein

"Chemometrics in Action"

Program and Abstracts

Organising Committee

Paul Eilers, *Erasmus MC Rotterdam.*
Gerben Mooiweer, *JDEcoffee/R&D, Utrecht.*
Bülent Üstün, *Aspen Oss BV, Oss.*
Jeroen Jansen, *Radboud University Nijmegen, Nijmegen.*
Jos Hageman, *Wageningen University, Wageningen.*
Theo Reijmers, *Kinesis Pharma, Breda*.

Program

9.15   Registration and coffee.

10:10   **Paul Eilers** – Opening.

10.15   **Rudolf Kessler**
        "_Perspectives in process Analysis: remarks on robustness of spectroscopic_
        methods for industrial applications**"**

11.15   Coffee/tea break.

11.30   **Ewoud van Velzen**
        "_Robustness: A key parameter in the validation of multivariate sensory_
        regression models"

12.00   **Andrei Barcaru**
        "_Thinking Bayes: creating new ways of interpreting GCxGC(-MS) data_"

12.30   Walking lunch.
        Demonstration of microNIR spectrophotometer.

14.00   **Cyril Ruckebusch**
        "_Chemometric analysis of chemical images_"

14.45   **Olga Lushchikova**
        "_Development of an alarm system for water quality control based_
        on flow cytometry data of algal populations"

15.15   Coffee/tea break.

15.30   **Jan Geretzen**
        "_A novel and effective way for preprocessing selection based on design of_
        experiments"

16.00   Closing and drinks.

Abstracts (in order of presentation)

Rudolf Kessler *Perspectives in Process Analysis: Remarks on Robustness of Spectroscopic Methods for Industrial Applications*

Intelligent manufacturing has attracted enormous interest in recent years, not least because of the Process Analytical Technology/Quality by Design platform of the FDA and now also by the initiative Industrie 4.0. The future of industrial automation will be "arbitrarily modifiable and expandable (flexible), connect arbitrary components of multiple producers (networked), enabling its components to perform tasks related to its context independently (self-organizational) and emphasizes ease of use (user-oriented)". Optical spectroscopy will play a major role in the sensor technology as it provides simultaneously chemical and morphological information.

The sensitivity of an analyte can be described by the quantum mechanical cross sections of the molecule which are the effective area that governs the probability of an event of e.g. elastic scattering, or absorption, or emission (e.g. fluorescence or Raman) of a photon at a specified wavelength. Another very important criterion for an analytical method is its capability to deliver signals that are free from interferences and give "true results" and is called selectivity resp. specificity. The responses are based on interactions usually evaluated in a mathematical domain by e.g. chemometrics, giving what has been called "computational selectivity". Besides these key elements, many other aspects must also be taken into account to ensure robustness of the inline measurement.

I will discuss the advantages and disadvantages of spectroscopic inline techniques. Examples demonstrate robust optical set ups in different industrial applications, e.g. biotechnology, pharmacy and manufacturing industry. I illustrates the opportunity to receive robust spectroscopic information with causal relations to the target value. It is important to emphasize that inline quality control by spectroscopic techniques is a holistic approach. Process chemists, process engineers, chemometricians, and many other technologists must work together, where multimodality will be a bedrock supporting the production of smart materials in smart factories.

Ewoud van Velzen: *Robustness: a key parameter in the validation of multivariate sensory regression models*

International guidelines define robustness of an analytical procedure as a measure of its capacity to remain unaffected by small variations in experimental conditions. Robustness is generally not considered in validation protocols even though it provides an important indicator of the fitness of an analytical procedure during normal use. Also in the validation of multivariate sensory regression models robustness is seldom included. Particularly in "Sensoromics" analysis, where multivariate relationships are assessed between descriptive (consumer) data and compositional (chemical) data, we state that a measure of robustness is a critical parameter for establishing model reliability and model suitability.

We demonstrated the importance of robustness testing in a sensory study in which the chemical compositions of natural food products was associated with the descriptive scores of a trained consumer panel. This multivariate association was modelled by means of PLS. The data was rather complex as (i) the predictor set was compiled of various analytical datasets with different error structures, and (ii) the predictor set contained missing data. Furthermore, the data was sparse given the enormous sample diversity in the training set. By integration of predefined noise simulations in the double cross-validation and the permutation testing, estimates of the relevant model statistics (e.g. Q2-values) could be obtained. The aim of this approach is that the susceptibility of multivariate models against experimental errors can be tested. We found that robustness testing is of great additional value in the validation of multivariate prediction models. An important prerequisite, however, is that the error structures of the dependent and independent variables should be well known in advance.

Andrei Barcaru: *Thinking Bayes: creating new ways of interpreting GCxGC(-MS) data*

Nowadays, Bayesian statistics becomes increasingly popular in application fields as diverse as psychometrics, geophysics, sociology, medicine and (last but not least) analytical chemistry. The objectivity and robustness of Bayesian analysis made it an important tool in the WW II. Nowadays it has become the mathematical representation of how our brain works.

In the CHROMAMETRICS project, we used Bayesian statistics to create powerful and robust algorithms for processing data originating from chromatographic and mass spectrometry instrumentation. Some useful applications developed within the CHROMAMETRICS project, based on Bayesian paradigm, will be discussed in this communication. For example, Bayesian statistics has been applied to peak tracking in two-dimensional gas chromatography. Peak tracking refers to the matching of two peak tables. When a sample is analysed in two different conditions (different inlet pressure and/or temperature program), the goal is to match the compounds from one chromatogram with those in the other. The application of Bayesian analysis to this problem results in a paradigm shift compared to the way traditional peak matching algorithms work. Instead of a single (deterministic) answer, a collection of answers about possible peak matching results is achieved, with a probability assigned for each peak assignation. Another example is the comparison of two samples analysed in GCxGC-MS, to find meaningful differences among them. In this case, the objective is to find compounds that are present in one sample but absent in the other. The application of Jensen-Shannon divergence within a Bayesian framework proved to give meaningful results, as well as robust against peak misalignments.

In this communication, emphasis is given in the paradigm shift that using Bayesian analysis supposes, against traditional data analysis methods. Moreover, the methods developed in this project integrate different tools from different fields such as information technology and image analysis within the Bayesian framework.

Cyril Ruckebusch: *Chemometric analysis of chemical images.*

Chemical imaging aims to create a visual image of the different chemical compounds in a sample. This can be achieved by simultaneous measurement of spatial, spectral and/or temporal information. Image processing is required to obtain information regarding the mapping or distribution of these compounds in a biological or chemical image. In this talk, we will present our recent efforts to process hyperspectral images and super-resolution wide-field fluorescence nanoscopy data. We will show how the results obtained can be relevant to analyze other types of imaging modalities.

Olga Lushchikova: *Development of an alarm system for water quality control based on flow cytometry data of algal populations*

Rijkswaterstaat (RWS) applied flow cytometry for the on-line biomonitoring of algal cells in the river water to perform an accurate and fast water quality control. This biomonitoring is based on the measurement of optical properties of every single cell, leading to a large amount of data. This is why we developed a dedicated chemometric method, which is called DAMACY, to analyse the spectroscopic data. DAMACY is a multivariate method, which first describes the measured fingerprints of the algal communities. Secondly, this method predicts the normal environmental conditions such as the water temperature or day length. If the predicted natural factor shows a large deviation, the sample can be examined more carefully to find the non-natural cause. Previously the results were obtained using offline methods, such as microscopy. Because of this procedure the monitoring could not be frequent whereupon the small water pollutions were hidden. This warning system will detect specific water pollution events early and robustly.

Jan Gerretzen: *A novel and effective way for preprocessing selection based on design of experiments*

The aim of data preprocessing is to remove data artifacts—such as a baseline, scatter effects or noise—from the data and to enhance the analytically relevant information. Many preprocessing methods exist to do either or both, but it is not at all clear beforehand which preprocessing methods should optimally be used. Appropriate preprocessing selection is a very important aspect in data analysis, since we have shown that multivariate model performance is often heavily influenced by preprocessing.

Current preprocessing selection approaches, however, are time consuming and/or based on arbitrary choices that do not necessarily improve the model. It is thus highly desirable to come up with a novel, generic preprocessing selection procedure that leads to a proper preprocessing within reasonable time. In this presentation, a novel and simple preprocessing selection approach will be presented, based on Design of Experiments (DoE) and Partial Least Squares (PLS). I will show that preprocessing selected using this approach improves both improves model performance and model interpretability. The latter will be illustrated using variable selection. The focus of the presentation will be on spectroscopic data.

This work is part of the 'Analysis of Large data sets By Enhanced Robust Techniques' project (ALBERT), funded by TI-COAST, which aims to develop generic strategies and methods to facilitate better and more robust chemometric and statistical analyses of complex analytical data.

This meeting is sponsored by: