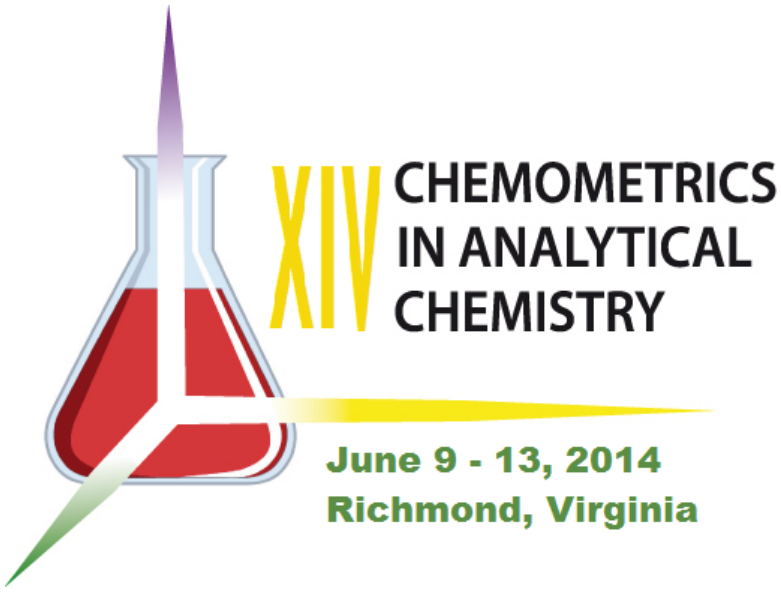
Abstracts

Chemometrics and Analytical Chemistry 2014



PL1

PERSPECTIVES ON THE INTERDISCIPLINARY NATURE OF CHEMOMETRICS AND THE FUTURE OF ITS IDENTITY AS A DISCIPLINE

Paul J. Gemperline[1], Maryann Cuellar[1], and Paul Trevorrow[2]
[1]*Department of Chemistry, East Carolina University, Greenville, NC 27858*
[2]*John Wiley & Sons Limited, The Atrium, Southern Gate, Chichester, West Sussex, United Kingdom. PO19 8SQ*

Chemometrics as an identifiable discipline is about 40 years old. It is characterized by its specialized jargon and distinctive mathematical methods. In the pioneering years it was identified as a sub discipline of analytical chemistry. It experienced rapid growth during the 1990's and 2000's, during which time publications grew at a linear rate while citations grew nearly exponentially, a dramatic indication of chemometrics' growing impact in a broad range of multidisciplinary fields. By these measures, chemometrics has been highly successful as a discipline. However, as the use of data analytics has become ubiquitous in the past decade, is research in chemometrics at risk of becoming irrelevant? Has the pace of innovation and development of new chemometric methods stalled, or worse yet, is it in decline? Are there still new methods to be discovered and invented, or, as knowledge and expertise in mathematics and computational methods has risen to new levels throughout the world, have all novel and innovative mathematical tools been discovered? If chemometrics is to remain a viable discipline, where will the next innovations come from?

This perspective will examine the impact of chemometrics in three specialized research areas as bellwether indicators of the discipline and its future, including work on use of Raman spectroscopy for monitoring bioprocesses. The perspective will end with an examination of human creativity and innovation and speculate on some of the possible roles that the multidisciplinary approach of chemometrics might play in helping to solve some of the grand challenges facing the human race and our planet.

PL2

CRITICAL ISSUES IN ANALYZING METABOLOMICS DATA

Age K. Smilde

Biosystems Data Analysis, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
Clinical Epidemiology, Biostatistics and Bioinformatics (KEBB), Academic Medical Center, Meibergdreef 9, 1100 DE Amsterdam, The Netherlands
Department of Food Science, University of Copenhagen,
Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.
a.k.smilde@uva.nl

Metabolomics is the new kid on the block in functional genomics and relies heavily on advanced instrumental techniques such as GC-MS, NMR and LC-MS. It can be used to probe metabolism in cellular organisms, to analyze metabolites in body-fluid samples, plant extracts and food to name a few examples. The purpose of these measurements is dictated by the biological question underlying the study. In analyzing metabolomics data several crucial steps have to be undertaken; one of those is processing the data in such a way that the biological question is answered. Due to the large amount of data and the high demands posed on the quality of the results of the study, data processing is a very important step. To this end techniques as PCA and PLS can be used but these are not sufficient, e.g., also association networks are becoming very popular. There are, however, some critical issues in processing metabolomics data that need to be addressed. Without presenting an exhaustive list a few will be commented upon. First, the measurement error is considerable and heteroscedastic. Secondly, the measurement error can be correlated for comprehensive metabolomics data. Thirdly, many metabolites are not informative and should be removed. Fourthly, the data are usually highly structured according to an experimental design with fixed and random factors. Fifthly, data sets have to be fused to arrive at a global data set since usually measurements are performed on more than one analytical platform and sometimes even on other omics platforms. These issues will be discussed, illustrated with real-life examples and a partial solution will be provided.

PL3

CHEMOMETRICS IN THE LQTA-UNICAMP BRAZIL: RECENT THEORETICAL ADVANCES AND NOVEL APPLICATIONS

Marcia M. C. Ferreira
LQTA – Laboratory for Theoretical and Applied Chemometrics
Institute of Chemistry, University of Campinas, Campinas, SP 13084-971, Brazil
marcia@iqm.unicamp.br

LQTA was founded at the end of 1996 when I returned from a postdoc at CPAC under the supervision of Prof. Bruce Kowalski. During these eighteen years our group was involved in several applications and methods were developed to attend our needs.

LQTA-QSAR[1] is an example. This software was developed for building 4D-QSAR models and named after our lab (***L**aboratório de **Q**uimiometria **T**eórica e **A**plicada*). This new methodology explores jointly the main features of CoMFA and 4D-QSAR paradigms. It is based on the generation of a conformational ensemble profile, CEP, for each compound, followed by the calculation of 3D descriptors. GROMACS free package is used for molecular dynamics simulations and generating CEP. The module **LQTAgrid** calculates intermolecular interaction energies at each grid point considering different probes and all aligned conformations from MD simulations. These interaction energies are the descriptors employed in the QSAR analysis.

QSAR-Modeling is the module used for building and validating PLS regression models. Models are thoroughly validated applying the leave–*N*–out cross-validation and y-randomization methods. The ordered predictor selection, **OPS**, algorithm[2] also developed in our laboratory, is used for feature selection in the construction of the PLS models. Applications will be presented.
Other methods and their applications will also be presented.

References

[1] Martins JP; Barbosa E; Pasqualoto KF; Ferreira MMC. *J. Chem. Inf. Comput. Mod*. **2009**, *49*, 1428.
[2] Teófilo RF; Martins JP; Ferreira MMC. *J. Chemom.*, **2009**, *23*, 32.

K1

GENE SETS: STROLLING THROUGH A (RANDOM) FOREST

Dan Jacobson[1], Philip Young[1], Erik Alexandersson[2], Melane Vivier[1]
[1]Institute for Wine Biotechnology, Stellenbosch University, Stellenbosch, South Africa 700
5[2]Department of Plant Protection Biology, Swedish University of Agricultural Sciences, Alnarp, Sweden
jacobson@sun.ac.za

Agricultural field studies often take place in very heterogeneous environments and, as such, the experimental design and sampling strategies can have a profound effect on experimental results. Plants are incredibly plastic organisms and adjust their gene expression and metabolism in response to their immediate environment. The fact that soil, wind and shade conditions can vary dramatically in a field even within the space of a few meters (or indeed millimetres) presents significant challenges. Systems biology approaches to field studies are already generating large omics data sets that routinely contain tens of thousands to millions of variables. In this talk we will examine the spatial environmental patterns in a vineyard field trial and their effect on small (dozens of variables) and large (35,000 variables) multivariate datasets. Even when experimental design and sampling is done properly, the orthogonal variables at play in the field can wreak havoc with statistical data analysis. Here we explore experimental designs and sampling strategies to account for this heterogeneity and an intersection between non-parametric agglomerative and machine-learning-based data analysis methods to deal with the challenging results that we are often faced with in such Fieldomics studies.

K2

CHEMOMETRICS MODELS OVER THE DECADES: STRATEGIES FOR LONG-TERM SUPPORT

Mary Beth Seasholtz, Wendy Flory, and Serena Stephenson
The Dow Chemical Company, 1897 Building, Midland MI 48667
mseasholtz@dow.com

Chemometrics has been applied to spectroscopy data for various process samples at The Dow Chemical Company since the early 1990's. The old adage "watch what you wish for" applies to us because now we have to support the hundreds of PLS models that are installed at manufacturing facilities across the world, and the original developers are no longer around for help! Chemometrics models are part of a calibrated method (as opposed to an absolute method), and so they need occasional validation, maintenance, etc., but the challenge is how to keep these models among the living over the long haul.

This talk will cover the kinds of requests that are received, and how we are putting a system in place to so we can address the requests in a reasonable amount of time and effort. Further, opportunities for improvement will be pointed out.

K3

LET THE DATA DO THE TALKING, COMBINING CHEMOMETRICS AND SPECTROSCOPY TO EXPLORE BIOLOGICAL SYSTEMS

Renee Jiji
Department of Chemistry, University of Missouri-Columbia, 601 S. College Avenue, Columbia, MO 65211-7600
jijir@missouri.edu

Biological systems and the data they generate carry a natural variance that can make quantitative predictions difficult. Furthermore, spectroscopic analysis frequently carries with it, its own set of challenges stemming from broad overlapping peaks, irregular baselines and low signal-to-noise. Chemometric methods can be used to address these issues and augment spectral analysis of biological molecules and macromolecules, such as proteins. The use of data fusion strategies to improve calibration and the power of multiway methods for exploratory analysis of biological systems will be presented.

K4

CHEMOMETRIC APPROACHES TO MAXIMIZE INTERPREPATION OF GC – TOFMS AND GC × GC – TOFMS DATA

Brendon A. Parsons, David K. Pinkerton, Brian D. Fitz, Brooke C. Reaser, and Robert E. Synovec
Department of Chemistry, Box 351700, University of Washington, Seattle, WA 98198, USA
synovec@chem.washington.edu

For the analysis of complex samples, gas chromatography and comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC – TOFMS and GC × GC – TOFMS) are powerful instrumental platforms. Chemometric approaches play a pivotal role in data analysis. We are exploring the development of several approaches to maximize data interpretation. Recent advances in the Fisher-ratio (F-ratio) analysis will be presented, a data mining technique to discover compounds that distinguish sample classes. Specifically, a tile-based F-ratio approach is presented that substantially improves chemical selectivity in the discovery process. PARAFAC has been shown to confidently analyze GC × GC – TOFMS data for many analytical studies, readily providing compound deconvolution, identification, and quantification. We have been exploring the theoretical and experimental boundaries for confident application of PARAFAC. Finally, we have been exploring new approaches to substantially improve chemometric resolution of GC – TOFMS and GC × GC – TOFMS data sets.

K5

A BIRD IN HAND: THE CENTRAL ROLE OF CHEMOMETRICS IN THE EXPLODING FIELD OF HANDHELD INSTRUMENTATION

Christopher D. Brown, PhD
908 Devices, 27 Drydock Ave., Boston MA 02210
chris@908devices.com

Ten years ago handheld spectrometers remained the realm of Star Trek fiction. Today, you can purchase a handheld Raman, FTIR, NIR, XRF and MS system, and fit all of them in your carry-on bag. It is estimated that last year Raman and XRF handhelds outsold laboratory systems 3-to-1, and yet the proliferation of these systems, and their enabling technologies remains unknown to most analytical chemists and chemometricians.

The use cases for these handheld systems pose a fascinating opportunity in chemometrics, encompassing hardware automation, signal processing and decision theory, and in almost every way the opposite of textbook chemometric problems: the problem space is unbounded and ill-defined, the environment is dynamic and system non-stationary, and perhaps most critically the overarching objective is not to convert captured data to information, but to capture data in a way that dictates action.

I'll discuss the emergence of this remarkable hardware, and my perspective on the analytic approaches crucial for their past and continued success.

K6

OPPORTUNITIES OF COMPOSITIONAL DATA ANALYSIS IN CHEMOMETRICS

P. Filzmoser

Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria

P.Filzmoser@tuwien.ac.at

Compositional data are defined as data where the relevant information is in the ratios between the variables. Accordingly, a statistical analysis of the raw data values is inappropriate, but the focus has to be on analysing the ratios. This can conveniently be done by applying a so-called log-ratio transformation to the data matrix, and proceeding with the statistical method on the transformed data. An interpretation of the results usually has to be based on back-transformed values in the original space.

We demonstrate the usefulness of this approach to chemical compositions, and outline the differences to the standard approach of analysing raw data. We also emphasize the limitations of this approach for high-dimensional data as they are typical in chemometrics, in particular with respect to variable selection, and provide possible solutions.

O1

TOWARDS A DATA PRE-PROCESSING STRATEGY

Jan Gerretzen[a,b]; Jeroen J. Jansen[a]; Ewa Szymańska[a,b]; Jacob Bart[c]; Henk-Jan van Manen[c]; Lutgarde M.C. Buydens[a]

[a]Radboud University Nijmegen, Institute for Molecules and Materials, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

[b]TI-COAST, P.O. Box 18, 6160 MD Geleen, The Netherlands

[c]AkzoNobel, Supply Chain, Research & Development, Zutphenseweg 10, 7418 AJ Deventer, The Netherlands

l.buydens@science.ru.nl

Data pre-processing is an essential part of data analysis, which aims to remove unwanted variation from the data in order to highlight variation of interest. Recently, we have shown that multivariate model performance is heavily influenced by pre-processing [1]. Current pre-processing selection approaches are time consuming and/or based on arbitrary choices that do not necessarily improve the model [1]. It is thus highly desired to come up with a novel, generic pre-processing selection procedure that leads to a proper pre-processing within reasonable time. In this presentation, a possible way to come up with such a pre-processing selection strategy will be presented.

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

Reference

[1] Engel, J. et al. *Trends in Anal. Chem.*, 2013, **50**, pp. 96-106

O2

STATISTICAL HOMOGENEOUS CLUSTER SPECTROSCOPY (SHOCSY): AN OPTIMISED STATISTICAL APPROACH FOR CLUSTERING OF ^{1}H NMR SPECTRAL DATA TO IMPROVE CLASSIFICATION AND ROBUST BIOMARKERS SELECTION

Xin Zou[1], Elaine Holmes[2, 3], Jeremy K Nicolson[2, 3], Ruey Leng Loo[1, 2]

[1] *Medway School of Pharmacy, Universities of Kent and Greenwich, Chatham Maritime, Kent, UK*

[2] *Section of Biomolecular Medicine, Department of Surgery and Cancer, Imperial College London, UK*

[3] *MRC-HPA Centre for Environment and Health, Imperial College London, UK*

r.loo@kent.ac.uk

Background: Chemometric analysis methods, such as Orthogonal Partial Least Square Discrimination Analysis (OPLS-DA), have been widely used in metabonomic studies for extracting metabolic biomarkers. However, it may generate sub-optimal results when the samples in each biological class are highly heterogeneous. Here, we propose a novel method, Statistical HOmogeneous Cluster SpectroscopY (SHOCSY), to reduce the variation within biological datasets and to enhance the biomarker selection process.

Method: In SHOCSY, i) supervised OPLS-DA is first applied to identify the potential metabolic biomarker features. ii) K-means approach is applied to cluster the spectra with similar biomarker features together. iii) An enrichment test is employed to associate the clusters to specific biological classes. This identifies the spectra that constitute homogenous cores for the specific biological class. The three steps are performed iteratively until the predictive ability of the OPLS-DA model built on the homogenous spectra of biological classes is maximal. The SHOCSY algorithm has been evaluated using a simulated and an animal ^{1}H NMR spectra datasets. The simulated data was designed to emulate control and Paraquat toxicity with 30 spectra in each class.

Results: SHOCSY can distinguish the homogeneous responders from those idiosyncratic ones. Using the homogeneous subsets, the model predictive ability is improved; and allows the biomarkers not be obscured by idiosyncratic responders.

Conclusion: SHOCSY iteratively 'learns' the metabolic features best representing the biological classes and identifies samples lacking these features. This enables robust biomarker selection process. SHOCSY has wide applicability and can be used to analyses other data, e.g., Mass Spectrometry (MS).

O3

TOWARDS THE DISEASE BIOMARKER IN AN INDIVIDUAL PATIENT USING STATISTICAL HEALTH MONITORING

Jasper Engel[a,b]; Lionel Blanchet[a,c]; Udo F.H. Engelke[b]; Ron A. Wevers[b]; Lutgarde M.C. Buydens[a]

[a] *Radboud University Nijmegen, Institute for Molecules and Materials, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands*

[b] *Laboratory of Genetic Endocrine and Metabolic Diseases at the Department of Laboratory Medicine, Radboud University Medical Centre, Geert Grooteplein 10, Nijmegen, the Netherlands*

[c] *Department of Biochemistry, Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Centre, Geert Grooteplein 10, Nijmegen, the Netherlands*

j.engel@science.ru.nl

In metabolomics, identification of complex diseases is often based on application of (multivariate) statistical techniques to the data. Commonly, each disease requires its own specific diagnostic model, separating healthy and diseased individuals, which is not very practical in a diagnostic setting. Additionally, for orphan diseases such models cannot be constructed due to a lack of available data. An alternative approach adapted from industrial process control is proposed in this study: statistical health monitoring (SHM).

In SHM the metabolic profile of an individual is compared to that of healthy people in a multivariate manner. Abnormal metabolite concentrations, or abnormal patterns of concentrations, are indicated by the method. Subsequently, this biomarker can be used for diagnosis. A tremendous advantage here is that only data of healthy people is required to construct the model. The method is applicable in current – population based - clinical practice as well as in personalized health applications.

In this study, SHM was successfully applied for diagnosis of several orphan diseases as well as detection of metabotypic abnormalities related to diet and drug intake.

O4

BIOMARKER IDENTIFICATION: WHAT CAN GO WRONG AT THE DATA NORMALIZATION STEP?

P. Filzmoser[1], B.Walczak[2]
[1]*Institute of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria*
[2]*Institute of Chemistry, University of Silesia, Katowice, Poland*
beata.walczak@us.edu.pl

Nowadays, hyphenated chromatographic techniques (such as, e.g., HPLC-DAD, LC-MS, or UPLC-MS) have become standard analytical tools for studying complex biological samples. Generated chromatograms (fingerprints) require extensive pre-processing, prior to comparative statistical analysis. The preprocessing step consists of signals enhancement (de-noising and background elimination), warping and removal of the so-called 'size effect', associated with a different sample volume and/or sample concentration. Our study focuses on the removal of 'size effect', i.e., on the data normalization. Although there are many methods which can be applied at this step of data analysis, the choice of a right method is not obvious, and it depends on data characteristics and the problem at hand. It can happen that the methods which are considered as alternative approaches lead to different results and different conclusions. In this study the influence of the data normalization methods on identification of 'biomarkers' is discussed. Among the approaches considered, there are the log-ratios based methods (belonging to the arsenal of the compositional data analysis tools).

Our study demonstrates that models with good performance can lead to wrong conclusions. This means that although it is possible to identify features which ensure good model performance, identification thereof can be meaningless and many hypotheses based on identification of 'biomarkers' should be revisited. It also means that performance of a model should not be used to guide the choice of a normalization method.

O5

FLOW CYTOMETRY FOR COMPREHENSIVE DISEASE DIAGNOSIS, USING NOVEL DEDICATED CHEMOMETRICS

Jeroen J. Jansen[1], Bart Hilvering[2], Leo Koenderman[2], Oscar van den Brink [3], Lutgarde M.C. Buydens[1]

[1] *Analytical Chemistry, IMM , Radboud Universiteit Nijmegen*
[2] *Respiratory Medicine, University Medical Centre Utrecht*
[3] *TI-COAST, Science Park 904 Amsterdam*
jj.jansen@science.ru.nl

Flow Cytometry (FC) provides a wealth of real-time data on the status of the immune system. It analyses fluorescent labels on surface markers expressed on the wall of individual, suspended cells (*e.g* blood). Technological developments in laser and FC technology have considerably increased the diversity in surface markers that can be simultaneously analysed on one cell, the speed at which multi-cellular samples can be measured and thereby the amount of multivariate data in a typical FC dataset. The great potential this brings for (1) discovering immunological disease mechanisms and (2) comprehensive and personalized disease diagnosis however requires a novel, advanced and dedicated chemometrics approach.

Conventionally, Flow Cytometry analyses the cellular expression of one or two specific properties (*e.g.* CD4 in the diagnosis of AIDS). We however developed the FLow cytometric Orthogonal Orientation for Diagnosis (FLOOD), for a specifically FC-based disease diagnosis and identification based on many (relations between) surface markers involved in the disease mechanism, for all cells in a blood sample. FLOOD follows familiar concepts from Process-Analytical Technology (PAT), to *first* model the cellular properties from known healthy individuals. *Then* the model 'mis-fit' for analogous profiles of diseased individuals provides a quantitative disease diagnosis. The same mis-fit will contain disease-specific patterns that can be separately modelled for further biomedical understanding based on the entire patient, from which the *individual cells* can be pin-pointed with a disease-related surface marker pattern. This breakthrough functionality of FLOOD unlocks the full technological potential of FC for quantitative, personalized and mechanism-based disease diagnosis.

O6

MULTIVARIATE TECHNIQUES FOR REAL-TIME, IN-SITU TISSUE IDENTIFICATION DURING SURGERY USING RAPID EVAPORATIVE IONIZATION MASS SPECTROMETRY

Julia Balog[1,2], Laszlo Molnar[2], Peter Varga[2] and Zoltan Takats[1]
[1] *Department of Surgery and Cancer, South Kensington Campus, Imperial College London, London SW7 2AZ.*
[2] *Medimass ltd, 2. Remenyi Ede street, Budapest, Hungary 1033.*
j.balog@imperial.ac.uk

Rapid Evaporative Ionization Mass Spectrometry (REIMS) is an emerging technique allowing near real-time, in-vivo characterization of tissue by mass spectrometric analysis of the aerosol released during electrosurgical dissection. The tissue characterization workflow includes the construction of a tissue specific spectral database and using a multivariate classification algorithm and spectral identification algorithm. Our aim is to separate healthy and cancerous tissue, to characterize the tumor and detect metastases in lymph nodes based on the REIMS fingerprint of each tissue type. In this study 4 different datasets were used containing 3,158 spectra. The acquired spectra were subjected to different multivariate statistical methods after simple pre-processing, in order to compare the performance of different linear (including principal component analysis (PCA), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA, OPLS-DA)) and non-linear classification methods (support vector machines, neural networks, random forests, K* algorithm). The results were compared using leave-one patient out cross-validation. The selected algorithm comprising of PCA and LDA was then further tested on a dataset containing 1,007 spectra. The specificity and sensitivity of correct tissue identification was 98.6% and 99.3%. PCA-LDA models were then built for real-time detection of different tissues, and tested with an independent validation set resulting in 85-99.9% correct identification. Statistical analysis of the various tissues proved that the mass spectrometer coupled intelligent surgical knife is capable of providing real-time identification of intra-operative pathology which could significantly influence ‘on table’ decision-making.

O7

WEIGHTED MONTE CARLO SAMPLING FOR VARIABLE SELECTION BASED ON MODEL POPULATION ANALYSIS

Qing-song Xu[1], Bai-chuan Deng, Yun-yong Huan[2], Hong-dong Li[2], Yi-zeng Liang[2]

[1]School of Mathematics & Statistics, Central South University, 932 Lushannanlu, Changsha 410083, P.R. China

[2]Research center of modernization of traditional Chinese medicines, Central South University, 932 Lushannanlu, Changsha 410083, P. R. China

qsxu@csu.edu.cn

Selecting a small subset of informative variables plays an important role in regression and discrimination for chemical data. By weighted Monte Carlo sampling in the variable space, a large number of sub-models can be established on one dataset. The model population analysis (MPA) is developed by statistically analyzing some parameters (regression coefficients, prediction errors, etc.) of these sub-models. A measure of importance is yielded for each variable. The Monte Carlo sampling gives a high weight to the important variable, resulting in the larger probability to select vital variable. The proposed method alleviates some of the limitations of other variable selection methods such as the uninformative variable elimination, lasso and related methods noted in chemometrics: it may select the different variables even with one sample changed in dataset. The resulting prediction accuracy is competitive or superior compared to the alternatives. We illustrate the proposed method by several chemometric data.

O8

(sMC) SIGNIFICANT MULTIVARIATE CORRELATION: EVALUATING VARIABLE IMPORTANCE IN PARTIAL LEAST SQUARES IN REGRESSION AND CLASSIFICATION

Thanh N. Tran[1,3,*], Nelson Lee Afanador [2,3], Lutgarde M.C Buydens[3], Lionel Blanchet[3,4]

[1]*Center for Mathematical Sciences, Merck, Sharp, & Dohme, Oss, the Netherlands*

[2]*Center for Mathematical Sciences, Merck, Sharp, & Dohme, West Point, PA, USA*

[3]*Institute for Molecules and Materials, Analytical Chemistry, Radboud University Nijmegen, the Netherlands*

[4]*Department of Biochemistry, Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen, the Netherlands*

thanh.tran@merck.com,thanh.tran@science.ru.nl

Identifying important variables in PLS regression has been a difficult task due to the complexity of PLS with varies combinations of rotations and projections in different sub-spaces. This presentation will provide a better understanding of the relationship between the PLS regression coefficients, orthogonally decomposed variances, and the impact of biased regression in PLS. This better understanding will help illustrate the reasons for the difficulty in the interpretation of the important variables in PLS. With this new understanding in mind, the authors propose the following new method, *significant multivariate correlation* (sMC), for statistically assessing variable importance in PLS regression and classification.

In its application to both simulated and real data sets (NIR spectroscopy and NMR metabolomics), the outstanding properties of sMC over several commonly used methods, such as regression coefficient bootstrapped confidence intervals, Variable importance in the projection (VIP) and Selectivity Ratio (SR), will be demonstrated.

O9

COMPARISON OF FOUR DIFFERENT FEATURE SELECTION METHODS IN SPECTRAL DATA: DISCRIMINATION FOR A BINARY OUTCOME

H. Nocairi, V. Michaut and F. Leroy
L'Oréal Research and Innovation, Aulnay-Sous-Bois, France
hnocairi@rd.loreal.com

The methods of classification and discrimination between two or several groups are of paramount importance in analyzing the spectral data of vibrational type spectroscopy (Infrared, Raman, NIRS, ultraviolet-visible (UV-vis), or nuclear magnetic resonance (NMR) …). However, the issue of spectra is very particular since they frequently show uncontrolled variations of global intensity eing affected by both the length of the optical route through the sample, and the physical properties of the sample (size and distribution of particles ...). As a consequence, specific methods have been developed for spectrum analysis. The most known statistical method of discrimination is Fisher's Canonical Discriminant Analysis (FCDA), but not appropriate to analyze spectral data because of multicollinearity between wavelengths.
Since recent years, a wide set of alternative methods has emerged. Some of these are developed procedures to identify wavelengths that contribute to provide valuable informations. Wavelength The selection of wavelengths or spectrum zones is a critical step in spectral data analysis.

In this presentation, the performances of three different feature selection methods for discriminating between two groups were compared. The feature selection algorithms included Sparse Partial Least Squares Discriminant Analysis (Sparse-PLSDA), Evolving Window Zone Selection method and interval partial least squares regression (I-PLS). They were used for being compared to our approach based on PLSDA. The potential of our approach is discussed and the four methods of analysis are compared through a real data set.

O10

GENETIC HYBRIDATION (HYBRIDGEN): A COOPERATIVE COEVOLUTION ALGORITHM FOR VARIABLE SELECTION

C. Cernuda[1], E. Lughofer[1], P. Hintenaus[2], W. Märzinger[3], J. Kasberger[4]

[1] Department of Knowledge-Based Mathematical Systems, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040, Linz, Austria [2] Software Research Center, Paris Lodron University Salzburg, Kapitelgasse 4-6, 5020, Salzburg, Austria

[3] i-RED Infrarot Systeme GmbH, Hafenstrasse 47-51, 4020, Linz, Austria

[4] Recendt GmbH, Science Park 2, Altenbergerstrasse 69, 4040, Linz, Austria

carlos.cernuda@jku.at

Variable selection methods are crucial for reaching accurate and computationally feasible final models due to the size of the data we handle. Metaheuristics have been widely used in Chemometrics for that purpose [1,2,3]. Their main pros are simple implementations and diverse searching space coverage, and their cons are the impossibility of ensuring global maxima, possible slow convergence, and risk of getting trapped in local minima. We propose a procedure that keeps the pros and avoids the cons.

HybridGen has been developed following the concept of cooperative coevolution [4]. Its goal is to combine, by means of genetic operators, the searching power of the pure entities, coming from different metaheuristics, leading to hybrids which are better than their predecessors. A prerequisite is isomorphic coding, thus entities can be compared and combined. The general idea is a cyclic process consisting on i) letting each metaheuristic run on a subpopulation, ii) coevolve all the entities together as a whole big population using the genetic operators, and iii) split the coevolved population onto subpopulations, going back to i). We act on all three steps, achieving better solutions with faster convergence and escaping from local minima, by i) controlling each metaheuristic separately, ii) experimenting with the genetic operators during the overall process, and iii) modifying the way the subpopulations are obtained from the joint general population.

The procedure was tested on real-world NIR spectral data from melamine resin and viscose production coevolving ant colony optimization and particle swarm optimization. We achieve better solutions, with faster convergence, than the single metaheuristics.

[1] Niazi, A., Leardi, R. (2012), Genetic algorithms in Chemometrics. J. Chemometrics, 26(6): 345-351

[2] Shamsipur, M., Zare-Shahabadi, V., Hemmateenejad, B., Akhond, M. (2006), Ant colony optimisation: a powerful tool for wavelength selection. J. Chemometrics, 20: 146–157

[3] Cernuda, C., Lughofer, E., Hintenaus, P., Märzinger, W. (2014), Enhanced genetic operators design for waveband selection in multivariate calibration based on NIR spectroscopy. J. Chemometrics, DOI: 10.1002/cem.2583

[4] Luke, S. (2013), Essentials of Metaheuristics (2nd Edition). Lulu Publishers, chap. 6, pp. 109-131

O11

PETROLEOMICS BY ELECTROSPRAY IONIZATION FT-ICR MS ALLIED TO PLS WITH VARIABLE SELECTION METHODS: PREDICTION OF TOTAL ACID NUMBER OF CRUDE OILS

Luciana A. Terra[a], Paulo R. Filgueiras[a], Lílian V. Tose[b], Wanderson Romão[b,c], Douglas D. de Souza[d], Eustáquio V. R. de Castro[b], Lize M. S. L. de Oliveira[e], Júlio C. M. Dias[e], Ronei J. Poppi[a]

[a]*Institute of Chemistry, University of Campinas, Campinas, SP, Brazil*
[b]*LabPetro-Federal University of Espirito Santo, Vitoria, ES, Brazil*
[c]*Federal Institute of Espirito Santo, Vila Velha, ES, Brazil*
[d]*Institute of Physics "GlebWataghin", University of Campinas, Campinas, SP, Brazil*
[e]*CENPES/PETROBRAS, Rio de Janeiro, RJ, Brazil*
lucianaassisterra@gmail.com

Negative ion mode electrospray ionization, ESI(-), with Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was coupled to partial least squares (PLS) regression and the variable selection methods to estimate the total acid number (TAN) of thirty-four Brazilian crude oil samples. ESI(-)-FT-ICR mass spectra presents a power of resolution corresponding to approximately 500.000:1 and a mass accuracy less than 1 ppm, producing a data matrix containing a total number of variables of 5703. They correspond to heteroatom-containing species detected as deprotonated molecules, $[M-H]^-$ ions, being identified primarily as naphthenic acids, fenols and carbazole analog species. TAN values for all samples range from 0.06 to 3.61 mg of KOH g^{-1}. A PLS model was built using twenty-five samples for calibration and nine for validation. To facilitate the spectral interpretation, three methods of variable selection were studied: variable importance in the projection (VIP), interval partial least squares (iPLS) and elimination of uninformative variables (UVE). UVE proved the most suitable method for variable selection, reducing the number of the variables from 5703 to 183 and producing a root mean square error of prediction (RMSEP) of 0.32 mg of KOH g^{-1}. By reducing the size of the data it was possible to relate the selected variables with their corresponding molecular formulas, thus identifying their relation with the TAN.

O12

rPLS FOR VARIABLE SELECTION IN CLASSIFICATION PROBLEMS

Åsmund Rinnan and Søren B. Engelsen
Department of Food Science, Faculty of Science, University of Copenhagen, Denmark
aar@food.ku.dk

Rinnan et al. [1] recently suggested a novel technique for variable selection for predictive models. This technique is a recursive technique where the regression coefficient successively is multiplied by the original X matrix in order to focus the prediction on the most important variables of the dependent variables in X. This study will focus on the use of the same technique on classification problems, both in the regular case of PLS-DA [2], but also in the framework of ECVA [3] classification. In the ECVA algorithm the idea can readily be added in two different steps of the algorithm: inside the PLS loop that is inside the ECVA, but also on the outer canonical weights estimated as the output of the algorithm. The manuscript will thus give an in depth view of how the recursive technique can be applied to classification techniques, and discuss the strengths and weaknesses of the different implementations of the idea. The datasets used in the study will include NIR and NMR metabolomics data.

References

[1] Rinnan, Å.; Andersson, M.; Ridder, C.; Engelsen, S.B. Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS, Journal of Chemometrics 2013, DOI: 10.1002/cem.2582
[2] Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. Journal of Chemometrics 1987, 1, 185—196.
[3] Nørgaard, L.; Bro, R.; Westad, F.; Engelsen, S. B. A modification of canonical variate analysis to handle highly collinear multivariate data. Journal of Chemometrics 2006, 20, 425—435.

O13

AN EFFICIENT CHEMOMETRIC STRATEGY FOR SIZE REDUCTION OF MULTI CAPILLARY COLUMN - ION MOBILITY SPECTROMETRY (MCC-IMS) DATA

Ewa Szymańska[1,2], Emma Brodrick[3], Mark Williams[3], Anthony N. Davies[3,4], Henk-Jan van Manen[4], Lutgarde M.C. Buydens[2]

[1]TI-COAST, P.O. Box 18, 6160 MD Geleen, The Netherlands

[2]Radboud University Nijmegen, Institute for Molecules and Materials (IMM), P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

[3]School of Applied Sciences, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, CF37 1DL, UK

[4]AkzoNobel N.V., Supply Chain, Research & Development, Expert Capability Group - Measurement & Analytical Science, P.O. Box 10, 7400 AA, Deventer, The Netherlands

E.Szymanska@science.ru.nl

Ion mobility spectrometry (IMS) is increasingly in demand for medical applications, process and environmental control, as well as food quality and safety. In all these applications the breath or air samples are extremely complex mixtures often including numerous volatile organic compounds (VOCs). Through a combination of IMS with multi-capillary column chromatography (MCC-IMS), a fast, robust, non-invasive and easy-to-use system for qualitative and quantitative analyses of VOCs at low ppb and ppt levels is now available. However, MCC-IMS spectra present chemometric challenges because of the high dimensionality (more than a million variables per sample) and redundancy of information (several variables associated to a single VOC).

In our study, an effective data analysis strategy has been developed and employed in the analysis of MCC-IMS spectra from 264 breath and ambient air samples. This strategy includes several data pre-processing (background correction, alignment, denoising and compression by wavelet transform) and variable elimination tools (mask construction and sparse partial least-squares-discriminant analysis (sparse-PLS-DA)). The influence and compatibility of data reduction tools was studied by applying different combinations to the data sets. Loss of information after pre-processing was evaluated, *e.g.*, by comparing the performance of classification models for different classes of samples. Finally, interpretability of classification models was evaluated and regions of spectra that are the related to the identification of potential analytical biomarkers were successfully determined. This work will greatly enable the standardization of analytical procedures across diverse instrumentation types promoting the adoption of this new technology in a wide range of diverse application fields.

O14

VARIBLE IMPORTANCE IN PLS IN THE PRESENCE OF AUTOCORRELATED DATA: CASE STUDIES IN INDUSTRIAL MANUFACTURING PROCESSES

Nelson Lee Afanador [1,3], **Thanh N. Tran**[2,3], **Lutgarde M.C. Buydens**[3]
[1]*Center for Mathematical Sciences, Merck, Sharp, & Dohme, West Point, PA, USA*
[2]*Center forMathematical Sciences, Merck, Sharp, & Dohme, Oss, Netherlands*
[3]*Institute for Molecules and Materials, Analytical Chemistry, Radboud University Nijmegen, Netherlands*
nelson_afanador@merck.com

An integral part of interpreting atypical process performance in manufacturing processes is a multivariate understanding of process parameters and their relationship to a product's critical quality attributes. In this endeavor, the use of Partial Least Squares (PLS) has greatly advanced the analysis of data that exhibits a high level of multicollinearity, but has not fully explored the impact to a model in the presence of autocorrelation in the manufacturing time domain, wherein a current observation is correlated to some degree with the previous observation(s). This autocorrelation provides an additional challenge to understanding model performance and the associated important variables. For example, when this autocorrelation is predominantly positive, it will have the effect of biasing downward the mean squared error (MSE), hence methods that employ the MSE for important variable determination may see an increase in the false positive rate. This paper introduces the application of an autocorrelation correction formulation to PLS in an attempt to address this concern. Recently, the Significant Multivariate Correlation (sMC) method was introduced and displayed favorable results in identifying important variables when compared to other commonly used importance metrics. Given the reliance of the sMC on a correct estimate of a MSE for important variable determination, we will compare the performance of the correction factor in this method to various important variable selection methodologies in the presence of both multicollinearity and autocorrelation, while varying signal-to-noise ratio. This pioneer study in this area will help further promote research in the future.

O15

EXPLORING AND SEGMENTATION OF HYPERSPECTRAL IMAGES USING SPECTRAL FEATURES CALCULATED FOR PIXEL GROUPS

Sergey Kucheryavskiy
Department of Biotechnology, Chemistry and Environmental Engineering,
Aalborg University, Niels Bohrs vej 8, 6700, Esbjerg, Denmark
svk@bio.aau.dk

Hyperspectral images contain large amount of data, therefore specific methods for processing and analysis are needed to reveal useful information from the images. Chemometrics provides such methods gathered under the umbrella of multivariate image analysis (MIA).

One of the MIA's drawbacks is that most the methods consider image pixels as independent objects without taking into account their spatial relations. So the image is in fact treated as a large set of spectra. In the meantime, methods that use the spatial information are also in developing. One of them [1] was recently proposed for classification and discrimination of objects on hyperspectral images by calculating spectral features taking into account all pixels belonging to a particular object.

In the present work the method is further developed in order to be applicable to cases, where no objects can be detected/segmented a priory. Two approaches are suggested — for high-resolution images features are calculated for adjacent pixel blocks, while if resolution is limited the spectral features are calculated for every pixel by considering the pixel's neighborhood. Both approaches can be used for exploratory analysis of hyperspectral images as well as for segmentation purposes. Several real case examples will be shown to demonstrate the performance of the approaches.

References:

[1] S. Kucheryavskiy, Chemometrics and Intelligent Laboratory Systems, vol. 120, 126-135 (2013).

O16

STRATEGIES FOR SINGLE-MOLECULE FLUORESCENCE IMAGING DATA ANALYSIS

Cyril Ruckebusch,[1] Romain Bernex,[1] Michel Sliwa,[1] Franco Allegrini,[2] Johan J. de Rooi,[3] Paul H.C. Eilers[3]

[1]*LASIR CNRS, Université Lille Nord de France, Sciences et Technologies, France*
[2]*Departamento de Química Analítica, Instituto de Química de Rosario (IQUIR-CONICET), Universidad Nacional de Rosario, Argentina*
[3]*Department of Biostatistics, Erasmus Medical Center, the Netherlands.*
Cyril.ruckebusch@univ-lille1.fr

Functional super-resolution microscopy in biological imaging can be achieved by different methods and data analysis may involve data fitting, signal deconvolution or other alternatives. The most straightforward single-molecule fluorescence imaging techniques rely on the sparse activation of individual fluorophores within a sample. The activated fluorophores are ideally observed individually on different images and their localization process is repeated over thousands of frames. Summing up all these positions provides the rendered image of the sample. However, the requirement to work in conditions of low density of active fluorophores goes along with some limitations regarding potential applications. The development of methodologies capable of detecting emitters in non-sparse conditions, i.e. when active fluorophores are numerous and their emissions are highly overlapping, as in living cells imaging, is still a main issue.

Chemometrics provides alternative strategies that have not yet been investigated to tackle the difficulty of single-emitters localization in high-density labeled samples. We propose an approach based on the combination of a dissimilarity criterion (originally developed for peak purity assessment in spectroscopic mixtures) and Gaussian point spread function fitting. Also, we highlight the potential of penalized estimation using a L_0-norm penalty for sparse deconvolution. We present the results obtained for live cells expressing a construct encoding a fluorescent protein. We discuss in terms of resolution enhancement (both methods provide sub-diffraction limit optical resolution), background elimination and contrast enhancement.

O17

MONITORING AND MODELING THE CHEMICAL ADAPTATION OF MICROALGAE CELLS TO SHIFTING ENVIRONMENTAL CONDITIONS

Frank Vogt
Department of Chemistry, University of Tennessee, 552 Buehler Hall
Knoxville, TN 37996-1600, USA
fvogt@utk.edu

With an increase of industrialization, the production of CO_2 is rising and the fate of this greenhouse gas has become a serious concern. Research regarding CO_2 sequestering has also focused on phytoplankton as about half of the global primary carbon production is due to algal photosynthesis. It has been hypothesized that a relation between environmental conditions and the chemical composition of microalgal biomass exist.

FTIR spectroscopy has become a versatile tool for studying microalgae cells; however, for monitoring adaptation processes, the cells need to be kept in an aqueous environment and thus, FTIR-ATR spectroscopy has been employed. Obtained data contain information about (i) the build-up of a mciroalgal biofilm and (ii) its chemical adaptation to changing nutrient conditions. To investigate chemical processes within the biomass, a model has been derived which describes wavenumber-dependent the loss of reactants and the formation of intermediate and final products. For the given data, the wavenumber- and time-dependent, nonlinear model function comprises >2,000 fit parameters and is fitted to time series of FTIR-ATR spectra containing >61,000 data points per nutrient situation. In the presentation's first part, a novel approach for nonlinear least-squares is introduced by means of which the impediment of local minima of the Residual Sum of Squares can be considerably mitigated. In the second part, this method is applied to derive model parameters describing how microalgae cells spectro-chemically adapt to different nutrient situations. Results from this nonlinear hard-modeling are used to assess the impact of nutrient conditions on microalgae's sequestration of inorganic compounds.

O18

MULTISAMPLE AND MULTITECHNIQUE IMAGE ANALYSIS: JOINING IMAGES WITH DIFFERENT SPATIAL PROPERTIES

Anna de Juan[1], Sara Piqueras[1,2], Marcel Maeder[3], Víctor Olmos[1], Romà Tauler[2]

[1]*Chemometrics group, Universitat de Barcelona. Diagonal, 645, 08028 Barcelona, Spain*

[2]*IDAEA-CSIC. Barcelona, Spain*

[3]*Dept. Chemistry, The University of Newcastle, Newcastle, Australia*

anna.dejuan@ub.edu

Hyperspectral images are massive measurements containing spectral and spatial information. Resolving (unmixing) a hyperspectral image is done to understand the nature and spatial distribution of sample constituents. Multiset analysis (multisample and/or multitechnique) is the proper choice when several images need to be treated simultaneously [1].

Multisample image analysis enables coupling images obtained with the same technique with different size, geometry and, interestingly, different spatial resolution, because the pixel direction can be completely different among images. This last option allows combining full images of objects with zoomed Regions Of Interest (ROIs) with a very high spatial resolution.

Multitechnique image analysis is a challenging problem because it needs a common pixel mode among imagesand, therefore, correcting translation/rotation among images and balancing differences in spatial resolution is required. Downsampling (binning) images with high resolution produces a loss of spatial detail, whereas oversampling (interpolation) images of low resolution implies an artificial assumption of the spatial behavior of sample constituents. In this sense, we want to propose the use of a special kind of multiset structure, formed by an L-shaped data arrangement, in which augmentation in the row direction is performed with images with a common pixel mode with the lowest spatial resolution, and in the column direction with the highest resolution image appended with its binned version. Analyzing this data arrangement requires the recently introduced variant of MCR-ALS algorithm for incomplete data sets [2], but respects the real spatial resolution of each image taking advantage of the combined information of the different spectroscopic techniques.

References

[1] A. de Juan, S. Piqueras, M. Maeder, T. Hancewicz, L. Duponchel, R. Tauler. Chemometric tools for image analysis, in Image Analysis in Infrared and Raman Spectroscopic Imaging,Wiley-VCH (2nd ed.). 2014.

[2] M. Alier and R. Tauler. Chemom. Intell. Lab. Sys. 127 (2013) 17-28.

O19

IDENTIFICATION OF DOCUMENT FORGERY BY ADDING TEXT USING NIR HYPERSPECTRAL IMAGE AND CHEMOMETRICS

Carolina S. Silva[a], M. Fernanda Pimentel[a], Ricardo S. Honorato[b], Celio Pasquini[c], José M. Prats-Montalbán[d], Alberto Ferrer[d]

[a]*Department of Chemical Engineering, Universidade Federal de Pernambuco, Rua Prof. Arthur de Sá S/N, Cidade Universitária, 50740-521 - Recife, PE, Brazil*

[b]*Department of Federal Police,Superintendência Regional em Pernambuco, Av. Cais do Apolo, 321, Bairro do Recife 50030230 - Recife, PE, Brazil*

[c] *Chemistry Institute, Department of Analytical Chemistry, Universidade Estadual de Campinas, Cidade Universitária 13084-971 - Campinas, SP, Brazil*

[d]*Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Camino de Vera s/n, Edificio 7A, 46022 Valencia, Spain*

carolina.santossilva@ufpe.br, mfp@ufpe.br, saldanha.rsh@dpf.gov.br, pasquini@iqm.unicamp.br, jopramon@eio.upv.es, aferrer@eio.upv.es

One major problem at police departments is document forgery. One typical forgery is the addition of text over original documents. In this work, this fraud was simulated using one pen to write a number on a sheet of bank checks that was subsequently modified using another pen. Pens with the same shades of black and tip thickness, of different brands and types were employed to produce surrogate samples. Twenty-two samples were produced simulating this type of forgery. Near Infrared Hyperspectral Image Analysis (NIR-HI) was used, since it is necessary to consider a fast and non-destructive method of analysis. Different pre-processing techniques were tested. PCA and MCR-ALS were carried out on the image spectra to identify whether more than one pen had been used to produce the document. SNV showed the best results as a pre-processing technique. A PCA and MCR-ALS models were built. Analyzing the images retrieved from the scores values and the distribution maps, only 4 out of 22 fake samples were not discriminated. In those 4 exceptions, it was not possible to differentiate the inks from the paper due high absorbance of cellulose in the NIR range, making difficult to differentiate inks and paper spectra. Nevertheless, 82% of the forged samples were successfully identified, demonstrating the potential of NIR-HI associated with chemometrics to approach this type of forensic issue.

O20

POLYHYDROXYALKANOATE GRANULES QUANTIFICATION IN MIXED MICROBIAL CULTURES: SUDAN BLACK B *VERSUS* NILE BLUE A STAINING

Daniela P. Mesquita[1], A. Luís Amaral[1,2], Eugénio C. Ferreira[1]

[1]*Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal*

[2]*Instituto Politécnico de Coimbra, ISEC, DEQB, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal*

Polyhydroxyalkanoates (PHAs) are intracellular granules found in a wide variety of microorganisms under limited nutrient conditions when carbon source is available in excess. These polymers, usually from lipid nature, are used as carbon and energy sources for metabolic synthesis and growth. Despite the important role of PHAs in cell physiology, they are regarded as potential substitutes of traditional petrochemical plastics with the additional advantage of being completely biodegradable and produced from mixed microbial cultures (MMC). PHA quantification is regularly accomplished using a digestion step prior to chromatography analysis which is a labor and time-consuming technique.

To overcome these limitations in polymers quantification, the present work investigates two methods for PHA granules identification based on quantitative image analysis (QIA) procedures in an enhanced biological phosphorus removal (EBPR) system operated for three months. MMC were analyzed for PHA granules detection by Sudan Black B (SBB) and Nile Blue A (NBA) staining using bright-field and epifluorescence microscopy, respectively. The captured color images were evaluated through QIA and the image analysis data was further processed using multivariate statistical analysis. Quite satisfactory partial least squares (PLS) regressions (R^2) of 0.85 for NBA and 0.86 for SBB were established between PHA concentrations predicted from QIA parameters and determined by the standard analytical method. Although SBB staining procedure was found to provide a somewhat higher estimation of PHA concentrations in MMC, the consistency between PLS results allowed to conclude that both SBB and NBB staining methods combined with QIA procedures demonstrated the capability to estimate PHA concentrations. Concluding, both staining procedures are promising alternative for a faster PHA assessment relatively to the laborious standard PHA quantification.

O21

VINEYARD HARVEST FORECASTING USING HYPERSPECTRAL AND METABOLIC IMAGING

Rui C. Martins[1,2], Nuno C. Sousa[1,2], António C. Silva-Ferreira[3,4]
[1]Life and Health Sciences Research Institute, School of Health Sciences, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[2]ICVS/3B's - PT Government Associate Laboratory, Braga-Guimarães, Campus de Gualtar, 4710-057 Braga, Portugal
[3]Stellembosh University, Private Bag X1, Matieland, 7602, Stellembosh, South Africa
[4]Escola Superior de Biotecnologia, R. Dr. António Bernardino de Almeida, 4200-072, Porto, Portugal
rui.martins@ecsaude.uminho.pt

VinePAT is a process analytical technology system for the 'in-vivo' hyperspectral imaging and metabolic imaging of vineyards, for high-precision viticulture. It comprises a portable miniaturized uv-vis-swnir (200-1000nm) hand-held device, hyperspectral spectroscopy operating system, metabolic imaging client and server systems. Here in, we show a case study of how to use hyperspectral and metabolic imaging for harvest forecasting, taking into account the metabolic image predictions throughout maturation and knowledge based pattern recognition from a maturation database stated in 2007. The system is able to detect dynamical patterns of maturation and compare to the existing ones in the database, using techniques such as warping for temporal synchronization and tensor analysis for dynamic comparison of maturation patterns. VinePAT is designed so that the producer develops its own knowledge base and PAT models, ensuring that restricted, producer only desired metabolic fingerprint of grapes is attained. We further demonstrate how to use the system to schedule harvesting for different grades of desired quality.

O22

SERS IMAGING AND MCR-ALS IN THE STUDY OF THE CHEMICAL DISTRIBUTION OF CONTROLLED-RELEASE POLYMERIC FILMS CONTAINING PARACETAMOL

Mónica B.Mamián-López, Ronei J. Poppi
Institute of Chemistry, University of Campinas, P.O. Box 6154, 13084-971, Campinas, SP, Brazil
monlopez@iqm.unicamp.br

Four controlled-release polymeric films containing paracetamol (between 5.1 and 19.4%), were prepared on a commercial nanostructured gold surface using the solvent-evaporation method. Briefly, adequate amounts of each constituent (Paracetamol, HPMC, PEG and PVP) were dissolved in water, deposited over the metallic substrate and let to dry to 60 ºC. Then, the hyperspectral SERS images were acquired and processed with MCR-ALS (with a closure constraint) assisted by SIMPLISMA to estimate the pure responses from each compound. The chemical distribution was assessed by building the respective images with the scores values obtained from MCR-ALS. Using the same information, their mean concentration was also calculated, reaching relative errors as low as 3.6% for paracetamol. In addition it was possible to discern interactions among the constituents and the homogeneity of the polymeric system. This approach combines the spectral-spatial information with the high detectability offered by SERS spectroscopy giving a very simple, versatile procedure that lets us study qualitative and quantitatively systems like controlled-release films where is necessary to simultaneously evaluate both distribution and composition.

O23

A NOVEL APPROACH FOR ASSOCIATING ODOURS TO COMPOUNDS IN GC-MS/O DATA

Jan Gerretzen[a,b]; Lutgarde M.C. Buydens[a]; Ariette Tromp – van den Beukel[c]; Elisabeth Koussissi[c]; Eric Brouwer[c]; Jeroen J. Jansen[a]; Ewa Szymańska[a,b]

[a]Radboud University Nijmegen, Institute for Molecules and Materials, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
[b]TI-COAST, P.O. Box 18, 6160 MD Geleen, The Netherlands
[c]Heineken Supply Chain BV, P.O. Box 510, 2380 BB Zoeterwoude, The Netherlands
j.gerretzen@science.ru.nl

Gas Chromatography-Mass Spectrometry/Olfactometry (GC-MS/O) is indispensable to associate chemical information about volatile odorants in a sample to odour descriptions given by a panel of human assessors. However, interpretation of GC-MS/O data is considerably hampered in practice due to instrumental artefacts and variation inherent to using humans as detectors. More specifically, the following three challenges have to be dealt with when analysing GC-MS/O data:

1. elution time differences between GC-MS/O runs;
2. detection time differences between mass spectrometer and olfactometer, and
3. heterogeneity among odour descriptions for identical odorants.

We have developed a novel approach that overcomes all these limitations, leading to a robust and interpretable association of odours to compounds in GC-MS/O data.
In short, signal alignment via COW (Correlation Optimized Warping) is used to deal with challenge 1, while challenge 2 is overcome by automated detection of odour areas. The solution to challenge 3 involves the novel Total Odour Count (TOC) concept to provide insight into the odour description heterogeneity. The approach is demonstrated on data sets of an alcoholic beverage, one data set containing beverage with spiked odorants. All spiked odorants were detected by the approach and associated to proper odour descriptions, as validated with an odour database. Moreover, several other (non-spiked) odorants present in the beverage were detected.

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

O24

A BAYESIAN INFERENTIAL METHOD FOR PEAK DETECTION IN CHROMATOGRAPHIC SIGNALS

Martin Lopatka[a,c], Gabriel Vivo-Truyols[b], Marjan Sjerps[a,c]

[a] *Kortweg de Vries Institute for Mathematics, University of Amsterdam, Postbus 94248, 1090 GE, Amsterdam, The Netherlands*

[b] *Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Postbus 94248, 1090 GE, Amsterdam, The Netherlands*

[c] *Netherlands Forensic Institute, Postbus 24044, 2490 AA, Den Haag, The Netherlands*

We present a novel algorithm for probabilistic peak detection in first-order chromatographic data. Unlike conventional methods that deliver a binary answer pertaining to the expected presence or absence of a chromatographic peak, our method calculates the probability of a point being affected by such a peak. The algorithm makes use of chromatographic information (i.e. the expected width of a single peak and the standard deviation of baseline noise). As prior information of the existence of a peak in a chromatographic run, we make use of the statistical overlap theory. We formulate an exhaustive set of mutually exclusive hypotheses concerning presence or absence of different peak configurations. These models are evaluated by fitting a segment of chromatographic data by least-squares. The evaluation of these competing hypotheses can be performed as a Bayesian inferential task. We demonstrate this approach for peak detection using multiple chromatographic systems providing examples of both improved performance and increased flexibility.

O25

HSI-NIR AND CHEMOMETRICS IN FORENSIC SCIENCE: DETECTION OF EXPLOSIVE RESIDUES IN HUMAN HANDPRINTS

Mª Ángeles Fernández de la Ossa[1], Carmen García-Ruiz[1] and José Manuel Amigo[2]

[1] Inquifor Research Group, Department of Analytical Chemistry, Physical Chemistry and Chemical Engineering and University Institute of Research in Police Sciences (IUICP), University of Alcalá, Ctra. Madrid-Barcelona km 33.600, 28871 Alcalá de Henares (Madrid), Spain.

[2] Department of Food Sciences, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.

jmar@life.ku.dk

Forensic scientists are continuously looking for novel tools that offer reliable evidence to the judicial system. A current challenge is the detection of explosive residues directly in people suspected of preparation of improvised explosive devices (IEDs). Here, we put forward a methodology capable of detecting trace concentrations of six explosives commonly used as part of IEDs (ammonium nitrate, black powder, single- and double-base smokeless gunpowders and dynamite) in human handprints by using near infrared hyperspectral imaging (HSI-NIR) in combination with linear classification methods (e.g., partial least squares discriminant analysis - PLS-DA). The results are extremely promising, obtaining levels of sensitivity and specificity for all classes close to 100% in the cross-validated models. This allowed us the detection of microparticles of explosives in diverse human handprints. Moreover, a comprehensive study of the possible interferences (sweat and common dirtiness) has been made, demonstrating the possible viability of the proposed methodology in more real situations.

O26

MONITORING OF POLYMER CURING PROCESSES WITH RAMAN SPECTROSCOPY

Thomas Jørgensen, Sergey Kucheryavskiy
Department of Biotechnology, Chemistry and Environmental Engineering,
Aalborg University, Niels Bohrs vej 8, 6700, Esbjerg, Denmark
svk@bio.aau.dk

Today, many products, from pipes to wind turbines, are made from different kinds of resins or related polymers. The most important stage in the production is curing, when liquid resin reacts with additives and forms a solid matrix resulting in a light material with excellent mechanical properties. In the curing process the main focus is on the point when the liquid resin is hard enough and thereby cured enough to be used for further processing. By determining the degree of cure, the time, which the material has to after cure at the production facility, can be more precisely estimated. Besides that, measuring the degree of cure on different parts of the material allows to assess homogeneity of cure.

The current techniques for measuring the degree of cure involve different analytical and mechanical tests. However all state-of-art methods are time consuming and destructive, which makes use of them very limited. Therefore inventing new non-destructive methods for monitoring and control of polymer curing processes is essential for cost, time and quality effective production.

In the present work a feasibility study of using Raman spectroscopy coupled with chemometric methods for monitoring of degree of cure of commercial resin system for composite application has been investigated. For each case a theoretical molecular model of curing process was built and compared with experimental results to explain behavior of the Raman spectra peaks. The obtained results show that spectra show clear difference even for the small changes happened after several hours of curing.

IN SITU ATR-FTIR MONITORING OF 1,2-BUTYLENE OXIDE POLYMERIZATION AND CHEMOMETRIC DATA ANALYSIS

Xiaoyun Chen, Randy Pell, Sagar Sarsani, Brian Cramm, Carlos Villa, and Ravindra Dixit
The Dow Chemical Company, Analytical Sciences, 1897 Bldg., Midland, MI 48667, USA
xchen4@dow.com

There has been rapid growth in the application of in situ optical spectroscopy techniques for reaction and process monitoring recently in both academia and industry. Vibrational spectroscopies such as mid-infrared, near-infrared spectroscopy, and Raman spectroscopy have proven to be versatile and informative. Accurate determination of concentrations, based on highly overlapped spectra, remains a challenge. As an example, 1,2-butylene oxide (BO) polymerization, an important industrial reaction, initiated by propylene glycol (PG) and catalyzed by KOH, is studied in this work in a semi-batch fashion by using in situ attenuated total reflectance Fourier transform infrared spectroscopy (ATR FT-IR) monitoring. The weak BO absorbance, the constantly changing interference from the product oligomers throughout the course of the reaction, and the change in BO spectral features with system polarity posed challenges for quantitative spectral analysis based on conventional methods. An iterative concentration-guided classical least-squares (ICG-CLS) method was developed to overcome these challenges. Taking advantage of the concentration-domain information, ICG-CLS enabled the estimation of the pure oligomer product spectra at different stages of the semi-batch process, which in turn was used to construct valid CLS models. The ICG-CLS algorithm provides an in situ calibration method that can be broadly applied to reactions of known order. Caveats in its applications are also discussed.

COMPARISON OF BOOTSTRAP AND ASYMPTOTIC CONFIDENCE LIMITS FOR CONTROL CHARTS IN BATCH MSPC STRATEGIES

Hamid Babamoradi, Frans van den Berg, Åsmund Rinnan
University of Copenhagen, Faculty of Science, Department of Food Science,
Spectroscopy & Chemometrics section
Rolighedsvej 30, DK-1958 Frederiksberg, Denmark
hamba@food.ku.dk

Multivariate Statistical Process Control (MSPC) is widely used as a Process Analytical Technology (PAT) tool in many industries ranging from food and pharmaceuticals to chemicals. Estimation of confidence, warning and control limits plays a key role in process monitoring and fault diagnosis. Traditional confidence limits are estimated by assuming predefined distributions for the data and residuals during the modelling stage. In most cases, neither data nor residuals follow these distributions since measurements come from designed and controlled spaces. Thus, uncertainty associated with the model and predictions cannot be well explained by such an inflexible prior assumption, potentially inflating type I and II errors. Moreover, traditional confidence limits are not available for contribution plots, which are used to identify the cause after a special event has been detected. This greatly reduces the applicability of contributon plots. Bootstrap re-sampling, on the other hand, can be compatible with the complexities of process data, since it allows us to estimate uncertainty distributions directly from data under NOC [1]. We are suggesting new bootstrap confidence limits as an alternative to asymptotic methods, built on predictions for the control charts in different online PCA-based batch MSPC strategies. Performance of the bootstrap and asymptotic confidence limits for commonly applied methods will be compared based on Overall Type I and II errors. Bootstrap confidence limits will also be formulated for contribution plots.

[1] H. Babamoradi, F. van den Berg, A. Rinnan, Comparison of bootstrap and asymptotic confidence limits for control charts in batch MSPC strategies, Chemometrics Intellig.Lab.Syst. 127 (2013) 102-111.

O29

A NOVEL FOURTH-ORDER CALIBRATION METHOD BASED ON ALTERNATING QUINQUELINEAR DECOMPOSITION ALGORITHM FOR PROCESSING HPLC-DAD-KINETIC-pH DATA OF NAPTALAM HYDROLYSIS

Xiang-Dong Qing, Hai-Long Wu, Xi-Hua Zhang, Hui-Wen Gu, Yong Li, Ru-Qin Yu

State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China;

hlwu@hnu.edu.cn

Five-way HPLC-DAD-kinetic-pH data was obtained by recording the kinetic evolution of HPLC-DAD signals of samples at different pH values and a new fourth-order calibration algorithm, alternating quinquelinear decomposition (AULD) based on pseudo-fully stretched matrix forms of the quinquelinear model, was developed. It has been successfully applied to investigate quantitatively the kinetics of naptalam (NAP) hydrolysis in three separate systems containing (1) indole-3-acetic acid (IAA) and 1-naphthaleneacetamide (NAD); (2) IAA, NAD and unexpected component(s) and (3) extracts from soil samples, respectively, as interferents. The good recoveries of NAP were obtained in these samples by selecting the time region of chromatogram. The resolved elution time, spectral, kinetic time and pH profiles of NAP were in good agreement with experimental observations. It demonstrates the potential for the utilization of fourth-order calibration for some complex systems.

O30

MCR CALIBRATION VS. PLS CALIBRATION. ANALYSIS OF SPECTROPHOTOMETRY DATA

O.Ye. Rodionova[1], Y.V. Zontov[2], A.L. Pomerantsev[1,3]
[1]*Semenov Institute of Chemical Physics RAS, Kosygin 4, Moscow 119991, Russia*
[2]*National Research University Higher School of Economics, Myasnitskaya 20, Moscow, 101000, Russia*
[3]*Institute of Natural and Technical Systems RAS, Kurortny 99/18, Sochi, 354024, Russia*
rcs@chph.ras.ru

Quality of calibration models constructed using partial least squares (PLS) and multivariate curve resolution (MCR) approach has been compared previously for various data sets. The main advantageous of each of the method are well known. They are the more accurate prediction for the PLS regression and possibility of reconstruction of the spectra of pure components for the MCR calibration. The subtle differences are not so evident and ordinary depend on specific datasets.

The current study is devoted to the quantitative determination of the rear earth (RE) elements in aqua solution of nitric acid. Ultra-violet and visible (UV-Vis) spectroscopy has been used as a measurement technique. Several cases are considered:

1. Spectra of RE elements have several, reliable absorbance peaks and noise level is low.
2. Noisy spectra with highly overlapped and poorly measured due to detector saturation peaks.
3. Mixture of the first two extreme cases.

Analyses of various cases helps to reveal additional pros and cons of the PLS and MCR calibrations.

RE nitric aqua solutions has been investigated as a feasibility stage of the process analytical technology where the variability in the process is unavoidable. For this reason additional property of the PLS and MCR models, viz. model robustness has been investigated. Variations in nitric acid concentration and higher concentrations of RE elements, not considered in the modeling stage, have been used for these purposes.

O31

DETERMINATION OF MILK ADULTERATION BY SUCROSE USING FT-MIR SPECTROSCOPY AND CHEMOMETRICS METHODS

Bassbasi Elmahfoud , Souhassou Said and Oussama Abdelkhalek

Laboratory of Applied Spectro-chemometry and Environmental, Faculty of Sciences and Techniques of Beni Mellal, University Moulay Soulymane, Morocco.
BP 523 23000 beni Mellal
oussamaabdelkhalek@yahoo.fr

Vibrational spectroscopy has proven itself to be a valuable contributor in the study of various fields of science, primarily due to the extraordinary versatility of sampling methods. FT-MIR measurement gives the vibrational spectrum of the analyte, which can be treated as its “fingerprint,” allowing easy interpretation and identification. Over the last few years, there has been tremendous technical improvement in FT-MIR spectroscopy. Advances in the instrumental design of FT-MIR spectrometers coupled with chemometrics methods have also been described and this enable trace level detection and satisfactory analysis.

Fourier transform infra-red spectroscopy (FT-MIR) coupled with chemometrics methods have been applied for a fast and non-destructive quantitative determination of sucrose in raw milk. The partial least-squares regression (PLS) method was successfully applied to predict added sucrose in raw milk based on modeling FTIR spectral transmission measurements. From the obtained results (very low relative prediction error and limit of detection (LOD) around 5 % and 1.25 g/L respectively, it can be concluded that FT-MIR with chemometrics can be useful for accurate quantitative determinations of added sucrose in raw milk. The proposed procedure is fast, non-destructive, simple and easy to use.

O32

DPLS AND INTERACTING VARIABLES: A NEW STRATEGY TO REVEAL THEIR CONTRIBUTION TO THE MODEL

Jasper Engel, Geert J. Postma, Ingrid van Puifflik, Lionel Blanchet, and Lutgarde M.C. Buydens
Radboud University Nijmegen, Institute for Molecules and Materials (IMM), Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
g.j.postma@science.ru.nl

Distance Partial Least Squares (DPLS) is a simple regression and classification method for solving a wide range of problems [1]. It is capable of solving both linear and highly non-linear problems. Interestingly it is also able to solve problems involving interacting variables. The latter two problems cannot be solved using PLS alone. DPLS is based on a distance matrix, calculated from the original data, followed by Partial Least Squares (PLS) regression. However, due to the calculation of a distance matrix the relation to the original variables is lost, meaning that the interpretation of the model is still impossible.

We have demonstrated in the past that the so-called pseudo sample trajectories can be applied to open such 'black-box' models and facilitate interpretation [2, 3]. However, in case of interacting variables the basic pseudo sample trajectory method is still not effective. An improvement of the methodology is proposed and tested. Interaction between variables was successfully detected and visualized. This facilitates the interpretation of variables which up to now hardly seem to be incorporated in the search after variables possibly relevant for e.g. a chemical or biochemical model (i.e. model interpretation).

References

[1] Zerzucha, P., M. Daszykowski, and B. Walczak, Chemometrics and Intelligent Laboratory Systems, 2012, 110(1), 156 162.
[2] Postma, G.J., P.W.T. Krooshof, and L.M.C. Buydens, Analytica Chimica Acta, 2011, 705(1-2), 123-134.
[3] Smolinska, A., L. Blanchet, L. Coulier, K.A.M. Ampt, T. Luider, R.Q. Hintzen, S. Wijmenga, L.M.C. Buydens, PLoS ONE, 2012, **7**(6), e38163.

O33

DATASET INTEGRATION: MODELING STRUCTURE BETWEEN HIGH DIMENSIONAL DATASETS

I. Montoliu

Analytical Sciences Department, Nestlé Research Center,Route du Jorat 57, 1000 Lausanne 26, Switzerland.
ivan.montoliuroura@rdls.nestle.com

The advent of the '-omics' during the last years has been a challenge for a big part of the analytical chemistry community. Metabolomics, in particular, has offered an interesting way to describe and provide new insights into the modulation of regulatory processes in complex biological systems. Holistic approaches have been widely used with this purpose, attempting to maximize the number of descriptors. This fact, together with the use of high throughput analytical techniques, has placed Chemometrics as a key element in the Metabolomics pipeline.

In the last years we have seen how Chemometrics tackled these challenges with success, performing feature identification through the association of metabolic profiles (spectroscopic, quantitative) with phenotypes. Moreover, the increase in throughput of many analytical methods has allowed the availability of a wealth of data from same samples. As a consequence, key question is now moving towards connecting the outcomes of these results. Integration onto a 'systems biology' approach poses new serious challenges, going from the integration of datasets to the identification of relevant elements. The problem can be tackled using several multivariate analysis approaches such multi block modeling, PLSR, Canonical Correlation Analysis or self-organizing maps. In this presentation we show how such a kind of modeling can be performed using a two-step modeling based in Random Forests for feature extraction and Elastic Nets for profile integration.

O34

LOCAL MODELING REVISITED: NOVEL APPROACHES FOR NON-LINEAR REGRESSION AND CLASSIFICATION

Marta Bevilacqua[1], Rasmus Bro[2], Federico Marini[1]
[1]Dept. of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy
[2]Faculty of Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
federico.marini@uniroma1.it

The evolution of the analytical instrumentation is such that an increasing number of high throughput fingerprinting platforms is available, resulting in the possibility of dealing with more and more complex real world problems. From the data analytical standpoint, this consideration translates to the possibility that many sources of variation, other than the one(s) of interest, affect the instrumental signals, resulting in a non-linear relationship between the dependent and independent blocks or, in the case of pattern recognition, in a non-linear separability of the categories in the feature space.

In such cases, the use of non-linear modeling approaches can help tackling with the complex relationships between the data blocks to be analyzed. However, most of the commonly used supervised non-linear methods have strict requirements in terms of the samples to variables ratio, and are more prone to overfitting. One way of effectively coping with these limitations is to implement the non-linearity through the training of locally linear models [1-2]. In this communication, the possibility of extending the Locally Weighted-PLS approach originally proposed by Centner and Massart [2] in different ways, e.g. to deal with non-linear classification problem (Locally Weighted PLS-DA) or to properly include replicated measurements in the model building phase by means of a modified bootstrap procedure, will be presented. Moreover, other key points such as the evaluation of the influence of the weighting and/or scheme adopted and the possibility of interpreting the model through the inspection of the local loadings will also be addressed.

References

[1] W.S. Cleveland, S.J. Devlin, J. Am. Stat. Assoc. **1988,** 83, 596-610.
[2] T. Naes, T. Isaksson, B.R. Kowalski, Anal. Chem. **1990**, 62, 660-673.
[3] V. Centner, D.L. Massart, Anal. Chem. **1998,** 70, 4206-4211.

O35

ARE INDEPENDENT COMPONENT ANALYSIS (ICA) AND MINIMUM VOLUME SIMPLEX ANALYSIS (MVSA) APPROPRIATE FOR MULTIVARIATE CURVE RESOLUTION IN ANALYTICAL CHEMISTRY?

Roma Tauler[a], Xin Zhang[a] and Hadi Parastar[b]
[a]*Institute of Environmental Assessment and Water Diagnostics (IDEA-CSIC), Barcelona, Spain*
[b]*Shiraf University, Teheran, Iran*

In this presentation we will show and discuss the possibilities of alternative methods proposed in different fields for similar purposes than multivariate curve resolution in chemometrics and analytical chemistry. The concepts of independence and of minimum volume simplex will be interpreted in terms of their meaning for the purpose of resolving component response profiles with chemical meaning, usually expressed as concentration and spectra profiles. Advantages and limits of these approaches will be compared with those associated with Multivariate Curve Resolution methods. Typical data sets in Chemistry, with different levels of complexity, selectivity and correlated information will be tested and conclusions from these comparisons will be given[1,2].

References

[1] H. Parastar, M.Jalali-Heravi, R.Tauler, Trends in Analytical Chemistry, 31 (2012) 134-143
[2] X.Zhang, R. Tauler, Analytica Chimica Acta 762 (2013) 25– 38

O36

ROBUST AUGMENTED CLASSICAL LEAST-SQUARES

Mohammad Goodarzi and Wouter Saeys
Department of Biosystems, Faculty of Bioscience Engineering, KU Leuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium
mohammad.godarzi@gmail.com

A series of multivariate calibration methods called Augmented Classical Least Squares (ACLS) has been proposed leading to an improvement in multivariate spectral analysis where there are strong interferences. These techniques couple both linear additive (e.g., CLS) and linear unmixing models (e.g., PCA) in order to overcome the normal limitation of CLS by incorporating different sources of prior information, such as measured pure component spectra, in the model. Since PCA is sensitive to outliers, the added value of using robust PCA in the augmentation has been evaluated in this study. Furthermore, both ACLS and robust ACLS (RACLS) have been compared to well-known biased regression techniques such as Partial Least Squares (PLS) and Ridge Regression. The prediction performance of RACLS and ACLS was found to be comparable. Note that, in our previous studies, it was found that incorporation of pure component spectra of the interferents in the ACLS framework led to a reduction of their effect, especially when a different interferent structure was present in the test samples. However, in this study, we also tried to couple a multivariate curve resolution method with CLS instead of a PCA with CLS. This is done by either SIMPLISMA or incorporating pure component spectra of the inteferents as an initial guess.

O37

STRATEGIES FOR TWO DIFFERENT CALIBRATIONAL SPACES

Yi-Zeng Liang[1], Yong-Huan Yun[1], and Qing-Song Xu[2]

[1]*College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P. R. China*

[2]*School of Mathematics and Statistics, Central South University, Changsha 410083, P. R. China*

yizeng_liang@263.net, qsxu@csu.edu.cn

Two chemical modeling spaces, say component spectral space and measured variable space in analytical chemistry, are firstly defined, respectively. From this point of view, classical calibration and inverse calibration can be two kinds of multivariate calibration in chemical modeling. It is worth noting that the intrinsic difference between these two calibration models is not fully investigated. The net analyte signal (NAS) proposed by Lorber based on orthogonal projections can be regarded as the theoretic summary for classic calibration. Also, the tensor calibration for high dimensionally linear data is its natural extension. However, in the case of complex analytical systems, NAS cannot be well defined in inverse calibration due to the existence of uninformative and/or interfering variables. Therefore, application of the NAS cannot improve the predictive performance for this kind of calibration, since it is essentially a technique based on the full-spectrum. From our perspective, variable selection can significantly improve the predictive performance through removing uninformative and/or interfering variables. In this study, we first clarify the intrinsic difference between these two calibration models and then use a new perspective to intrinsically prove the importance of variable selection in the inverse calibration model for complex analytical systems. In addition, we have experimentally validated our viewpoint through the use of one UV dataset and two generated near infrared (NIR) datasets.

O38

TOPOLOGICAL DATA ANALYSIS: DATA HAS SHAPE AND SHAPE HAS MEANING

Ludovic Duponchel
LASIR, Université Lille – Nord de France, 59655 Villeneuve d'Ascq Cedex, FRANCE.
ludovic.duponchel@univ-lille1.fr

An important feature of analytical chemistry is that data of various kinds is being produced at an unprecedented rate. This is mainly due to the development of new instrumental concepts and experimental methodologies. It is also clear that the nature of the data we are acquiring is significantly different. Indeed in chemometrics, we are given data in the form of always longer vectors, where all but a few of the coordinates turn out to be irrelevant to the questions of interest, and further that we don't necessary know which coordinates are the interesting ones. "Big data" in chemometrics is a future that might be closer than any of us suppose. It is in this sense that new tools have to be developed in order to explore and valorize such datasets. Topological data analysis (TDA) is probably one of these.[1] It was developed only few years ago considering that a data set is a sample or "point cloud" taken from a manifold in some high-dimensional topological space. The sample data are used to construct simplices, generalizations of intervals, which are, in turn, "glued" together to form a kind of wireframe approximation of the manifold. This manifold represents the "shape" of the data from which you can discover sub-populations of samples that cannot be observed with conventional techniques like PCA or cluster analysis. For this presentation, an analytical chemistry dataset will be explored in order to reveal the potential of this new concept.

[1] Extracting insights from the shape of complex data using topology, P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson & G. Carlsson, *Scientific Reports*, 3, 1236, doi:10.1038/srep01236

O39

EVALUATION OF CONTAMINATION PROFILE OF PERSINTENT ORGANIC POLLUTANTS IN FAT FROM TROPICAL DETRITIVOROUS FISH FROM BRAZIL BY KOHONEN NEURAL NETWORK

Gilmare A. da Silva[1], João P. M. Torres[2], Claudio E. de Azevedo e Silva[2, 3], Rodrigo O. Meire[2], Mauro de F. Rebelo[2], Olaf Malm[2], José Lailson-Brito Jr[3], Wanderley R. Bastos[4], Wilson de F. Jardim[5], Ricardo B. Rios[6], Juan C. Colombo[7], Gualberto Gonzalez-Sapienza[8], Luz Claudio[9], Bernhard Henkelmann[10], Kalr-Werner Schramm[10,11]

[1]Departamento de Química - Universidade Federal de Ouro Preto (UFOP), Campus Morro do Cruzeiro, S/N, 35400-000, Ouro Preto/MG, Brasil.

[2]Instituto de Biofísica Carlos Chagas Filho - Universidade Federal do Rio de Janeiro (UFRJ), 21941-901, Rio de Janeiro, Brasil.

[3]Faculdade de Oceanografia - Universidade Estadual do Rio de Janeiro (UERJ), 20550-103, Rio de Janeiro/RJ, Brasil.

[4]Laboratório de Biogeoquímica Wolfgang Christian Pfeiffer - Universidade Federal de Rondônia (UNIR), 76801-974, Porto Velho/RO, Brasil.

[5]Instituto de Química - Universidade Estadual de Campinas (UNICAMP), 13083-970, Campinas/SP, Brasil.

[6]Centro EULA, Universidad de Concepcion, Concepcion, Chile.

[7]LAQAB, Universidad de La Plata, La Plata, Argentina.

[8]Faculdad de Quimica, Universidad Nacional do Uruguay, Montevideo, Uruguay.

[9]Community and Environmental Medicine Department, Mount Sinai School of Medicine, New York, USA.

[10]Helmholtz Center Munich-German Research Center for Environmental Health (GmbH), Institute of Ecological Chemistry, Ingolstädter Landstr.1, D-85764 Neuherberg, Germany.

[11]TUM, Wissenschaftszentrum Weihenstephan fuer Ernaehrung und Landnutzung, Department fuer Biowissenschaftliche Grundlagen, Weihenstephaner Steig 23, 85350 Freising, Germany.

gilmare@iceb.ufop.br

Environmental contamination by persistent organic pollutants (POP) - polychlorinated biphenyls (PCB) and organochlorine pesticides (OCP) - was investigated in different river basins of Brazil, using fat extracts of detritivorous fish of the Genus Prochilodus. The fishes were dissected in the field and brought frozen to the laboratory, where non-polar extracts of fish dorsal with muscle were obtained. After a two-step clean-up, the extracts were analyzed by high resolution gas chromatography / high resolution mass spectrometry and the data treated by Kohonen neural network, with the aim to get the relations among samples, POP, and the influence of POP in samples. The contamination pattern in fish tissue varied from region to region and, in general, it indicated to be related with the human activity. Different types of OCP were presented in the majority of the samples. Hexachlorocyclohexane (HCH) isomers had a more marked presence in the samples collected in Manaus and dichloro-diphenyl-trichloroethane (DDT) metabolites in Corumbá. The PCB indicators were found in all samples from all stations with a more marked contamination obtained from São Paulo state. In different regions of Brazil the pattern of contamination was not similar, but seems clear, at least for PCB, that the more contaminated fish was observed at the most developed part of the country. The ubiquitous presence of POP legacy in South American detritivorous fishes was easily exhibit by Kohonen network; this study provided a general overview of POP dispersion in Brazil and highlighted the potential of “fish approach” to survey this contamination.

Acknowledgements:
Supporting Agencies: FAPEMIG; MCT/CNPq & MEC/CAPES (Brazil); FONDECYT (Chile); Fogarty International Center/NIH (USA).

O40

PROJECTION PURSUIT REVISITED: EXPLORATORY DATA ANALYSIS OF MULTICLASS DATA

Peter D. Wentzell[1], Siyuan Hou[1], Carolina S. Silva[2], M. Fernanda Pimentel[3]
[1]*Department of Chemistry, Dalhousie University, PO Box 15000, Halifax, NS B3H 4R2, Canada*
[2]*Departamento de Química Fundamental, Universidade Federal de Pernambuco, Recife, PE, Brazil*
[3]*Departamento de Engenharia Química, Universidade Federal de Pernambuco, Recife, PE, Brazil*
peter.wentzell@dal.ca

Projection pursuit (PP) is a technique for the exploratory analysis of multivariate data in which "interesting" subspace projections are obtained on the basis of optimizing a "projection index." Unlike principal components analysis (PCA), which is a variance-based projection method, PP is not constrained in the criteria used to seek informative projections and a variety of approaches have been employed. One of the most useful strategies for unsupervised class separation has been the use of kurtosis, the fourth statistical moment, as a projection index. Minimization of kurtosis as a means of detecting non-normality can lead to (unsupervised) class separation not achievable by PCA. However, the routine implementation of PP has been hindered by the slow nonlinear optimization methods. Recently, however, more efficient algorithms based on quasi-power methods have been developed. In this presentation, a range of examples will be presented to illustrate the power of PP as an exploratory data analysis tool. Issues associated with the implementation of PP, algorithmic alternatives, sample-to-variable ratio, unbalanced classes, and convergence will be addressed, and recent modifications to the original algorithm will be described.

P1

CANCER DETECTION USING MICROARRAY GENE EXPRESSION DATA SET: COMBINING DATA DIMENSION REDCUTION AND VARIABLE SELECTION TECHNIQUE

Sadegh Karimi
Department of Chemistry, College of Sciences, Persian Gulf University, Bushehr –Iran
karimi.sadegh@gmail.com

Cancer classification using gene expression data is known to contain the keys for addressing the fundamental problems relating to cancer diagnosis. LDA is one of the most used traditional classification techniques [1]. Since a pooled covariance matrix is calculated, the number of objects must be greater than the number of variables. Variable selection methods, e.g. Genetic Algorithms (GAs), cannot appropriately handle a large number of variables without either severe overfitting. Therefore, a data dimension reduction can be useful when dealing with high-dimensional data. To do so, an algorithm for the extraction of significant features from high-dimensional data was proposed. First, the data are reduced using PCA on cluster of the original gene expression profile [2]. The dataset is then represented by the significant scores retained from these local clusters. Afterwards; the reduced dataset can be analyzed through the specified PC selection (GA) techniques coupled with classification methods (LDA). Two gene expression datasets (Leukemia and SRBCT) are used. The performance of the presented classification models was based on Non-Error Rate (NER), evaluated both on cross-validation and external test samples. Seven SOM networks from 2–8 were checked for both data sets. The statistical parameter reveal that GA-LDA model obtained from five Kohonen nodes (25 clusters) is the optimum one for both goodness of fit and prediction ability. GA-LDA model of this cluster size (which used 5 PCs out of 85 extracted PCs) possesses very high degree of correctly assigned sample (NER) 1.00, 1.00 and 0.94 for calibration, cross-validation and prediction, respectively.

References

[1] WJ Krzanowski . Principles of Multivariate Analysis. Oxford University Press: Oxford, 1988.
[2] B Hemmateenejad,. S.Karimi., *Journal of Chemometrics*. 25 (2011) 139.

P2

X-METABOLOMICS: A COMPREHENSIVE SOFTWARE PLATFORM FOR DATAMING OF MASS SPECTROSCOPY AND BATCH PROCESS ANALYTICAL TECHNOLOGY

Rui C. Martins[1,2], António C. Silva-Ferreira[3,4], and Nuno C. Sousa[1,2]

[1]Life and Health Sciences Research Institute, School of Health Sciences,University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

[2]ICVS/3B's - PT Goverment Associate Laboratory, Braga-Guimarães, Campus de Gualtar, 4710-057 Braga, Portugal

[3]Stellembosh University, Private Bag X1, Matieland, 7602, Stellembosh, South Africa

[4]Escola Superior de Biotecnologia, R. Dr. António Bernardino de Almeida,4200-072, Porto, Portugal

rui.martins@ecsaude.uminho.pt

Metabolomics datamining and pattern recognition is an ever expanding area of applications in biotechnology and health sciences. Here in, we present a mass spectroscopy metabolomics pipeline implemented in our software platform X-metabolomics that is able to perform at-line processing of MS chromatograms. Chromatogram batches are imported from MS platforms in NetCDF format, and pre-processed: baseline correction, peak extraction, warping, quantification and identification of extracted peaks into an internal database, and performing filtering techniques before any interpretation of metabolic results, such as trying to remove from chromatograms non-biological compounds and artifacts, saturated peaks from any further analysis. Data can thereafter be used in supervised or un-supervised classification for correlation between compounds, or performing time-course analysis for identifying metabolic modes and possible biomarkers. Process analytical technology, such as, multivariate control charts, contribution plots and diagnostic plots are also implemented for rapid diagnostic of chromatograms and understanding metabolic deviation from control samples.

P3

DETECTION OF PESTICIDE RESIDUES IN TOMATO PEEL BY SERS IMAGING AND CHEMOMETRIC METHODS

Carlos Diego L. Albuquerque and Ronei J. Poppi
Institute of Chemistry, University of Campinas (Unicamp), P.O. Box 6154, 13084-971 Campinas, SP, BR
carlostaek2@yahoo.com.br; carlos.albuquerque@iqm.unicamp.br

Surface enhanced Raman spectroscopy (SERS) has been successfully applied in analytical chemistry in recent years, mainly to achieve low limits of detection. Moreover, its application in real samples is limited by sample complexity, presenting overlapping of peaks between the analyte and interfering. Non-Negative Matrix Factorization (NMF) and Multivariate Curve Resolution MCR are curve resolution methods designed to solve this problem. In this work, the SERS image technique in conjunction with curve resolution methods was applied for detection of malathion in tomato peel in concentrations in the range of 12.3 to 0.123 mg L^{-1}. The methodology also makes possible to recover the pure spectra, in order to certify the pesticide residues present in the sample. Both curve resolution methods were employed with Alternating Least Square (ALS) algorithm, non-negativity and closure constraints and the Asymmetric Least Square (AsLS) algorithm and normalization by row mean were used for pre-processing of data. NMF results were better than MCR, with superior malathion spectra recuperation. The distribution maps of concentration for malathion and tomato are complementary and heterogeneous. Furthermore, the total concentration of malathion was increasing in the mapping area at high levels of pesticide, in form of a non-linear curve (scores versus concentration expected). The proposed methodology is fast, efficient and no requires sample manipulation.

P4

DETECTION OF PESTICIDE IN PEEL FRUIT IN THE PRESENCE OF HETEROSCEDASTIC AND CORRELATED NOISE BY SERS AND MCR-WALS

Carlos Diego L. Albuquerque and Ronei J. Poppi
Institute of Chemistry, University of Campinas (Unicamp), P.O. Box 6154, 13084-971 Campinas, SP, BR
carlostaek2@yahoo.com.br; carlos.albuquerque@iqm.unicamp.br

The multichannel spectroscopic measurements, such as Surface-Enhanced Raman Spectroscopy (SERS) imaging is susceptible to heteroscedastic and/or correlated noise, due to the variations in the colloid homogeneity, signal intensity variation and sample complexity. Multivariate Curve Resolution (MCR) with Weighted Least Squares (WALS) has been used to solve this problem. In this work, the SERS image technique in conjunction with MCR-WALS was applied for detection of malathion in peel Damson plum in concentrations range of 12.3 to 0.123 mg L^{-1}.The Asymmetric Least Squares (AsLS) algorithm was used for data pre-processing. In order to obtain a better understanding on the noise structure, it was analyzed the measurement error covariance matrix through the generation of error covariance and correlation matrices from real experimental replicates. According to these two matrices, the SERS imaging measurement are strongly influenced by offset and multiplicative offset noise, presenting heteroscedastic and correlated noise. MCR-WALS spectral decomposition confirmed the existence of fluorescence signal (adding heteroscedastic noise) and overlapping of malathion and Damson plum spectra in several regions. The distribution of concentration map presented high complementarity and heterogeneity. The methodology proposed permitted the detection of pesticide in peel fruit, even in the presence of heteroscedastic and correlated noise in the SERS signal, with advantages to be fast, efficient and no requirements of sample manipulation.

P5

A PCA MODEL OF ATYPICAL ANDERSEN CASCADE DATA

Lauren Seabrooks, Jennifer Wylie and Justin Pennington
Respiratory Product Development, Merck, 181 Passaic Ave, Summit, NJ 07901, USA
lauren.seabrooks@merck.com

This research details a novel application of principal component analysis (PCA) to aid in the root-cause determination of atypical Andersen cascade impaction (ACI) data. Andersen cascade impaction is a technique used to measure the aerodynamic particle size distribution of medical inhalers an important indication of drug deposition in the lung. It is one of several methods which provide fractionation of particles based on the differences in inertia. A particle passes through numerous stages each with decreasing nozzle apertures, increasing linear velocity, and either remains entrained in the laminar flow paths or breaks through the stream and impacts on a collection surface located beneath each stage[1]. Although information rich this technique is labor intensive, involving many steps, and ultimately susceptible to errors. Therefore a prominent goal in the scientific community is to identify whether differences in data are genuine or the result of non-product quality factors causing variability in ACI measurements[2]. For this study, several factors related to analyst error and test design were identified. Next, a design of experiments approach was used to randomize experiments and determine which effects caused statistically significant ($p<0.05$) changes to the data. Significant effects were then modeled using PCA where each error defined a discrete class and each size fractionated group represented a distinct variable. Two components were sufficient to describe 80% the total variance and separate each class. The model was further successfully validated with three data sets. This research therefore establishes that PCA may aid in the investigation of atypical ACI data.

References

[1] Mitchell, J. and Nagel, M (2004). Particle size analysis of aerosols from medical inhalers. KONA. 22: 32-65.
[2] Glaab, V., Goodey, A., Lyapustina, S., and Mitchell, J. (2011). Efficient data analysis for MDIs and DPIs: failure mode effect analysis. Respiratory Drug Delivery Europe. 225-236.

P6

DETERMINATION OF DETERGENT AND DISPERSANT ADDITIVES IN GASOLINE USING *RING-OVEN* PRECONCENTRATION, NIR HYPERSPECTRAL IMAGING AND MULTIVARIATE ANALYSIS

Livia R. e Brito[1], Michelle P. F. da Silva[1], Jarbas J. R. Rohwedder[2], Celio Pasquini[2], Fernanda A. Honorato[3], and Maria Fernanda Pimentel[3]

[1]Departamento de Química Fundamental, Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil.
[2]Instituto de Química, Universidade Estadual de Campinas, Campinas, São Paulo, Brazil.
[3]Departamento de Engenharia Química, Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil.
liviarb.ufpe@gmail.com

In Brazil, the addition of detergent and dispersant additives in gasoline will be mandatory from 2015. Typically, the concentration of additives in gasoline is 200mg/kg, and the development of methods for their determination is urgent. In this work the ring-oven technique, proposed by Weisz in the 1950's, was revisited and employed to preconcentrate the additives. The gasoline containing the additives (in concentrations ranging from 50 to 1000 mg/kg) was dropped on a filter paper kept in a circular heated oven (150 ^{o}C). Thus, the volatile compounds are evaporated, while the remaining are spread by capillarity from the center of the paper creating a ring like region. The NIR hyperspectral image of the filter paper was obtained. The spectra of the pixels related to the presence of additive in the ring need to be separated from the most frequent pixels associated to the paper matrix. This was made using three different approaches based on: the concentration maps of the MCR-ALS, values of the PCA's scores, and the histograms of the PCA's scores. After selection, the informative spectra were averaged, and used to construct PLS regression models. The results show that it is possible to use the ring-oven technique and NIR hyperspectral imaging to quantify the additives in gasoline with values of RMSEP ranging from 30 to 100 mg/kg, depending on the type of additive.

P7

FAST DISCRIMINATION OF FRAGRANCES USING RAMAN SPECTROSCOPY AND CHEMOMETRIC METHODS

Robson B. Godinho[1,2], Mauricio C. Santos[2], and Ronei J. Poppi[1]
[1]*Institute of Chemistry, State University of Campinas, P.O.B 6154, Campinas, SP, 13083-970, Brazil.*
[2]*Givaudan do Brasil Ltda, Av. Eng. Billings, 1729 Ed. 31, São Paulo, SP, 05321-010, Brazil*
ronei@iqm.unicamp.br

Perfume has always aroused great fascination in humans for centuries and its use has become increasingly relevant in contemporary society. Due to this global demand, thousands of batches of these complex mixtures are produced and traded among companies which invest in the multibillion-dollar business. In this context, the IFRA (International Fragrance Association), to preserve the quality of fragrances produced and avoid cross-contamination, determines in its 'Code of Practice' that every fragrance produced must be properly sampled and tested to verify its compliance with sensorial and analytical specifications for a quality control department using predefined procedures prior to its commercialization. Given the large possibility of combination of organic molecules for creating of a perfume, the aim of this study was to develop an analytical methodology using Raman spectroscopy combined with chemometric methods for direct, fast and non-destructive analysis for discrimination of fragrances according to their composition. The supervised classification method used was the soft independent modeling of class analogies (SIMCA). In this procedure, 73 samples were identified according five odor classes previously defined and distributed in two data sets. The training set with 48 spectra obtained directly from samples stored in a vial was pre-processed and the data submitted for SIMCA modeling. The set used to validate the model showed a high success rate for the five classes of fragrance, with 100% of correct classification. This study demonstrates the wide applicability of the method for discrimination and classification of volatile organic mixture of high complexity for purpose of quality control.

P8

QUALITY CONTROL OF RAW COCOA BEANS BY NEAR INFRARED SPECTROSCOPY AND CHEMOMETRICS

Juliana C. Hashimoto[1], Jéssica C. Lima[1], Alessandra B. Nogueira[2], Juliana A. L. Pallone[1], Priscilla Efraim[1], and Ronei J. Poppi[3]

[1]*School of Food Engineering, University of Campinas - UNICAMP, 80 Monteiro Lobato Street, 13083-862, Campinas, São Paulo, Brazil*

[2] *Faculty of Chemistry, Pontifical Catholic University of Campinas – PUCC, Road D. Pedro I, km 136, 13083-862, Campinas, São Paulo, Brazil*

[3] *Institute of Chemistry, University of Campinas - UNICAMP, P.O Box 6154, 13083-970, Campinas, São Paulo, Brazil*
ronei@iqm.unicamp.br

Quality control in the cocoa industry is time consuming, expensive and generates chemical waste due to the use of conventional analytical methods. At this step, the parameters moisture, pH, acidity, lipid and shell content should be assessed as indicative of the quality, storage stability and process yield of raw cocoa beans. This study shows the application of Near-Infrared Spectroscopy (NIR) and chemometrics to overcome this challenge, providing a simple, rapid and reliable method for the quality control of cocoa beans. A total of 83 samples were analyzed by AOAC reference methods (Horwitz, 2006) and the shell content was measured by gravimetric assay. Ground and sieved samples were directly analyzed by diffuse reflectance in the near infrared region (10000-4000 cm^{-1}). NIR spectra were preprocessed by Savitzky-Golay first derivative and mean-centered. Partial Least Squares Regression (PLS) and interval-PLS calibration models were developed for the above mentioned parameters with 55 samples for calibration and 28 samples for external validation. All the correlation coefficients (r) were higher than 0.8 and the mean relative prediction errors ranged from 1.7% (lipid content) to 15% (acidity). The evaluation of the calibration, cross-validation and prediction errors lead to the conclusion that the NIR in conjunction to multivariate calibration models is able to predict with satisfactory accuracy these important quality parameters of cocoa beans. The findings of this work indicate NIR combined with chemometrics as a powerful tool for the quality control in the cocoa industry, making it feasible at different processing steps in a faster and cheaper way.

Reference

Horowitz, W (Ed.). **Official methods of analysis of the Association of Official Analytical Chemists**. 18 ed. Maryland : AOAC, 2006.

P9

CARBAMAZEPINE-NICOTINAMIDE COCRISTALS QUANTIFICATION AMONG ITS PRECURSORS USING FOUR SOLID STATE ANALYTICAL TECHNIQUES

Frederico L. F. Soares, Renato L. Carneiro
Department of Chemistry Federal University of São Carlos, BR-13560
São Carlos, 13565-905, SP, Brazil
renato.lajarim@ufscar.br

Cocrystals are mixed crystals designed by the addition of two or more different molecules in a same crystallographic pattern. These cocrystals presents unique physical and chemical proprieties, usually different from their precursors, being drug solubility the most important properties for the pharmaceutical industry. Although, carbamazepine is a widely used drug for treatment of epilepsy and trigeminal neuralgia, it presents a low solubility and bioavailability, leading to the necessity of a treatment using high doses. The cocrystallization of carbamazepine with highly soluble molecules, like saccharine and nicotinamide, have been widely studied as an alternative to enhance the solubility and bioavailability of this compound. Several analytical solid state techniques were used to characterize the carbamazepine-nicotinamide co-crystal, such as Fourier transformed infrared (ATR-FTIR), differential scanning calorimetry (DSC), X-ray powder diffraction (XRPD) and Raman spectroscopy. The suitability of these four solid state analytical techniques were evaluated to quantify a ternary mixture of carbamazepine-nicotinamide, and its coformers. Raman spectroscopy was the technique, which showed the best results for quantification, based on the errors of cross validation and prediction. The PLS regression model gave mean errors of cross validation around 2 (% wt/wt), and errors of prediction between 2.5 – 9.0 (% wt/wt) for all components presented in the ternary mixture. The quantification of cocrystal compounds among its coformers, it is necessary to determine the yield of the cocrystallization reactions and the purity of the product obtained in each stage.

P10

CHROMATOGRAPHIC SIMILARITY EXTRAPOLATION BY ADAPTED KALMAN FILTER RESPONSE

Martin Lopatka[1,3], Gabriel Vivo-Truyols[2], Marjan Sjerps[1,3]
[1]*Kortweg de Vries Institute for Mathematics, University of Amsterdam, Postbus 94248, 1090 GE, Amsterdam, The Netherlands*
[2]*Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Postbus 94248, 1090 GE, Amsterdam, The Netherlands*
[3]*Netherlands Forensic Institute, Postbus 24044, 2490 AA, Den Haag, The Netherlands*
martin.lopatka@gmail.com

In the context of forensic chemical profiling, the assertion of similarity between two signals is often the basis of evidence. We approach this task as an iterative two-stage state estimation problem, taking as input two continuous chromatographic signals. Adopting the underlying assumption that the chromatograms represent two noisy measurements of a single underlying process allows us to formulate their similarity as a function of the iteratively computed Kalman gain at each point in the time series. We adapt the Kalman filter approach to include dynamically-updated state covariance extrapolation using probabilistic peak detection which allows the necessary flexibility when chromatograms originate from significantly different samples.

Both global and local similarity scores can be calculated for each pairwise chromatogram comparison. In the case of chromatograms originating from different samples, the assumption that a single underlying process is responsible for both signals is violated. In order to arrive at a state estimate that minimizes estimated error covariance, the Kalman gain acts as a slack parameter. Given an accurate model of the system noise, this parameter can be compared to the degree of variation expected when multiple measurements are taken of a single process. In this way we derive a robust similarity metric for comparing full chromatographic signals useful for forensic profiling tasks.

P11

MULTI-PRODUCT PLS CALIBRATION MODELS OF ULTRAVIOLET SPECTRA: DETERMINATION OF TOTAL ACIDITY IN RED AND WHITE WINES

Luiza Mariano Leme, Cristiane da Silva Morais, Patrícia Valderrama, Paulo Henrique Março
Technological University Federal of Paraná (UTFPR), C.P. 271, CEP 87301-006, Campo Mourão-PR, Brazil
paulohmarco@gmail.com

A multi-product Partial Least Squares (PLS) model based on ultraviolet (UV) spectra (350-400nm) was proposed to determine total acidity in red and white wines. At this UV range it is observed the absorption of poliphenols as resveratrol, chalcones and tannins. Furthermore, in this spectral region the absorbance intensity was below one, so stray-light problems with the instrument were not observed and Beer's law was obeyd. The spectra were collect in a 1mm quartz cuvette and the reference method to acidity determination was acid-base titration. A total of 120 samples for calibration and 58 for validation were used. Outliers were evaluated based on leverage, unmodeled residuals in spectra and in dependent variable. The mean center PLS model with 20 latent variables show accuracy represented by Root Mean Squares Error of Calibration (RMSEC) and Prediction (RMSEP) 5.54 and 6.32, respectively. Fitting represented by correlation between reference and predicted values was R^2 of 0.74, which is very common when the reference value is provided from titration. Detection and quantification limits were 19.32 and 58.56, respectively. The inverse of analytical sensitivity allows for the establishment of a minimum concentration difference that could be detected to the model in the concentration range where it was investigated. In this case it is possible to distinguish wine samples with acidity differences from 5.85 $meqL^{-1}$. However, this is an optimistic estimate that considers the spectral noise as the larger source of error and does not take into account the lack of fit of the model.

P12

MCR-ALS AND NIRS APPLIED ON THE EVALUATION OF THE LEIDIUM MEYENII ANTIOXIDANT ACTIVITY

Daiane R. Soares, Rhayanna P. Gonçalves, Patrícia Valderrama, and Paulo Henrique Março
Federal Technological University of Paraná (UTFPR), C.P. 271, CEP 87301-006, Campo Mourão-PR, Brazil
paulohmarco@gmail.com

MCR-ALS was used as a tool to resolve NIR spectra acquired after heating oil with and without *Lepidium meyenii* flour to evaluate its antioxidant activity. *Lepidium meyenii*, known as Maca Peruana, is thought to be a very nutritive food besides lots of medicinal properties. Because of the lack of information, it was proposed to study its antioxidant activity comparing the products that appears after heating soya oil containing Maca Peruana flour and soya oil without this additive. The oils were heated first at 25°C and then at 30°C and so it was kept being heated until 170°C, which is considered to be the frying temperature. At each increase of 10 degrees in the temperature one sample of 5 mL was collected and stored until room temperature of 25°C and so, a VisNIR spectrum (range from 400 to 2500nm) was measured. From this, the best strategy found to evaluate the results was to use PCA at first, in order to determine the matrix rank. In the case of oil containing Maca Peruana, PCA results indicated that 3 principal components (PCs) explained most of the variance (98%) while to oil without Maca, 2 PCs explained the same variance (98%). One spectra resolved by MCR-ALS to the oil containing Maca is different from the ones resolved for “pure” oil, and the relative concentration shows that the 2 similar spectra behaves significantly different: when Maca is added, the oil resists some more to degradation compared with the other in a difference of 20 degrees.

P13

DETECTION OF ETHINYLESTRADIOL IN SEWAGE TREATMENT USING UV-VIS SPECTROSCOPY, MCR-ALS AND ICA

Cíntia M. Ritter, Suzana M. M. Curti, Letícia B. da Silva, Paulo H. Março, Patrícia Valderrama
Technological University Federal of Paraná (UTFPR), C.P. 271, CEP 87301-006, Campo Mourão-PR, Brazil
pativalderrama@gmail.com

The synthetic estrogen ethinylestradiol (EE) is an active component of oral contraceptives, considered as an endocrine disrupting compound. It is excreted from humans and released via sewage treatment plant (STP) effluents into aquatic environments. EE apparently affects the phosphate cycle, which is indicated by the increasing of phosphate concentrations in water and, as any environmental pollutant, once incorporated into an organism, EE affects the hormonal balance of various species including humans. Nonetheless, its presence in the environment is gaining considerable importance, mainly in water quality evaluation. In order to detect EE in STP, ultraviolet-visible spectroscopy (200-900 nm) coupled with Multivariate Curve Resolution with Alternating Least Squares (MCR-ALS) with column augmentation was proposed. To this, Independent Component Analysis (ICA) was used as initial estimative. The study was conducted in three months (March, April and May/2012) where two samples/months (three collect points each day) were collected in the same week. Pure spectra and kinetics of the EE present at STP's were recovered. The spectra recovered by MCR-ALS were identical to the spectra of EE reported in previous study. Others species like detergents and soups, edible oils residuals and organic matter were also present, and their spectra and kinetic were recovered and identified.

P15

HYBRID HARD-SOFT-MODELING OF UNFOLDING PROCESSES INVOLVING G-QUADRUPLEX AND *I*-MOTIF DNA STRUCTURES

Sanae Benabou and Raimundo Gargallo
Department of Analytical Chemistry, University of Barcelona, Martí i Franqués 1-11, 08028 Barcelona, Spain.
sbenabou_13@ub.edu

The stability of complex DNA structures like G-quadruplex or *i*-motif depends on experimental factors like pH, temperature or ionic strength, among others. Biophysical studies, such as melting experiments, provide information about the influence of these factors on the relative stabilities of these structures. In the case of melting experiments, heating induces DNA unfolding.

Traditionally, melting experiments have been monitored spectrophotometrically in a univariate way, i.e., being the result of the measurement a vector of absorbance values as a function of temperature. Appropriate univariate methods have been developed to analyze such data [1]. Modern spectrophotometers can record easily a collection of spectra as a function of temperature. Multivariate methods, when applied to these data sets, may be able to discern the existence of intermediates which may be unnoticed when melting experiments are monitored just from a univariate point of view.

In previous works [2,3], a mathematical procedure based on hard-modeling was proposed for the determination of [illegible] melting experiments. The method makes use of previously developed equations for univariate analysis. Here, a modification of the method, which is now based on hybrid hard-soft modeling, is proposed to deal spectral contributions not related to DNA species.

Acknowledgement: Funding from Spanish government is acknowledged.

References:

[1] K.J. Breslauer, Methods in Enzymology 259, 221 (1987).
[2] S. Fernandez, R. Eritja, A. Aviñó, J. Jaumot, R. Gargallo, Int. J. Biol. Macromol. 49, 729 (2011)
[3] R. Gargallo, Chemometrics in Analytical Chemistry (2012), poster number 138

MULTIVARIATE METHODS APPLIED TO THE ANALYTICAL STUDY OF THE INTERACTION OF THE STAINS-ALL DYE WITH DEOXYOLIGONUCLEOTIDES

Sanae Benabou, Daniel García, Ramon Eritja, Anna Aviñó, and Raimundo Gargallo
Department of Analytical Chemistry, University of Barcelona, Martí i Franqués 1-11, 08028 Barcelona, Spain.
sbenabou_13@ub.edu

Guanine- and cytosine-rich regions of DNA are capable of forming complex and characteristic structures known as G-quadruplex and *i*-motif, respectively. It has been shown the formation *in vitro* of such structures in DNA sequences corresponding to the end of telomeres and to the promoter regions of several oncogenes, such as *n-myc* [1]. The stability of folded structures of DNA may depend strongly on the interaction with inorganic and organic ligands. Hence, biomedical research is being carried out to find ligands which could modulate *in vivo* the stability of characteristic DNA structures, such as G-quadruplex and *i*-motif.

In previous studies it has been shown that the interaction of the cationic carbocyanine dye Stains-All with different DNA structures produces different colored dye-DNA complexes [2].
In the present study we have focused our attention on the study of the interaction of oligonucleotide sequences that could form complex DNA structures, such as G-quadruplex and i-motif, with this dye. Circular dichroism and molecular absorption spectroscopies have been used to monitor the experiments carried out. Multivariate data analysis based on soft- and hard-modeling methods has been used to recover qualitative and quantitative information about the species and conformations present in all experiments.

The results have shown an unexpected low stability of the dye in aqueous solution. Moreover, the dye has shown a high specific interaction with parallel intramolecular G-quadruplex and *i*-motif structures.

Acknowledgement: We acknowledge funding from both the Spanish government (CTQ2012-38616-C02-02) and Catalan government (2009 SGR 45).

References

[1] S. Benabou, R. Ferreira, A. Aviñó, C. González, S. Lyonnais, M. Solà, R. Eritja, J. Jaumot, R. Gargallo. Biochim. Biophys. Acta Gen. Subjects, Vol. 1840, 41-52 (2014).
[2] P. Campbell; H. MacLennan; O. Jorgensen. The Journal of Biological Chemistry, Vol. 258, No. 18, Issue of September 25, pp. 11267-11273 (1983).

P17

SECOND-ORDER ADVANTAGE WITH DATA LOSING THE BILINEARITY IN A SINGLE SAMPLE. A NOVEL NON-BILINEAR ADAPTED PARTIAL LEAST SQUARES/RESIDUAL MODELING METHOD

Agustina V. Schenone, María J. Culzoni, and Héctor C. Goicoechea

Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral – CONICET, Ciudad Universitaria, Santa Fe (S3000ZAA), Argentina
hgoico@fbcb.unl.edu.ar

The most challenging data structure problem to achieve analyte quantitation from second-order data in the presence of uncalibrated components using multivariate calibration methods is the loss in bilinearity, i.e. different profiles for a component in a single sample (the shape of the profile changes with the change in the other sensor). Total synchronous fluorescence spectroscopy (TSFS) generates matrices which constitute a typical example of this kind of data. An approximation of the non-bilinear profile of the interferent can be partially achieved by modeling TSFS data with unfolded partial least squares with residual bilinearization (U-PLS/RBL). A new residual modeling for non-bilinear data is here presented, namely unfolded partial least squares with residual modeling of non bilinear data (U-PLS/RMNB). Simulated data show that the new model can conveniently handle the studied analytical problem with better performance than U-PLS/RBL. One example involving TSFS matrices illustrate the ability of the new method to handle experimental data as well: the determination of the anticancer doxorubicin in human plasma samples.

P19

DETERMINATION OF A THRESHOLD FOR NOISE RECOGNITION IN 2D CORRELATION SPECTROSCOPY

Mirjam Schmidt and Matthias Otto

Institute of Analytical Chemistry, Technische Universität Bergakademie Freiberg, Leipziger Straße 29, 09599 Freiberg, Germany

mirjam.schmidt@chemie.tu-freiberg.de

2D correlation spectroscopy (2D COS) is a versatile concept to reveal correlations in a dataset obtained under a specific perturbation [1, 2]. However, the interpretation of the calculated correlation peaks is still subjective; especially when an important correlation and a correlation originating from noise have to be distinguished. The application of a noise filter by means of PCA [3] was suggested as preprocessing for 2D correlation spectroscopy to eliminate artifacts. But for our dataset 2D correlation spectra remain complicate even after PCA denoising with different numbers of principal components used. Thus the problem of separating significant correlations persists. A way to make this determination more objective is the calculation of a suitable threshold for noise recognition.

Our dataset originates from simulated aging of coal at 200 °C in air atmosphere which yielded ten samples analyzed by ATR-IR spectroscopy. After applying 2D COS the 2700 – 2000 cm^{-1} range in IR spectra was chosen to evaluate the threshold. The signal-noise- and noise-noise-correlation in 2D correlation spectra were considered to calculate the maximal noise level as threshold. All correlations above this noise level were defined to be meaningful and thus analyzable. Data points lower than the calculated threshold were displayed with white color in 2D correlation spectra. As a result 2D correlation spectra become much easier interpretable and the determination between signal and noise is more objective. Furthermore the suggested threshold determination is easily applicable to various different datasets obtained from e.g. chromatography or spectroscopy.

References

[1] Noda; Y. Ozaki. *Two-Dimensional Correlation Spectroscopy*. John Wiley & Sons Ltd., 2004.

[2] Noda. Two-dimensional correlation spectroscopy - biannual survey 2007-2009. *Journal of Molecular Structure*, 974:3–24, 2010.

[3] Y.M. Jung. Principal component analysis based two-dimensional (pca-2d) correlation spectroscopy: Pca denoising for 2d correlation spectroscopy. *Bull. Korean Chem. Soc.*, 24:1345–1350, 2003.

P20

CHEMOMETRIC BASED, COMPACT, IN-LINE FUEL SULFUR ANALYZER FOR IMPROVED MANAGEMENT OF DESULFURIZERS

Andrew L. Wagner, Ted J. Amundsen, and Paul E. Yelvington
Mainstream Engineering Corporation, 200 Yellow Place
Rockledge, FL 32955, USA
pyelvington@mainstream-engr.com

Fuel cell technology has progressed to the point where fuel desulfurization and reforming have become limiting factors. Both solid-oxide and polymer-electrolyte-membrane fuel cells are poisoned by sulfur compounds in transportation fuels. Sulfur removal is particularly challenging for military fuel-cell applications because JP-8 jet fuel contains up to 3,000 ppm of sulfur, 200 times the highway diesel limit. To more efficiently manage desulfurization processes, improved in-line methods of determining the extent of sulfur contamination in fuel are required. Mainstream Engineering is developing a compact, in-line analyzer for measuring total sulfur in jet fuel for fuel cell applications. This analyzer is intended to allow less frequent desulfurizer regeneration events when compared to the current scheme of scheduling regenerations based on the worst-case sulfur limit. The sampling system is non-destructive and allows measurement of sulfur directly in the liquid fuel. The measurement technique uses multivariate analysis of infrared absorption or Raman intensity spectra, which provides a robust calibration, outlier detection, and identification of fuel type. Models using PLS1, PLS2, PCR, and PLSDA were tested in combination with forward and reverse calibration transfer. Fuel samples were measured repeatedly over 18 months to evaluate the robustness and accuracy of the regression models and instrument hardware. Laboratory flow experiments with sulfur adsorbent columns, which mimic the end use of the analyzer, have been performed. Several candidate desulfurizer control strategies that leverage the continuous measurement of fuel sulfur concentration are under evaluation.

P21

IN-LINE RAMAN SPECTROSCOPY AND MULTIVARIATE CURVE RESOLUTION FOR MONITORING CARBAMAZEPINE-NICOTINAMIDE COCRYSTALS

Frederico L. F. Soares and Renato L. Carneiro
Department of Chemistry Federal University of São Carlos, São Carlos, 13565-905, SP, Brazil
fredlfsoares@gmail.com

Cocrystals are multicomponent substances designed by the addition of two or more different molecules, that when crystallized together results in a single crystal with crystallographic motifs different from their precursors. Cocrystallization process may involves several molecular species, which are generally solid at room conditions. Thus it is necessary an accurate monitoring of different components that might appear during the cocrystallization process. For cocrystal synthesis and monitoring, it was used a crystallization reaction procedure. Thus, it was prepared a nicotinamide solution in water with a concentration near saturation level. This solution was added to the carbamazepine, in solid state. The reactions were performed at different temperatures and monitored by the Raman spectroscopy. It was used multivariate curve resolution (MCR-ALS) to recover concentration profiles and spectra profiles of the different reactions. The reactions made at room temperature and 40 °C showed a quickly conversion of carbamazepine form III into its hydrate form, hindering the formation of the cocrystal. At 60 °C and 80 °C it was possible to observe a cocrystallization with a full conversion of the initial compounds into the cocrystal. The employment MCR-ALS coupled with Raman spectroscopy enabled to observe distinctly the main points of the reactions, such as drugs dissolution, crystal nucleation and cocrystal crystallization, which can be easily employed on a routine procedure in the pharmaceutical industry.

P22

EVALUATION OF COPOLYMERS CONVERSION USING RAMAN SPECTROSCOPY AND MULTIVARIATE CURVE RESOLUTION

Gabriella R. Ferreira[1], Frederico L. F. Soares[2], Renato L. Carneiro[2], Alexandre P. Umpierre[1], and Fabricio Machado[1]

[1]Institute of Chemistry, University of Brasília, Brasília, 70910-000, DF, Brazil
[2]Department of Chemistry, Federal University of São Carlos, São Carlos, 13565-905, SP, Brazil
fredlfsoares@gmail.com

The metal-containing monomers (MCMs) can be defined as metal complexes with specific ligands, which undergo free-radical polymerization. For the synthesis of these materials, dispersion copolymerizations were carried out in a reaction medium containing 17 wt.% of aluminum methacrylate (MAcAl) and 83 wt.% of vinyl monomers, such as vinyl acetate (VAc), ethyl acrylate (EA) and methyl methacrylate (MMA). Raman spectra were acquired at different reaction times for each used monomer. It was used multivariate curve resolution (MCR-ALS) to recover concentration profiles of the reactions. The concentration profile of VAc/MAcAl reaction shows that the polymerization occurs slowly, in which the conversion stabilizes after 15 hours of reaction. The concentration profile of EA/MAcAl copolymerization shows that high conversions are achieved in a short reaction time. However, as the reaction proceeds (between 4 and 12 hours), it is possible to observe a decrease of the reaction rate, which is an indication that the modified monomer is polymerizing. The concentration profile of MMA/MAcAl copolymerization shows a high conversion in the first 2 hours of reaction, indicating a possible low incorporation of the modified metal monomer into the growing copolymer chains, due to the differences of reactivity between the monomer species. The use of Raman spectroscopy coupled with MCR-ALS allowed the evaluation of the conversion rate of three different organic-inorganic copolymers, although it is difficult to determine the conversion of hybrid polymers, the use of chemometric methods and Raman spectroscopy allowed to compare the conversion rate between the studied copolymers.

P23

APPLICATION OF EPA-PMF TO MULTIPLE SITE PARTICLE COMPOSITION DATA

Melik Kara[1], Philip K. Hopke[2], Yetkin Dumanoglu[1], Hasan Altiok[1], Tolga Elbir[1], Mustafa Odabasi[1], Abdurrahman Bayram[1]

Department of Environmental Engineering, Dokuz Eylul University, Tinaztepe Campus, Buca-Izmir, TURKEY
Center for Air Resources Engineering and Science, Clarkson University, Box 5708. Potsdam, NY 13699 USA
phopke@clarkson.edu

Source apportionment has most often been applied to a time series of data collected at a single site. However, in a complex airshed where there are multiple sources, it may be helpful to collect samples from multiple sites to ensure that some of them have low values of specific source contributions such that edges can be properly defined. In this study, samples were collected at multiple sites in the Aliaga region (38º40'-38º54'N and 26º50'-27º03'E) located in the western part of Turkey on the coast of the Aegean Sea. This area contains a number of significant air pollution sources including five scrap iron-steel processing plants with electric arc furnaces (EAFs), several steel rolling mills, a large petroleum refinery, a petrochemical complex, a natural gas-fired power plant, a fertilizer plant, ship breaking yards, coal storage and packing, scrap storage and classification sites, large slag and scrap piles, heavy road traffic, very intense transportation activities including ferrous scrap trucks and busy ports used for product and raw material transportation. The newest version of EPA PMF (V5.0) has the capability of handling multiple site data. A total of 456 samples of PM_{10} in six sampling sites and 88 samples of $PM_{2.5}$ in one site were collected for four seasons. Eight factors were identified as iron-steel production from scrap (23.4%), re-suspended and road dust (23.3%), crustal (20.5%), marine aerosol (14.4%), biomass and wood combustion (7.2%), salvage activities (4.7%), coal combustion (3.7%) and residual oil combustion (2.8%). The pattern of source contributions and conditional probability function analysis were consistent with the locations of the known sources. Thus, the multiple site data allowed for a comprehensive identification of the primary sources of PM in this region.

P24

KINETIC STUDY OF UNCATALYZED BROMATE OSCILLATOR WITH PHENOL BY MULTIVARIATE CURVE RESOLUTION APPLIED TO UV-VIS SPECTRAL DATA

Iván F. Robayo, Jesús A. Ágreda
Department of Chemistry, Universidad Nacional de Colombia
Av. Cra. 30 # 45-03, Bogotá, Colombia.
ifrobayom@unal.edu.co

The Uncatalyzed Bromate Oscillator (UBO) is a well known chemical oscillator, but its reaction mechanism is still unclear due to the fact that this reaction has plenty of intermediates. Experimental techniques such as High Performance Liquid Chromatography and Ion Selective Electrodes have been used to unravel the reaction mechanism of UBO, but these techniques alone do not provide adequate information, since their analytical signals are not taken *in situ* or they produce only a single signal. However, UV-VIS absorption spectroscopy enables to measure, simultaneously, many species of the reaction, and this characteristic can be improved by the use of multivariate curve resolution. Therefore, in this work, the UBO, using Phenol as the organic substrate (UBO-Phenol), was studied with an UV-VIS spectrophotometer, and the spectra data resolved with the Multivariate Curve Resolution-Alternating Least Squares algorithm (MCR-ALS). The starting point for the MCR-ALS was the pure spectra of intermediate species proposed by György, Varga, Körös, Field and Ruoff (the GVKFR model). Simulations of spectral data of UBO-phenol intermediates were also done, when it was necessary. Previously to apply the MCR-ALS algorithm to the GVKFR model some improves were made by sensitivity analysis. The final results resolved the spectra of the intermediates species of the reaction and showed an excellent agreement between the experimental data matrices and the simulated matrices. In the near future, it is expected that this protocol will help to elucidate the dynamics of other oscillators with species that absorb in the UV-VIS region of the electromagnetic spectrum.

ON THE USE OF EXTENTS FOR PROCESS MONITORING AND FAULT DIAGNOSIS

Sriniketh Srinivasan, Julien Billeter, Dominique Bonvin
Laboratoire d'Automatique, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
julien.billeter@epfl.ch

Process monitoring and fault diagnosis are broadly used to control quality and enforce safety compliance in industrial processes. Processes are commonly monitored by online spectroscopy, with either PCA or calibration techniques such as PLS being used to predict abstract or physical process variables. By comparing these variables to historical data measured under normal operating conditions, possible faults are detected based on deviations from statistical thresholds [1]. One then tries to identify the causes of these faults, identification being easier when monitoring involves a calibration that predicts physical process variables and when fault detection uses a model to relate controlled and manipulated variables [2].

Exploiting the structure of balance equations, a transformation can separate multivariate data into decoupled variant/invariant states, which can be investigated individually to identify rate laws and reconstruct unmeasured quantities [3]. A convenient linear transformation uses the generalized concept of extents [4, 5], which coincides with a time-invariant transformation used to model rank-deficient spectroscopic data [6]. This transformation requires only limited process information, namely, the reaction stoichiometry, the species transferring between phases, the composition of inlet flows and the initial conditions. Moreover, this transformation was adapted to handle calorimetric and spectroscopic data [7, 8].

This contribution addresses the applicability of the transformation to extents for process monitoring and fault diagnosis. By comparing the extents computed from measurements of the current batch with either their prediction or the extents computed from previous batches, significant deviations can point at possible faults and provide a systematic way of identifying their causes.

References

[1] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, K. Yin, Comp. Chem. Eng. 27 (2003) 327
[2] V. Venkatasubramanian, R. Rengaswamy, K. Yin, S. N. Kavuri, Comp. Chem. Eng. 27 (2003) 293
[3] S. Srinivasan, J. Billeter, D. Bonvin, IFAC Proceedings Vol. 1 (2013) 102
[4] M. Amrhein, N. Bhatt, B. Srinivasan, S. Bonvin, AIChE J. 56 (2010) 2873
[5] N. Bhatt, M. Amrhein, D. Bonvin, Ind. Eng. Chem. Res. 49 (2010) 7704
[6] J. Billeter, Y.-M. Neuhold, K. Hungerbühler, Chemom. Intell. Lab. Syst. 95 (2009) 170
[7] S. Srinivasan, J. Billeter, D. Bonvin, Chem. Eng. J. 207-208 (2012) 785
[8] J. Billeter, S. Srinivasan, D. Bonvin, Anal. Chim. Acta 767 (2013) 21

P28

CHARACTERIZATION OF THE SOILPUCHUNCAVÍ (CENTRAL CHILE) USING THE PMF MULTIVARARIATE APPROACH

Sonnia Parra[1], Manuel Bravo[1], Waldo Quiroz[1] , Teresa Moreno[2], Angeliki Karanasiou[2]

[1]*Department of Analytical Chemistry and Environmental, Pontificia Universidad Católica de Valparaíso, Avenida Brasil 2950, Valparaíso, Chile*

[2] *Institute of Environmental Assessment and Water Studies''IDAEA", CSIC, C/Jordi Girona 18-26, 08034 Barcelona, Spain*. *Sonnia.parra@ucv.cl*

The use of statistical techniques applied to recognition and identification of sources contamination has become an increasingly important tool. The chemical composition of soil samples provides a datasets suitable for the application of these statistical techniques (Positive Matrix Factorization (PMF), and Principal Components Analysis (PCA)). In this paper we apply PMF to characterize the pollutant source in a set of soil samples taken in the Valley Puchuncaví (site characterized by a high concentration of heavy metals). Each sample has been analyzed for major and minor elements (using inductively coupled plasma atomic emission spectroscopy and inductively coupled plasma mass spectrometry); Mercury (using a atomic absorption spectroscopy with gold amalgam); nitrate and sulfate (using a high performance liquid chromatography and conductivity detection). Analysis of the soils using PMF resulted in a success allowing the identification of the chemical profile of each source and the relative contribution. Combining these results with a PCA approach successfully demarcated the main source of the heavy metal contamination in the soil of the Valley of Puchuncaví (Central Chile).

P29

MULTIVARIATE CURVE RESOLUTION MODELING OF LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY DATA IN A LIPIDOMIC STUDY OF JEG-3 CELLS EXPOSED TO CONTAMINANTS

Eva Gorrochategui [1], Sílvia Lacorte [1], Josefina Casas [2], Romà Tauler [1]

[1] *Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, 08034, Catalonia, Spain.*

[2] *Department of Biomedicinal Chemistry, Institute of Advanced Chemistry of Catalonia (IQAC), Barcelona, 08034, Catalonia, Spain.*

egmqam@cid.csic.es

A lipidomic study was developed in human placental choriocarcinoma JEG-3 cells exposed to tributyltin (0.1 □M) and a method was based on the application of multivariate curve resolution alternating least squares (MCR-ALS, [1]) to data sets obtained by ultra-performance liquid chromatography coupled to time-of-flight mass spectrometry (UPLC-TOF-MS). Lipids from exposed JEG-3 cells were solid-liquid extracted and analyzed by UPLC-TOF-MS in full scan mode, together with control samples. Raw UPLC-TOF-MS data of samples were organized in a column-wise augmented data matrix, where every sample was arranged in an individual data matrix (samples x elution times, m/z values). This augmented data matrix was subdivided into 20 distinct chromatographic regions, giving 20 new augmented data matrices which were then modeled by MCR-ALS. A total of 86 components were resolved and a statistical comparative study of their elution profiles showed distinct responses for the lipids of exposed versus control cells, evidencing a lipidome disruption attributed to the presence of the xenobiotics. Finally, PLS-DA analysis allowed the determination of more significant variables in prediction (VIPs, [2]) and, therefore, the potential biomarkers. Identification of those components was based on the combination of mass accuracy and high spectral accuracy parameters [3].

Acknowledgements: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 32073.

References

[1] Parastar, H. and R. Tauler, *Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: A new insight to address current chromatographic challenges.* Analytical Chemistry, 2014. **86**(1): p. 286-297.

[2] Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: A basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 109-130.

[3].Wang, Y. and M. Gu, *The Concept of Spectral Accuracy for MS.* Analytical Chemistry, 2010. **82**(17): p. 7055-7062.

P30

SIMCA AS A ONE-CLASS CLASSIFIER

A.L. Pomerantsev[1,2] **, O.Ye. Rodionova**[1]

[1]*Semenov Institute of Chemical Physics RAS, Kosygin 4, Moscow 119991, Russia*

[2]*Institute of Natural and Technical Systems RAS, Kurortny 99/18, Sochi, 354024, Russia*

forecast@chph.ras.ru

SIMCA, the method of soft independent modeling of class analogy, is known for 40 years [1]. This approach is a natural extension of the method of principal component analysis (PCA). SIMCA has been revised repeatedly [2-5]. Today it is very popular in analytical chemistry (chemometrics), but almost unknown outside. SIMCA provides a unique opportunity to make classification accounting both for the Type-I error □ (false rejection) and the Type □ (false acceptance) extremely rare. The SIMCA theoretical base is thoroughly developed, but most of the analytical studies contain gross errors, which are repeated from publication to publication.

The presentation is going to bridge the gap and to provide a general SIMCA concept as a data driven method. The following items will be considered:

- How PCA relates to SIMCA
- The distance measures in use: score distance, orthogonal distance, total distance
- What statistics are used in SIMCA and how this statistics are distributed
- How to estimate the parameters of these distributions
- How to make the decision at a given the Type I error □
- How to calculate the Type II error □
- How to make the decision at a given □

References

[1] Wold S. *Pattern Recognition* 1976; **8**: 127-139.
[2] Nomikos P, MacGregor JF. *Technometrics* 1995; **37**: 41–59
[3] Hubert M, Rousseeuw PJ, Vanden Branden K. *Technometrics*, 2005; **47**: 64-79
[4] Pomerantsev AL, *J. Chemometrics* 2008; **22**: 601-609
[5] Pomerantsev AL, Rodionova OY, *J. Chemometrics* 2013 (DOI: 10.1002/cem.2506)

P31

IN SILICO EXPERIMENTS DESIGNED TO EMULATE *IN VITRO* TESTS FOR ASSESSMENT OF ANTIOXIDATIVE POTENCY OF COMPOUNDS

Rok Martinčič,[1] Igor Kuzmanovski,[2] Alain Wagner,[3] and Marjana Novič[1]

[1] *Laboratory of Chemometrics, National Institute of Chemistry, Hajdrihova 19, POB 660, SI-1001, Ljubljana, Slovenia*

[2] *Institut za hemija, PMF, Univerzitet "Sv. Kiril i Metodij", PO Box 162, 1001 Skopje, Macedonia*

[3] *Laboratoire des Systèmes Chimiques Fonctionnels, UMR 7199, Faculté de Pharmacie, 74 route du Rhin, BP 24, 67401 Illkirch-Graffenstaden, France*

marjana.novic@ki.si

Reactive oxygen species (ROS) can cause serious cell damage and can compromise the integrity of the body's basic structures and functions. Nonlethal DNA mutations may cause cancers, replication and transcription errors may cause viral interactions, while irreparable DNA damage leads to apoptosis. Radiation and UV light exposure is one of the largest promoters of apoptosis. Antioxidants are molecules that can act as free radical scavengers and are thus investigated as effective radioprotective substances. In this study we have investigated a set of natural and synthetic compounds that show antioxidant potency and can be used for protection against radiation. Experimentally determined antioxidative potency is measured as a percentage of intact thymidine after oxidation in Fenton, UV or gamma assays. Since the *in vitro* assays demand a huge amount of work, we have developed *in silico* models based on available experimental data [1]. In order to obtain optimal new extended models we have used and compared three different modelling methods, multiple linear regression (MLR), counter-propagation artificial neural network (CP-ANN), and support vector machines (SVM). The models based on the SVM method show the best predictive power. The errors in prediction of thymidine protection [%] for test compounds were 14%, 6% and 14% for the Fenton, UV and gamma radiation assays, respectively. The model has been developed following the OECD principles for QSAR models [2], including a new approach to the assessment of the domain of applicability, which has not been reported in the literature yet for the SVM method.

Acknowledgement: The financial support by the EU inter-regional Slovenia-Italy structural funds (T2C project) and Ministry of Higher Education, Science and Technology (P1-017) is acknowledged.

References

[1] A. le Roux, I. Kuzmanovski, D. Habrant, S. Meunier, P. Bischoff, B. Nadal, S. A.-L. Thetiot-Laurent, T.le Gall , A. Wagner, M. Novič, J. Chem. Inf. Model., 51, 3050–3059 (2011)

[2] OECD Guideance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. ENV/JM/MONO(2007)2. www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en. Accessed January 2013.

P32

PRE-PROCESSING STRATEGY FOR LARGE DATASETS: ANALYSIS OF MULTI CAPILLARY COLUMN – ION MOBILITY SPECTROMETRY (MCC-IMS) DATA WITHIN THE ALBERT PROJECT

Ewa Szymańska[1,2], Emma Brodrick[3], Mark Williams[3], Jan Gerretzen[1,2], Femke Reijnen[1,4], Edwin R. van den Heuvel[4], Eric Brouwer[5], Eduard P.P.A. Derks[6], Jeroen J. Jansen[2], Geert Postma[2], Antony N. Davies[3,7], Henk-Jan van Manen[7], and Lutgarde M.C. Buydens[2]

[1]TI-COAST, P.O. Box 18, 6160 MD Geleen, The Netherlands
[2]Radboud University Nijmegen, Institute for Molecules and Materials (IMM), P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
[3]School of Applied Sciences, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, CF37 1DL, UK
[4]University Medical Center Groningen, The Department of Epidemiology, PO Box 30.001, 9700 RB Groningen
[5]Heineken Supply Chain BV, R&D, P.O. Box 510, 2380 BB Zoeterwoude, The Netherlands
[6]DSM Resolve, Process Analysis & Statistics, P.O.Box 18, 6160MD Geleen, The Netherlands
[7]AkzoNobel N.V., Supply Chain, Research & Development, Expert Capability Group - Measurement & Analytical Science, P.O. Box 10, 7400 AA, Deventer, The Netherlands
E.Szymanska@science.ru.nl

The Analysis of Large data sets By Enhanced Robust Techniques (ALBERT) project is developing generic strategies and methods to facilitate better and more robust chemometric and statistical analysis of complex analytical data. ALBERT is a multidisciplinary project studying diverse data from a range of analytical techniques provided by three TA-COAST consortium multinational industrial partners: AkzoNobel, DSM Resolve, and Heineken.

The MCC-IMS (multi-capillary column - ion mobility spectrometry) data set is an example of the large data sets within the ALBERT project. The MCC-IMS is capable of detecting volatile analytes at low ppb to ppt levels without the need for ultra-high vacuum mass spectrometers. The particular data sets used in this study have been generated with the main aim of finding patterns characteristic for different groups of air and breath samples, however, application of standard chemometric tools is hampered by the high dimensionality and redundancy of this data set. An efficient data pre-processing strategy is therefore required to cope with data peculiarities and to enable further analysis. The developed strategy in the ALBERT project has included several data pre-processing methods including background correction, alignment, denoising and compression by wavelet transform as well as variable selection by mask construction and sparse partial least squares-discriminant analysis (sparse-PLS-DA). Important pre-processing aspects such as efficiency of data reduction and level of information loss have been evaluated and will be discussed. Finally, the generality of the optimized strategy has been assessed in relation to other data sets available within the ALBERT project.

P34

INTEGRATION OF NEURAL NETWORK TECHNIQUE WITH RESPONSE SURFACE METHODOLOGY FOR OPTIMIZATION OF TECHNOLOGICAL PROCESS OF PIGMENT DYING

Natalja Fjodorova[1], Marjana Novič[1], and Tamara Diankova[2]
[1]*National Institute of Chemistry, Hajdrihova 19, SI- 1000, Ljubljana, Slovenia;*
[2]*St-Petersburg State University of Technology and Design, Bolshaya Morskaya st. 18, 191186, St. Petersburg, Russia*
natalja.fjodorova@ki.si

The competitive environment of today's marketplace in many manufacturing and service industries caused the increase in the popularity of experimental design strategies and optimization methods. The bottleneck feed-forward neural network (FFBN) as a mapping technique provides 2D map with setting operational input parameters overlapped with response output values at the same coordinates. This phenomenon makes possible the 2D visualization of technological process with multiple optima in the whole studied process. Integration of the FFBN mapping technique with the surface response design (SRD) offers multiple optima and ensures more reliable results. Implementation of both techniques in the process of dying of the aramid fibers is demonstrated in the study. It is illustrated how the response optimizer of a SRD model evaluated the desirability of all optimum solutions obtained using FFBN.

P35

THE CHEMOMETRIC METHODS IN ELECTRONIC TONGUE SYSTEM FOR RECOGNITION AND CLASSIFICATION OF PHARMACEUTICAL SAMPLES

M.Wesoły[1], J. Lisiecka[1], K. Sołłohub[2], K. Cal[2], W. Wróblewski[1], P. Ciosek[1]

[1] *Department of Microbioanalytics, Warsaw University of Technology, Noakowskiego 3, 00 – 664 Warsaw, Poland*

[2] *Department of Pharmaceutical Technology, Medical University of Gdansk, Hallera 107, 80-416 Gdansk, Poland*

pciosek@ch.pw.edu.pl

Electronic Tongues (ETs) are systems composed of a set of sensors (a sensor array) and a data analysis system (Pattern Recognition block, PARC) that allows one to extract useful information from sensor responses. They are dedicated to qualitative and quantitative analysis of liquid samples with complex matrix. The aim of such analysis is the recognition and classification of samples images.

The majority of Active Pharmaceutical Ingredients (APIs) present in oral drug products have a bitter taste. Appearance of such taste in drug formulations is undesirable and frequently influences negatively the efficiency of pharmacotherapy, especially a pediatric one.

The main goal of this study was to assess the effectiveness of taste masking in pharmaceutical samples using electronic tongue. Microencapsulation was chosen as a powerful tool for taste masking of various APIs. Eudragit L30D-55 and Eudragit L30D-55 with addition of SLS were used as a taste masking coat for Ibuprofen (API). Sensor array composed of standard ion-selective electrodes (ISEs) was used to form chemical images of samples before and after microencapsulation.

Principal Components Analysis (PCA) and Partial Least Squares (PLS) were applied to compare chemical images of the samples in order to detect the microencapsulation effect of API, which influences taste properties of pharmaceuticals. Application of chemometric methods to study taste – masking effect in Ibuprofen formulations measured with potentiometric sensor array was showed. The performed analysis confirmed the suitability of electronic tongue for the investigation of encapsulation efficiency.

Acknowledgement: This work has been supported from the financial resources for science in years 2013 – 2017 as a research project within a framework of "Diamond Grant" programme and by National Science Centre within a framework of OPUS project "Sensor arrays for the study of the release process of active pharmaceutical ingredients and excipients from pharmaceuticals".

P36

IMPROVING OUTLIER DETECTION BY FUSION OF OUTLIER DETECTION MERITS USING SUM OF RANKING DIFFERENCES

Brett R. Brownfield, John H. Kalivas
Department of Chemistry, Idaho State University,
Pocatello, ID 83209, USA
browbre2@isu.edu; kalijohn@isu.edu

Outlier detection is a key step in building a multivariate calibration model. Also important is determining if a new sample to be predicted by a calibration model is an outlier to the calibration domain. Outliers, especially in high-dimensional space, can be difficult to detect. Mahalanobis distance (MD) is a common merit used to identify outliers. However, two issues exist with MD. First, MDs are sensitive to the outliers themselves. Second, when calculating MDs, an inverse of the calibration covariance matrix is used and hence, MDs are typically eigenvector dependent. Sum of ranking differences (SRD) is a new tool that can simultaneously evaluate multiple merits. For example, instead of having to decide on how many eigenvectors to span the calibration space for the MD inverse calculation, all respective eigenvector based inverses can be simultaneously used with SRD. In combination with MD, additional merits can be used with SRD to assess samples as outliers. Examples are other distance measures, angles, and Procrustes analysis indicators which all measure similarities between a sample and the calibration space. This poster presents results from using multiple outlier merits with SRD to evaluate near infrared (NIR) spectroscopic calibration sets for outliers along with examining new validation samples for outlier behavior.

P37

VALIDATION OF PLS MODELS FOR BIODIESEL QUALITY PARAMETERS DETERMINATION IN DIESEL BY INFRARED SPECTROSCOPY

Werickson F.C. Rocha[1], Maurício G.Fonseca[1], Claudete N Kunigami[2], Luciano N Batista[1], Viviane F da Silva[1]

[1]National Institute of Metrology, Quality and Technology (Inmetro), Directorate of Industrial and Scientific Metrology, Chemical Metrology Division, 25250-020, Xerém, Duque de Caxias, RJ, Brazil

[2]National Institute of Technology (INT), Division of Analytical Chemistry, 20081-312, Rio de Janeiro, RJ, Brazil.

wfrocha@inmetro.gov.br

The application of analytical procedures based on multivariate calibration models has been limited in several areas due to requirements of validation and certification of the model. Procedures for validation are presented based on the determination of figures of merit, such as precision, accuracy, sensitivity, analytical sensitivity, selectivity, signal-to-noise ratio and confidence intervals for PLS models. An example is discussed based on the determination biodiesel quality parameters in diesel by diffuse reflectance spectroscopy. The results show that multivariate calibration models can be validated to fulfill the requirements imposed by industry and standardization agencies.

The models were developed for the following parameters: heating value, Iodine index, acid number index, density and carbon residue using 16 samples containing the biodiesel in diesel in the concentration range of 2% (v/v) to 10% (v/v).The diffuse reflectance spectra were obtained in a triplicate between 4000-10000 cm^{-1} at a resolution of $4cm^{-1}$. In the construction of the calibration models it was used the PLS-Toolbox-3.5, being employed 10 samples for the calibration and 6 for the validation sets. The pretreatment used was the multiplicative scatter correction.

This work demonstrated that purposed method is: (a) valid to determine quality parameters of biodiesel in diesel; (b)simple; (c) rapid; (d) sensitive; (e) economic.

Acknowledgements: The author thanks LAMOC (Laboratório de Motores e Combustíveis) for donating raw material for the studies. This research is entirely supported by INMETRO (Brazil).

P38

EXPLORING & ANALYZING GENOME WIDE ASSOCIATION DATA, A MULTIVARIATE APPROACH

P. Singh[1,2], J. Engel[2], J. Jansen[2],J. R. de Haan[1], and L. M. Buydens[2]
[1]Department of Bioinformatics, Genetwister Technologies B.V., Wageningen, Netherlands
[2]Department of Analytical Chemistry, Radboud University, Nijmegen, Netherlands
p.singh@genetwister.nl

Plant breeders are highly interested in finding genetic source of phenotypic variations. The advent of high throughput sequencing technology enables rapid re-sequencing of plant genomes & thus identification of SNPs, which brings this relation in direct reach. Genome wide association studies (GWAS) are now widely used to find the specific SNPs closely associated with phenotypes. Unfortunately, GWAS data are extremely high dimensional & complex. Presence of many uninformative predictive variables (noise), categorical nature, structure and family relatedness in data makes it more challenging to analyze and interpret. Genotype –environment interactions and abundance of missing values, sometimes add another dimension of complexity to the data.

Univariate techniques such as mixed-model are often used by researchers to analyze GWAS data. However, these models don't take the multivariate nature of data into account. Common used multivariate techniques such as PCA & partial least square (PLS) are however, often unable to handle above mentioned properties of GWAS data.
To address the above mentioned issues with GWAS data, we have investigated applicability of several multivariate techniques. In this study, we reviewed and compared the value of several (genetic) distance measures in combination with common chemometric techniques such as PLS to find relation between SNPs and phenotypes.

References

[1] A. Korte, A. Farlow, Plant Methods, 9:29 (2013)
[2] V Segura et al, Nature Genetics, 44, 825 (2012)
[3] X Zhou, M. Stephens, Nature genetics, 44, 821 (2012)
[4] H. Chun, D Ballard, J Cho, H Zhao, Genetic Epidemiology, 35, 479 (2011)
[5] O Libiger, C.M. Nievergelt, N.J. Schork, Human Biology, 81(4) , 389 (2009)

P40

STRATEGIES TO EVALUATE SIGNIFICANT FACTORS IN PLACKETT-BURMAN DESIGNS APPLIED IN GCxGC-qMS

Luciana F. Oliveira, Soraia C. G. N. Braga, Paulo R. Filgueiras, Fabio Augusto, and Ronei J. Poppi

Institute of Chemistry, State University of Campinas (Unicamp), CP 6194, 13084-971 Campinas, São Paulo, Brazil.
lufontesoliveira@gmail.com

The use of Designs of Experiment (DOE) has a significant application in analytical chemistry and it can found in literature different strategies to evaluate these designs. Plackett-Burman designs are a class of saturated designs that allows study at most 4n-1 (n=number of experiments) factors. In this study, half-normal probability plots, statistical interpretation based on dummies variables, Dong algorithm and permutation tests were applied to evaluate Plackett-Burman designs used to assess the significance of the results for determination of relevant parameters on the robustness of GCxGC-qMS analysis. In a first Plackett-Burman design the variables studied were: carrier gas flow rate, modulation period, temperature of ionic source, MS photomultiplier power, injector temperature and interface temperature. In a second design, the variables were: minimum and maximum limits of the scanned mass ranges, ions source temperature and photomultiplier power. The use of different strategies to analyze the effects lead to slightly different conclusions in the first design: while all strategies agree that photomultiplier power is significant, variations on the carrier gas flow rate were pointed out as significant in half-normal plots and confidence limits calculated through dummy variables, but results from Dong algorithm and permutation tests indicate that this parameter does not affect the quantitative robustness. For other study (second design), all strategies agree that photomultiplier power and lower limit of mass scanning range all have an impact upon robustness.

P41

A DATA FUSION METHOD EXPLOITING THE MULTIVARIATE ADVANTAGE

B.P. Geurts, J. Engel, B. Rafii, J.J. Jansen, L.M.C. Buydens
Radboud University Nijmegen, Institute for Molecules and Materials, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
b.geurts@science.ru.nl

The aim of fusing datasets from multiple sources or analytical platforms is to efficiently use all available information to improve the prediction ability. We will exploit the multivariate advantage to the fullest by taking into account the relations between variables, both from the same and different datasets. Current methods for data fusion struggle either with a large number of variables (low-level fusion) or with possible information loss in a dimension reduction (mid-level fusion). With this poster we introduce an improved method for data fusion that overcomes the drawbacks of both and combines their strengths.

Our method is a hybridization of low-level and mid-level fusion. In low-level fusion the data blocks are simply concatenated, where mid-level fusion includes a preceding dimension reduction [1]. Our method is midway between these two, because a dimension reduction takes place but the discarded information is subsequently explored for aspects that can still improve prediction. This allows us to combine data from multiple sources with a classification accuracy that exceeds the accuracy of low-level and mid-level fusion in the case of many variables. The method also provides more insight in the relationship between variables in different datasets.

This poster shows that our method outperforms conventional low- and midlevel fusion in prediction accuracy for both simulated and real data. We show that the novel method can achieve better prediction in scenarios that differ in data size, the presence of noise and of partial correlations between variables. Our method improves the integration of diagnostic information from multiple sources.

Reference

[1] Hall, D. L., *Mathematical Techniques in Multisensor Data Fusion*. Artech House: Boston, MA, 1992.

P42

MULTIVARIATE ANALYSIS OF DEEP-ULTRAVIOLET RESONANCE RAMAN AND CIRCULAR DICHROISM SPECTROSCOPIC DATA FOR SECONDARY STRUCTURE DETERMINATION

Olayinka O. Oshokoya and Renee D. JiJi
Department of Chemistry, Columbia, University of Missouri,
601 S. College Avenue, Columbia, MO 65211 USA
ooo294@mail.missouri.edu

Determination of protein secondary structure (α-helical, β-sheet, and disordered motifs) has become an area of great importance in biochemistry and biophysics as protein secondary structure is directly related to protein function and protein related diseases. While NMR and x-ray crystallography can predict placement of each atom in proteins to within an angstrom, optical methods (CD, Raman, IR) are the preferred techniques for rapid evaluation of protein secondary structure content. Such techniques require calibration data to predict unknown protein secondary structure content where accuracy may be improved with the application of multivariate analysis. We compare protein secondary structure predictions obtained from multivariate analysis of ultraviolet resonance Raman (UVRR) and circular dichroism (CD) spectroscopic data using classical and partial least squares, and multivariate curve resolution-alternating least squares is made. Based on this analysis, the suggested best approach to rapid and accurate secondary structure determination is a combination of both CD and UVRR spectroscopy.

A disadvantage of multivariate calibration methods is the requirement of known concentration or spectral profiles. Second-order calibration methods, such as parallel factor analysis (PARAFAC), do not have such a requirement due to the “second-order advantage”. An exceptional feature of UVRR spectroscopy is that UVRR spectra are also dependent on excitation wavelength as they are on secondary structure composition. Thus, higher order data can be created by combining protein UVRR spectra of several proteins collected at multiple excitation wavelengths. Here, PARAFAC has been used to analyse UVRR data collected at multiple excitation wavelengths on several proteins to determine secondary structure content.

P43

CONFIDENCE INTERVALS FOR SUPPORT VECTOR REGRESSION BY BOOSTING TYPE ENSEMBLE METHOD

Paulo R. Filgueiras[1], Luciana A. Terra[1], Eustáquio V. R. de Castro[2], Lize M. S. L. de Oliveira[3], Júlio C. M. Dias[3], Ronei J. Poppi[1]

[1]Institute of Chemistry, University of Campinas, Campinas–SP, P. O. Box 6154, 13083-970, Brazil.
[2]Laboratory Research and Development of Methodologies for Analysis of Petroleum (LabPetro), Department of Chemistry, Federal University of Espírito Santo – UFES, Vitória–ES, Brazil.
[3]CENPES/PETROBRAS, Rio de Janeiro–RJ, Brazil.
pauloiuna@hotmail.com

Currently there is increasing use of multivariate calibration methods in analytical chemistry, especially nonlinear regression methods such as support vector machines. However, in most analytical applications, beyond the predicted result, it is very important to provide an estimate of the confidence interval. This study aim to estimate the T10%, T50% and T90% of distilled volume in crude oils samples by ^{1}H NMR spectra and support vector regression (SVR). This analysis used 35 crude oil samples split into 24 calibration and 11 prediction samples. The confidence interval of the prediction samples was estimated by a boosting type ensemble method. In this methodology, the calibration samples are weighted during the development of the model, and its final weight is proportional to its contribution to the final model. The models presented reliable confidence intervals for the three measured properties. It was also possible to identify outliers in calibration samples by evaluating the weight of the boosting method, improving the accuracy of the prediction errors in the calibration models.

P44

FT-IR MICROSPECTROSCOPY AND CHEMOMETRICS TOOLS FOR EVALUATION OF LIGNOCELLULOSIC COMPONENTS DISTRIBUTION ON SURFACES OF WOOD PULPS

Rosario del P. Castillo[1,2] **and Juanita Freer** [2,3]

[1]*Faculty of Pharmacy, University of Concepcion.* [2]*Biotechnology Center, University of Concepcion.* [3]*Faculty of Chemical Sciences, University of Concepcion.*

Barrio Universitario s/n, Universidad de Concepción Chile.

rosariocastillo@udec.cl

Distribution of lignocellulosic components including cellulose, lignin and hemicelluloses and its effect on the simultaneous saccharification and fermentation process (SSF) in the production of bioethanol from wood was analyzed by application of multivariate methods on infrared spectra of micro-images obtained by attenuated total reflectance (ATR) - Fourier Transform Infrared microspectroscopy of pretreated *Pinus radiata* wood. Principal Component Analysis (PCA) and Alternating Least Squares (ALS) were used to evaluate the spectra and to generate the distribution images. Results show that chemometrics analysis facilities the identification of three principal components of the pulps in a fast way, and improve the results obtained by analysis of individual bands of spectra. Therefore, reconstructed images show that components distribution affects the SSF process. Combination of these techniques could be used as fast methodology to know in a first approximation the lignocellulosic components distribution of wood substrates for SSF process in bioethanol production, where chemical information is included unlike other microscopy techniques.

Acknowledgments: Authors thank the financial support of Fondecyt 11130388 project and to "Centro de Microscopía Avanzada (CMA Bio-Bio, ECM-12)" of the University of Concepcion, Chile.

P45

MULTIVARIATE RESOLUTION METHODS IN RAMAN IMAGING APPLIED TO EXPLOSIVES DETECTION

Mariana R. Almeida[1], *Deleon N. Correa*[1], *Jorge J. Zacca*[2], *and Ronei J. Poppi*[1]
[1]*Institute of Chemistry, University of Campinas, POB 6154, 13084-971, Campinas, SP, Brazil*
[2]*National Institute of Criminalistcs, Brazilian Federal Police, SAIS Quadra 07 Lote 23, 70610-200 Brasília, Distrito Federal, Brazil*
mariana.almeida@iqm.unicamp.br

The detection of explosives and their precursors, especially in the presence of complex backgrounds, has been a significant focus of research in the forensic field. In this study, the results obtained by independent component analysis (ICA) and multivariate curve resolution (MCR) applied in Raman imaging data of the banknotes after ATM explosions using the following explosives: TNT, HMX, PETN, RDX, ammonium perchlorate, picric acid, dynamite and gunpowder were compared. Data analyses were performed with MF-ICA and MCR-ALS algorithms using non-negativity constraints in concentration and spectral profile. The solutions were evaluated by variance explained, lack of fit and similarity criterion. The solutions obtained by MF-ICA and MCR-ALS were similar, except for TNT. For TNT it was observed that the profile obtained by MCR-ALS disagreed from reference profile, probably due to the low signal contribution of this component to the global model. To overcome this problem, interval Multivariate Curve Resolution, iMCR, was used and recovered the spectra correctly. In the set data studied, MF-ICA algorithm presented advantages when one of the components has low concentration. Moreover, the MCR method allows more constraints than ICA method. The main difficulty of ICA is the determination of the number of independent components, since it is fundamental for correct final solution.

P46

DETERMINATION OF INTERSECTING LINES IN QUESTIONED DOCUMENTS BY SURFACE-ENHANCED RAMAN SPECTROSCOPY IMAGING AND MCR-ALS

Mariana R. Almeida, Carlos Teixeira, Deleon N. Correa, and Ronei J. Poppi
Institute of Chemistry, University of Campinas, POB 6154, 13084-971, Campinas, SP, Brazil
mariana.almeida@iqm.unicamp.br

Analysis of questioned documents to reveal the sequence of intersecting lines has been a major request for forensic documents examiners. In this work, surface-enhanced Raman spectroscopy (SERS) imaging was used to assess the order in which intersecting graphic lines was drawn. Multivariate curve resolution (MCR) was required to resolve and identify SERS spectra and to create images. The samples studied were intersecting graphic lines made by ballpoint pens in different combinations and intersections. The intersections were created using different ballpoint pens and also junctions consisting of a line drawn by one ballpoint pen and a line created by a laser printer. Raman chemical imaging was enhanced by applying SERS using colloidal gold as substrate. Spectra were decomposed using the MCR-ALS method, and pure contributions were recovered. It was observed that when the pen ink line is below of the printer line, there is a gap in the concentration map of the pen ink in the intersection region, showing that printer ink was above. When, the pen ink is above of the printer ink the opposite behavior was observed, the concentration was low for the printer ink in the intersection region. For the intersection region with two inks pens, there is a predominance of the ink that is above in concentration map. The methodology presented proved to be a promising tool for determination of sequence of intersecting lines in questioned documents. Due to similarity in most commercial ink compositions, the use of multivariate curve resolution was necessary.

P47

EVALUATION OF MSPC TECHNIQUES AND MCR-ALS TO MONITOR THE BIODIESEL TRANSESTERIFICATION REACTION USING NIR SPECTROSCOPY

Rafaella F. Sales[a], Carolina S. Silva[b], Neirivaldo C. da Silva[b], Alianda D. de Oliveira[b], Dácio Vieira[a], Suzana M. de Lima[a], M. Fernanda Pimentel[a]

[a]Department of Chemical Engineering, Universidade Federal de Pernambuco, Rua Prof. Arthur de Sá S/N, Cidade Universitária, 50740-521 - Recife, PE, Brazil

[b]Department of Fundamental Chemistry,Universidade Federal de Pernambuco, Avenida Prof. Moraes Rego, 123, Cidade Universitaria, 50670-901 - Recife, PE, Brazil

rafaelladfs@gmail.com,carolina.santossilva@ufpe.br, neirivaldocavalcante@gmail.com, aliandaquim@hotmail.com, dacio.vieira@gmail.com, suzanamdl@gmail.com, mfp@ufpe.br

Biodiesel synthesis by transesterification is a two phase complex reaction of oil and alcohol, in which the main products are ester (biodiesel) and glycerin. The reaction is usually carried out in batch reactors and various factors affect the process: variability of feedstocks, type and speed of stirring, concentration and type of catalysts and temperature. Real time monitoring of transesterification allows a more efficient process control and undertanding. In this work, two chemometrics strategies were evaluated to treat the spectral data obtained from transesterification reactions of soybean oil with methanol, using in-line NIR spectroscopy. The reactions were carried out in a 500 mL batch reactor. In the first strategy, multivariate control charts based on latent variables were built using batch-wise and variable-wise approaches. Batches “out of control” were produced to detect different types of fault (raw material composition, temperature and agitation controls). In the second strategy, batches with different temperatures (20^0C, 45^0C and 55^0C) and catalyst concentrations (NaOH – 0.5%, 0.75% and 1%) were produced. MCR-ALS was employed to obtain the concentration profiles of each reaction component (oil, ester, methanol and glycerin) along the reaction. Unlike the variable-wise, the batch-wise approach was able to identify the reactions with different feedstocks. The variable-wise approach identified batches where transient changes in temperature and agitation occurred. Using MCR-ALS and SNV, the relative concentration profiles obtained followed the expected variation from each component, allowing observation of kinetic behavior and the equilibrium time under each experimental condition.

P48

EVALUATION OF NONLINEAR FEATURE EXTRACTION METHODS FOR VIBRATIONAL SPECTROSCOPIC DISCRIMINATION OF GEOGRAPHICAL ORIGINS FOR AGRICULTURAL SAMPLES

Sanguk Lee,[1] Hyeseon Lee[2] and Hoeil Chung[1]

[1] *Department of Chemistry, Hanyang University, Seoul, 133-791, Korea*

[2] *Department of Industrial & Management Engineering, Pohang University of Science and Technology, San 31 Hyojadong, Pohang 790-784, Korea*

hoeil@hanyang.ac.kr

Nonlinear dimensionality reduction methods, such as supervised neighborhood preserving embedding (SNPE), were employed to represent near-infrared (NIR) and Raman spectral features of agricultural samples (Angelica gigas, sesame and red pepper) and the constructed variables were used for discrimination of their geographical origins. These are enabling of recognizing non-linear spectral behaviors and minute spectral differences among classes by simultaneously preserving local relationships, so could be an alternative to widely adopted linear feature extraction (representation) methods such as principal component analysis (PCA) and partial least squares (PLS). For this purpose, diffuse reflectance NIR spectral datasets of Angelica gigas, sesames and red peppers, and Raman spectral dataset of the same red peppers were prepared. The spectra were represented into new variables (coordinates) using PCA, PLS, NPE (neighborhood preserving embedding) and SNPE, and then these were subsequently used to discriminate corresponding samples into two groups of either imported or domestic sample using *k*-nearest neighbor (*k*-NN) and support vector machine (SVM). The combination of SNPE and SVM provided the slightly better accuracy in discriminating geographical origins for the tested samples.

P49

KERNEL-BASED PREDICTION MODEL FOR QUANTIATIVE ANALYSIS BASED ON VIBRATIONAL SPECTROSCOPY

Junghye Lee,[1] Sanguk Lee,[2] Hyeseon Lee,[a] Hoeil Chung[2] and Chi-Hyuck Jun[1]

[a] *Department of Industrial & Management Engineering, Pohang University of Science and Technology, San 31 Hyojadong, Pohang 790-784, Korea*

[2]*Department of Chemistry, Hanyang University, Seoul, 133-791, Korea*

hyelee@postech.ac.kr

Kernel-based methods often find better feature space mapping which represents the relationship between target variable and the original input space. To evaluate the potential of kernel-based methods in chemometrics-based quantitative analysis, four different vibrational spectroscopic datasets (NIR dataset of naphtha, Raman dataset of naproxen tablet, Raman dataset of lube oil and Raman dataset of polyethylene pellet) were employed. For each spectral dataset, kernel-based partial least squares (K-PLS) regression was used to determine either sample composition or physical property. For kernel function, radial basis function (RBF), polynomial and sigmoid kernel were used, and compared with linear based model. The optimal parameters for the kernel functions were found by the grid search method. K-PLS with RBF kernel provided the improved accuracies in comparison with those resulted from other kernel functions. It was originated from the fact that RBF handled nonlinear relationship better with input space and needed fewer hyperparameters than polynomial and sigmoid kernel, especially for small number of features. Overall results suggest that K-PLS with RBF kernel is worthwhile to adopt as an alternative quantitative chemometric tool along with PLS, a most widely accepted method.

P50

CALIBRATION UPDATE STRATEGIES FOR AN ARRAY OF POTENTIOMETRIC CHEMICAL SENSORS

Alisa Rudnitskaya[1], Ana Maria S. Costa[1], Ivonne Delgadillo[2]
[1]*CESAM and Chemistry Department, Aveiro University, Campus Universitario de Santiago, Aveiro, 3810-193 Portugal.*
[2]*QOPNA and Chemistry Department, Aveiro University, Campus Universitario de Santiago, Aveiro, 3810-193 Portugal*
alisa.rudnitskaya@gmail.com

The electronic tongues – multisensor systems based on the arrays of cross-sensitive sensors and data processing tools - have been shown to be promising analytical instruments for a wide range of applications. One of the problems hindering practical use of the ETs is temporary drift of the sensors, *i.e.* gradual change of the sensor characteristics occurring in the process of their exploitation. Two approaches can be employed to deal with the sensor drift. One consists of regular re-calibration of the sensor array, which is effective but time and labor consuming. Alternatively, statistical methods can be used. While significant efforts have been directed to the development of the calibration transfer and update techniques, they were mostly applied to the near infrared spectroscopic instruments. Very few works addressed this issue for the potentiometric sensor arrays. In the present study, update of the calibration model for the lead quantification in multicomponent solutions using an array of 10 potentiometric sensors will be considered. Calibration update will be performed using slope/bias correction, expansion of PLS regression model, Tikhonov regularization and autoassociative back-propagation artificial neural networks and results will be presented and discussed. Suitability of different approaches for calibration maintenance of the array of potentiometric chemical sensors will be discussed.

P51

IMPROVED ALGORITHM FOR TUMOR TYPE IDENTIFICATION WITH RAPID EVAPORATIVE IONIZATION MASS SPECTROMETRY

Peter Varga[1], Julia Balog[1,2], and Zoltan Takats[2]

[1] Medimass ltd, 2. Remenyi Ede street, Budapest, Hungary 1033

[2]Department of Surgery and Cancer, South Kensington Campus, Imperial College London, London SW7 2AZ

peter.varga@medimass.com

The recently developed rapid evaporative ionization mass spectrometry opens the possibility of creating a novel instrument for in-situ MS guided surgery. The technique is based on the mass spectrometric analysis of the aerosol released during standard surgical procedures. The in-vivo tissue identification requires a histologically validated tissue specific spectral database and a multivariate classification algorithm, which can be used in milliseconds during surgery. The currently used algorithm, including principal component analysis (PCA) and linear discriminant analysis (LDA), is based on the unique lipid fingerprint of each tissue type, which is not only proved to be different in healthy and cancerous tissue, but also in different tumor types and subtypes. In order to identify different subtypes, grades, both the sensitivity of the algorithm has to be improved and the amount of noise reduced. Noise filtering is done by background subtraction and a well-chosen wave filter using Fourier transformation. Most tumor subtypes are only slightly different, not completely separable using linear methods, thus specific kernel method is applied to wrap the space and reach linear separation. After reducing the dimensions with kernel PCA, different classification algorithms including LDA and support vector machines (SVM) are tested. A dataset was generated from various tissues, and the results were calculated with cross-validation. With these improvements the classification of different tissue types has been increased significantly. With the proper application of these optimized pre-processing algorithms a method for real-time tumor subtype identification can be developed, which could help the on-table decision making of the surgeons.

P52

CHEMOMETRIC APPROACH TO IMPROVE ACCURACY AND PRECISION OF QUANTITATION IN TWO-DIMENSIONAL LIQUID CHROMATOGRAPHY USING DUAL DETECTORS AND MULTIVARIATE CURVE RESOLUTION

Daniel W. Cook and Sarah C. Rutan

Department of Chemistry, Virginia Commonwealth University, 1001 W. Main St Richmond, VA 23284-2006, USA

dwcook@vcu.edu

While the increased peak capacity of two-dimensional liquid chromatography (LC×LC) is appealing for complex samples, such as those resulting from metabolomic samples, the accuracy and precision of quantitation in LC×LC has been limited in comparison to the abilities of one-dimensional LC. One potential cause of this is the undersampling of the first dimension peak due to limitations of the second dimension speed. We introduce a strategy using dual diode-array detectors placed after the first and second dimension columns to take advantage of both the superior quantitation of one-dimensional LC and the superior resolution of LC×LC. To do this, the pure spectra are obtained from the two-dimensional chromatogram and are used to inform multivariate curve resolution on the first dimension chromatograms.

P53

HPLC STATIONARY AND MOBILE PHASE GRADIENT SIMULATIONS

Lena N. Jeong[1], Steven G. Forte[1], Sarah C. Rutan[1], Ray Sajulga[2], and Dwight R. Stoll[2]
[1]*Department of Chemistry, Virginia Commonwealth University, 1001 W. Main St. Richmond, VA 23284-2006, USA*
[2]*Department of Chemistry, Gustavus Adolphus College, 800 W. College Ave. Saint Peter, MN 56082*
jeongl@vcu.edu

Current approaches for method development in liquid chromatography (LC) typically involve consideration of one stationary phase material with variations in mobile phase chemistry. However, variations in stationary phase chemistry can also be utilized to improve the separation of mixtures. The goal of the present work is to design and implement code for simulation of chromatographic separations that is sufficiently flexible to enable the investigation of different chromatographic variables, including: simultaneous changes in mobile phase and stationary phase chemistry, injection volume, and injection solvent. The result of the simulations will provide guidance for the synthesis of novel gradient stationary phases, and a fuller understanding of the effects of method variables on chromatographic performance. The Craig distribution model is used for simulation, where a retention factor is assigned at each distance and time interval within the simulation. Retention factors can be calculated from the parameters S and k_w, the hypothetical free energy of transfer of a particular solute between the organic solvent and water and its retention factor extrapolated to a purely aqueous mobile phase, respectively. These parameters are extracted from experimental data and utilized in the simulation to predict the chromatographic behavior in gradient systems. The simulation program was validated by comparison of retention times of nineteen amphetamines and related compounds to predictions obtained using linear solvent strength (LSS) theory. The simulation code is demonstrated to be useful for simulating chromatograms for variety of different gradient shapes as well as simultaneous stationary and mobile phase gradients.

P54

DETERMINATION OF KINEMATIC VISCOSITY OF CRUDE OILS BY FTIR AND MULTIVARIATE CALIBRATION: EVALUATION OF TRENDS IN RESIDUALS BY PERMUTATION TEST

Karla P. Rainha[1], Paulo R. Filgueiras[2], Luciana A. Terra[2], Ronei J. Poppi[2], Lize M. S. L. de Oliveira[3], Júlio C. M. Dias[3], and Eustáquio V. R. de Castro[1]

[1]Laboratory Research and Development of Methodologies for Analysis of Petroleum (LabPetro), Department of Chemistry, Federal University of Espírito Santo – UFES, Vitória–ES, Brazil.

[2]Institute of Chemistry, University of Campinas, Campinas–SP, P. O. Box 6154, 13083-970, Brazil.

[3]CENPES/PETROBRAS, Rio de Janeiro–RJ, Brazil.

rainhapk@gmail.com

In this study, a nonparametric permutation test was used to evaluate trends in multivariate calibration residuals for the determination of kinematic viscosity of crude oils by FTIR. It was evaluated the residuals from principal component regression (PCR), partial least squares (PLS) and support vector regression (SVR). It was used 68 samples of crude oils with viscosity ranging of 57 to 429 mm^2s^{-1}. The data set was split into 48 samples for model development and 20 samples for validation. The permutation test was performed comparing the quadratic coefficient obtained from the adjusted curve of residuals of the original validation data against sample values, with the quadratic coefficients obtained from the adjusted curve of the permuted residues of validation against sample values. The significance was assessed by relationship between of the quadratic polynomial coefficient of the original validation residuals and the permuted coefficient greater than original coefficients. Calibration models obtained low prediction errors: 25 mm^2s^{-1}, 24 mm^2s^{-1} and 26 mm^2s^{-1} for PCR, PLS and SVR models respectively. However, using the permutation test, a quadratic trend in residuals was observed for PCR and PLS models.

P55

COMBINATION OF CHEMOMETRICS AND ELECTROPHORESIS CAPILLARY-DIODE ARRAY ABSORBANCE DETECTION FOR THE DEVELOPMENT OF SNAKE VENOMS FINGERPRINTS

Sílvia Mas[1], Anna de Juan[1] and Catherine Perrin[2]

[2]*Chemometrics Group. Department of Analytical Chemistry. Universitat de Barcelona. Av. Diagonal, 647. 08028 Barcelona, SPAIN*

[1] *Equipe Sciences Analytiques & Modélisation Moléculaire of Institut des Biomolécules Max Mousseron (IBMM). Unité Mixte de Recherche Universités Montpellier 1 & 2 et CNRS, 15 avenue Charles Flahault, 34093 Montpellier Cedex 5, FRANCE*

silviamas@*ub.edu*

Several drugs containing snake venoms are already commercialized covering a wide range of medical applications such as thrombosis and myocardial infarction. Quality control of these drugs is essential to ensure the efficiency of the treatments.

The main objective of the present study is to develop an analytical fingerprint approach to assess the quality control of snake venoms used in pharmaceutical products. A strategy combining chemometric methods with capillary electrophoresis diode-array absorbance detection (CE-DAD) is proposed.

In order to define a reference class based on fingerprint information for an acceptable snake venom-based product, analysis of a high number of snake venom-lots will be required. The reference class will give information on the compounds consistently present in a similar proportion and on the relative composition profile related to venoms with the required quality standards. First, preprocessing methods will be applied in order to handle the different electrophoretic problems occurred during CE-DAD analysis, such as the presence of baseline/background contributions. Second, multivariate curve resolution-alternating least squares (MCR-ALS) will provide pure component profiles. Pure spectral profiles will be considered for identification of chemical components and pure electrophoretic profiles will be used to obtain relative concentrations of compounds in the different samples. Finally, the compositional profile (fingerprint) obtained by MCR-ALS (based on peak areas) will be used as input information to define the reference class to asess quality venoms. SIMCA will be used to define the characteristics and boundaries of acceptable samples, and the model will be used to perform quality control in new test samples.

P57

APPLICATION OF MCR-ALS FOR THE PREDICTION OF DIFFERENT PETROLEUM PROPERTIES USING SPECTROSCOPIC DATA

V. Kartnaller[1], P.F. de Aguiar[1], E. de Castro[2], J.C.M. Dias[3], and L.M.S.L. de Oliveira[3]

[1]*Instituto de Química, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.*

[2]*Departamento de Quimica, Universidade Federal do Espírito Santo, Espírito Santo, Brasil.*

[3] *Centro de Pesquisas e Desenvolvimento Leopoldo Américo Miguez de Mello (Cenpes), Petrobras, Rio de Janeiro, Brasil.*

This work presents the use of multivariate curve resolution-alternating least squares (MCR-ALS) as a tool for the prediction of different petroleum properties in oil samples, which has never been reported before, as to the best knowledge of the authors.

It was used spectroscopic data of 99 different samples of Brazilian oils with already determined properties, where 30 were then separated as test samples so that these were still representative of the properties' range. The prediction was performed in two steps: (1) estimation of the spectrum that described the related property studied, calculated by the training set and (2) prediction of the property for the test set. In the first step, the training set (D_1) was initially used for the estimation of this illustrative S-matrix for the determined property, using the equation of MCR ($\hat{S}^T = C^+ D_1$). For the second step, the new calculated S-matrix served as an initial estimation for the MCR-ALS algorithm, where now the data matrix contains the spectra of the training set (D_1) and of the test set (D_2). For this step, it was used as constraints of the MCR model the non-negativity of the spectra and concentration and an equality constraint.

For the study, important physical-chemical properties and characteristics of oils were tested, such as °API, relative density (at 20°C/4°C), viscosity and SARA. Results showed that the MCR tool can be used as an analytical method for the quantification of these kinds of properties, showing a good power of prediction, varying for the different properties studied.

P58

USE OF NIR CHEMICAL IMAGES TO ANALYZE THE HOMOGENEITY OF THE POWDER MIXING PROCESS OF GLIBENCLAMIDE TABLETS

Fernanda V. C de Vasconcelos[1], Leandro de M. França[2], Claudete F. Pereira[2], Maria Fernanda Pimentel[3]

[1]Departamento de Química, Universidade Federal da Paraíba, João Pessoa, 58.051-970, Brazil
[2]Departamento de Química Fundamental, Universidade Federal de Pernambuco, Recife, 52.171-030, Brazil
[3]Departamento de Engenharia Química, Universidade Federal de Pernambuco, Recife, 50.740-521, Brazil
fveracruz@gmail.com

The mixing process is a critical step in the manufacturing process of solid pharmaceutical products (capsule and tablets), mainly when there is a low concentration of the active pharmaceutical ingredient. Evaluation of the mixing homogeneity is essential during the validation process to ensure that API content and the uniformity are within the specified standards. Glibenclamide is an oral hypoglycemic used in the treatment of diabetes. A glibenclamide tablet contains only 5 mg of the active pharmaceutical ingredient (API), which is 3 wt% of the total tablet mass, requiring precise quality control. We describe the use of near infrared hyperspectral images (NIR-HI) to evaluate homogeneity during the mixing process in the manufacture of glibeclamide tablets. Samples were collected from the top, middle and bottom of both sampling v-mixer sides at different mixing times (20, 50, 80 and 90 min). NIR hyperspectral images were acquired from the compressed samples. We used MCR-ALS and MIA techniques to assess the concentration distribution. The API optimized spectra by MCR-ALS, after 90 min mixing time, achieved a correlation greater than 0.80, for overall sampled points. The concentration distribution maps, score images and histograms demonstrated that both v-mixer sides were homogeneous during the whole batch mixing process. But the middle of both sampling v-mixer sides were more heterogeneous for the five evaluated compounds (4 excipients and API) at the beginning and homogeneous at the end of the batch mixing process. As described, NIR-HI and chemometrics methods are useful tools for validating the manufacture of glibenclamide tablets.

P59

OPTIMIZATION OF GAP DERIVATIVES FOR MEASURING BLOOD CONCENTRATION ON FABRIC FROM VIBRATIONAL SPECTROSCOPY

Stephanie A. DeJong, Stephen L. Morgan, and Michael L. Myrick
Department of Chemistry and Biochemistry, University of South Carolina, 631 Sumter Street, Columbia, SC 29208, USA
dejongs@email.sc.edu

Derivatives are common preprocessing tools, typically implemented as Savitzky-Golay (SG) smoothing derivatives. Gap derivatives (GDs) are an alternative to SG derivatives, approximating the analytical derivative by calculating finite differences of spectra without curve fitting. GDs offer an advantage of tunability for spectral data as the distance (or gap) over which this finite difference is calculated can be varied. Gap selection is a compromise between signal attenuation, noise amplification, and spectral resolution. This work discusses the implementation and optimization of fourth-order GDs as an alternative to SG derivatives for processing infrared spectra of blood on fabric prior to multivariate calibration. A discussion of the importance of fourth derivative gap selections is presented in the context of multivariate calibration, as well as a comparison to SG preprocessing and lower order gap derivatives.

P60

REGULARIZATION PROCESSES FOR COMBINING ROUGHNESS AND SMOOTHING IN A MULTIVARIATE CALIBRATION MODEL

Alister J. Tencate[1], John H. Kalivas[1], Erik Andries[2]

[1]Department of Chemistry, Idaho State University, 921 S. 8th Avenue, Pocatello, ID 83209, USA
[2]Center for Advanced Research Computing, University of New Mexico, Albuquerque, NM 87106, USA
[2]Department of Mathematics, Central New Mexico Community College, Albuquerque, NM 87106, USA
tencalis@isu.edu

Tikhonov regularization (TR) has been successfully applied to form spectral multivariate calibration models by augmenting the spectroscopic data with a regulation operator matrix. This matrix can be set to the identity matrix I (equivalent to the ridge regression method) yielding rough regression vectors. It can also be set to a derivative operator L forming smoothed regression vectors and thereby generating a smoothed prediction model. Two new regularization methods are proposed that factor both roughness and smoothing into the model. This combination occurs by augmenting calibration spectral data simultaneously with independently weighted I and L. The results of these two new methods are presented and compared to results from only using TR as ridge regression forming rough model vectors and only using a smoothing TR process. Two NIR spectral data sets are used. One consists of corn samples for the analysis of moisture content and the other is a three component system measured at different temperatures.

P61

ANALYTICAL VALIDATION OF A MULTIVARIATE CALIBRATION METHOD FOR DETERMINATION OF SOIL ORGANIC CARBON BY NEAR INFRARED SPECTROSCOPY

André M. de Souza[2], Maurício R. Coelho[2], Ademir Fontana[2], Thayane C. Barbosa Winkler[2], Natasha M. C. Jucá[2], Thais B. de Castro[2], Gabriela B. Shimidt[2], Patrícia Valderrama[1], and Ronei Jesus Poppi[1].

[1]Institute of Chemistry, University of Campinas (UNICAMP), P.O. Box 6154, 13084-971 Campinas, SP, Brazil.
[2]Embrapa Soils, Brazilian Agricultural Research Corporation, Rua Jardim Botânico 1024, CEP 22460-000, Rio de Janeiro (RJ), Brazil.
andremarcelo.souza@embrapa.br

The near infrared diffuse reflectance spectroscopy (NIR - DRS , 700-2500 nm) or NIR - DRS associated with the visible region (Vis/NIR - DRS , 400-780 nm) is considered the most promising analytical technique to replace current routine analysis of soil organic carbon (SOC) due to the numerous advantages of this technique over the traditional ones: it is cheaper, faster, non-destructive and "clean". However, one of the key challenges to enable the implementation of NIR-DRS spectroscopy as a routine method in soil laboratories is to build robust multivariate models from a large number of samples, sufficiently representative of the soil types found in a certain region. In this work, local and global multivariate calibration models were developed and validated for the determination of SOC on a representative set of Brazilian soil samples, as well as a critical view of a possible implementation of the NIR-DRS spectroscopy as a routine method of SOC in a near future in all laboratory of soil analysis in Brazil.

P62

MULTIVARIATE OPTICAL COMPUTING AND ITS APPLICATION TO THE TAXONOMIC CLASSIFICATION OF PHYTOPLANKTON

Shawna K. Tazik, Joseph A. Swanstrom, and Michael L. Myrick
Department of Chemistry and Biochemistry, University of South Carolina, 631 Sumter Street, Columbia, SC 29208
tazik@email.sc.edu

Multivariate optical computing (MOC) is a predictive spectroscopy technique that combines the processes of spectral collection and multivariate analysis into a single measurement. Using MOC, interference filters called multivariate optical elements (MOEs) are fabricated such that their transmission profiles mimic the functionality of a regression vector or linear discriminant vector. This decreases the complexity of the instrumentation needed to measure a chemical or physical property of an unknown sample. This work discusses the principles of MOC and the resulting MOEs as they are implemented in an imaging photometer configuration to taxonomically classify phytoplankton in the ocean.

P63

INVESTIGATION OF CORRELATION BETWEEN FLAVOR SENSORIAL DESCRIPTORS AND CHEMICAL ANALYSIS OF CACHAÇA USING MULTIVARIATE CALIBRATION WITH VARIABLE SELECTION

Gilmare A. da Silva[1], Daniela C. Cardoso[2], Robson J. de C. F. Afonso[1], Sandra R. Gregório[3], and Maurício X. Coutrim[1].

[1]Instituto Federal do Norte de Minas Gerais (IFNMG), Campus Salinas, C.P. 71, 39560-000, Salinas/MG, Brasil.
[2]Departamento de Química - Universidade Federal de Ouro Preto (UFOP), Campus Morro do Cruzeiro, S/N, 35400-000, Ouro Preto/MG, Brasil.
[3]Departamento de Tecnologia de Alimentos - Universidade Federal Rural do Rio de Janeiro (UFRRJ), BR 465, km 7, 23890-000, Seropédica/RJ, Brasil.
gilmare@iceb.ufop.br

Faced with the need to identify the compounds that characterize sensorial and chemically the Cachaça, a typical very well-known Brazilian product, the aim of this work was to investigate the correlation between sensorial evaluation of flavor and chemical composition of Cachaça samples produced in alembics of Salinas/Brazil, a very famous region of production, by means of multivariate calibration. It was collected 24 samples produced by the Association of Handmade Producers of Salinas Cachaça. The flavor sensorial parameters alcohol, acid, sweet, bitter, fruit, sugarcane bagasse, sugarcane, fermented juice, citric, tar, astringent and burning determined by quantitative descriptive analysis were correlated with alcohol content, acidity and 48 compounds of Cachaça using the ordered predictors selection (OPS) and partial least squares (PLS). Organic compounds were determined by high performance liquid chromatography with diode array detection and mass spectrometry (MS), in addition to gas chromatography with flame ionization detection and MS. Copper was determined by atomic absorption spectroscopy. Using OPS it was possible to verify the compounds correlated directly with the sensorial descriptors evaluated, with 7 of them presenting $R^2 > 0.90$ according to PLS models constructed with the selected variables. In the selection procedure, it was necessary 2 or 3 latent variables (RMSECV from 0.048 to 0.332) and was noticed a reduction of up to 80% of the original variables. This inedited study allowed finding correlations between chemical composition and flavor sensorial descriptors, both capable of discriminating Salinas alembic-made Cachaças, becoming a mean for monitoring new samples and evaluate quality process.

Acknowledgements:
Supporting Agencies: CAPES, CNPq and FAPEMIG. Brazilian federal institutions: IFNMG, UFOP and UFRRJ.

P64

APPLICATION OF MULTIVARIATE CALIBRATION WITH VARIABLE SELECTION IN THE ASSESSMENT OF CORRELATION BETWEEN CHEMICAL ANALYSIS AND AROMA SENSORIAL DESCRIPTORS OF CACHAÇA

Gilmare A. da Silva[1], Daniela C. Cardoso[2], Robson J. de C. F. Afonso[1], Sandra R. Gregório[3], and Maurício X. Coutrim[1].

[1]Instituto Federal do Norte de Minas Gerais (IFNMG), Campus Salinas, C.P. 71, 39560-000, Salinas/MG, Brasil.
[2]Departamento de Química - Universidade Federal de Ouro Preto (UFOP), Campus Morro do Cruzeiro, S/N, 35400-000, Ouro Preto/MG, Brasil.
[3]Departamento de Tecnologia de Alimentos - Universidade Federal Rural do Rio de Janeiro (UFRRJ), BR 465, km 7, 23890-000, Seropédica/RJ, Brasil.
gilmare@iceb.ufop.br

It was inedited investigated the correlation between aroma sensorial analysis and chemical composition evaluation of Cachaça sample, a typical very well-known Brazilian product, produced in alembic of Salinas - Brazil, a famous region of production, by means of multivariate calibration. To conduct the study 24 samples produced by the Association of Handmade Producers of Salinas Cachaça were collected and the parameters alcohol content, acidity and 48 compounds of Cachaça were determined; the organic compounds by high performance liquid chromatography with diode array detection and mass spectrometry (MS) besides gas chromatography with flame ionization detection and MS, and copper was determined by atomic absorption spectroscopy. The data were treated using the ordered predictors selection (OPS) and partial least squares (PLS). Using OPS and PLS it was possible to determine the compounds directly correlated with the aroma sensorial descriptors determined by quantitative descriptive analysis: alcohol, sugarcane bagasse, sweet, fruity, acid, citrus, sugarcane, molasses, woody, astringent and sour. Seven of them presenting $R^2 > 0.90$ according to PLS models constructed with the selected variables. In the selection procedure, it was necessary 2 or 4 latent variables (RMSECV from 0.070 to 0.146) and was noticed a reduction of up to 90% of the original variables. With the substances identified in this study, taking into account the great interest in recognize the compounds that characterize sensorial and chemically the Cachaça, certainly it is possible conclusively assess the Salinas alembic-made Cachaças sensorial attributes investigated and be able to discriminate new samples evaluate quality process.

Acknowledgements:
Supporting Agencies: CAPES, CNPq and FAPEMIG. Brazilian federal institutions: IFNMG, UFOP and UFRRJ.

VARIABLE SELECTION BASED ON PLS MODELING OF NIR SPECTROSCOPY FOR GLUCOSE MONITORING

Mohammad Goodarzi, Sandeep Sharma, Herman Ramon and Wouter Saeys
Department of Biosystems, Faculty of Bioscience Engineering, KU Leuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium
mohammad.godarzi@gmail.com

By controlling the blood glucose levels of diabetics permanently, diabetes-related problems such as blindness and loss of limbs can be delayed or even avoided. Therefore, many researchers have aimed at the development of a non-invasive sensor to monitor the blood glucose level continuously. As non-invasive measurement through the skin, the ear lobe or the gums has proven to be either unreliable or impractical, attention has recently turned to minimally invasive sensors which measure the glucose content in serum or interstitial fluid. Thanks to the development of on-chip spectrometers, minimally invasive, implantable devices are coming within reach. However, due to technical limitations, it is not possible to acquire a large number of wavelengths over a broad range. Therefore, the most informative combination of a limited number of variables should be selected. In this study, Genetic Algorithm, Forward iPLS, Backward iPLS and Moving Windows variable selection and combinations of these methods were used in order to address the question whether the first overtone band (1500-1800nm) or the combination band (2050 to 2300 nm) is the most informative for glucose measurements and which wavebands should be measured within these wavelength ranges. The four different data sets employed focus on the determination of (1) glucose in aqueous solutions over the range of 1-30 mM in presence of urea and sodium D-lactate, (2) glucose in aqueous solutions over the range of 2-16 mM, in presence of icodextrin and urea and (3) glucose in a human serum matrix. It was found that the first overtone band is most informative for aqueous solutions, while for glucose measurement of serum samples the combination band was found to be the better choice.

P66

HYPERSPECTRAL IMAGING COUPLED WITH CHEMOMETRICS IN FOOD SENSING

Mohammad Goodarzi, Janos Keresztes, Jeroen Van Roy, Mostafa Khojastehnazhand, Roberto Moscetti, Ainara López, Sergii Morshchavka and Wouter Saeys

Department of Biosystems, Faculty of Bioscience Engineering, KU Leuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium

mohammad.godarzi@gmail.com

Although Hyperspectral Imaging (HSI) was first developed for remote sensing applications, it has become a major tool in many fields such as food science, pharmaceuticals and medical diagnostics, thanks to its applicability by using both imaging and spectroscopy knowledge to further understand phenomena under investigation. A lot of research has been devoted to this field but still the question "which preprocessing technique should be used?" remains. In this study, we have tried to compare different preprocessing techniques in order to address the above question. Hereto we used three different data sets originating from detection of rainbow trout fish ageing, anomaly detection in hazelnuts, and fusarium detection in kernels samples. The HSI images were acquired by two cameras in the range of Vis-NIR and SWIR, 400-1000 nm and 1000-2500 nm, respectively.

P67

IMPLEMENTATION OF STATISTICAL ANALYSIS TOOLS FOR A HANDHELD RAMAN SPECTROMETER

Mark Mabry and Claire Dentinger
Rigaku Raman Technologies, Inc., 14 New England Executive Park, Suite 102 Burlington, MA 01803, USA

Increasing emphasis on lean manufacturing efficiencies and quality based decision making has created a need to bring analytical measurements out of the laboratory. This has resulted in tremendous growth in the area of portable analytical instrumentation, made possible by recent developments in opto-electronic technology. For example, hand-held Raman spectrometers are being implemented to confirm raw material identity in pharmaceutical manufacturing. Material identity can be quickly confirmed by spectral comparison against a library spectrum of a known reference materials. There are a number of data processing strategies which can be used to perform this library spectral comparison. In addition chemometric analysis can enable measurements of a semi-quantitative nature and discrimination between closely related materials. To date chemometric tools require development of the model at a separate workstation which must then be transferred back to the portable instrument for use.

A portable, long wavelength Raman spectrometer with on-board chemometric model development will be presented. An example of discriminating closely related pharmaceutical incoming raw materials using this on-board model development will be demonstrated. Advances in the area of library search methods will also be discussed.

PARALLEL ISOTOPIC TAG SCREENING: A CHEMOMETRICS APPROACH TO QUANTITATIVE PROTEOMICS

Peter D. Wentzell, Bjorn L.M. Wielens, Gemma Regan, Umesh Regmi
Department of Chemistry, Dalhousie University, PO Box 15000
Halifax, NS B3H 4R2, Canada
peter.wentzell@dal.ca

In conventional comparative proteomics using a "bottom-up" approach, tryptic digests of the proteins in two or more samples to be compared result in peptides that are then chemically labeled according to one of several protocols. These tags allow peptides from different sources to be distinguished from one another in the mass spectrum, and relative quantitation is achieved from the intensity ratios of peak clusters. Tandem mass spectrometry (MS/MS) is then used in conjunction with database searching to identify the parent protein. In the work presented here, an approach to quantitative proteomics using only low resolution MS^1 data is described. Isotopic patterns are calculated and used to generate several models for peak clusters based on single, double or triple charges with varying numbers of labels. Through liquid chromatography-mass spectrometry (LC-MS), peptides are located in the mass chromatogram by applying an algorithm that implements these models in parallel, referred to as parallel isotopic tag screening (PITS). The PITS algorithm incorporates stages of peak detection, clustering and quantitation. The labeling strategy employed uses an inexpensive technique for dimethylation of primary amines. A triple-labeling method is used in which a common reference is labeled with light and heavy tags to allow peptide detection, and a test sample is labeled with an intermediate tag to allow quantitation. This approach permits operation in MS-only mode, improving the instrument duty cycle, and permits the generation of matrices of data more conducive to chemometric methods such as multivariate curve resolution in longitudinal studies.

P70

UNTARGETED METABOLOMIC APPROACH TO THE STUDY OF EMT IN PROSTATE CANCER CELLS

Carme Bedia, Núria Dalmau, Romà Tauler, and Joaquim Jaumot
Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E-08034, Barcelona, Spain
joaquim.jaumot@idaea.csic.es

Epithelial-mesenchymal transition (EMT) is an essential process in cell biology involved in embryonic development, tissue remodeling and wound healing. In this process, epithelial cells reduce intercellular adhesion and increase invasive and migratory properties. Recently, EMT has been shown to play a crucial role in tumor invasion and metastasis. So, the aim of this work is to evaluate changes in the metabolome caused by different endocrine disruptors in acute and chronic treatments in order to detect biomarkers that allow characterizing the EMT transition. For that reason, an extract from the treated/control cells has been obtained and measured by means of high-resolution LC-MS.

From the measured chromatograms different approaches have been used to characterize the EMT transition. On one hand, a preliminary analysis has been attempted by considering the total ion chromatograms (TIC) and analyzing them by Principal Component Analysis in order to evaluate the differences between the control and the treated samples. On the other hand, Multivariate Curve Resolution (MCR) analysis of the full LC-MS has allowed the detection of potential biomarkers related to the EMT process induced by the different treatments. Finally, from the MCR list of potential biomarkers a reduced set of candidates has been obtained by means of considering the most influential variables in a PLS-DA model by linking the treatment and the biomarker concentration in the different samples.

Acknowledgements
The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 32073

P71

CHEMOMETRIC TOOLS APPLIED TO THE EXPLORATION OF MASS SPECTROMETRY IMAGING DATA

Joaquim Jaumot and Romà Tauler
Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E-08034, Barcelona, Spain
joaquim.jaumot@idaea.csic.es

In recent years a lot of attention has been focused on the development and application of chemometric tools to the study of hyperspectral images due to its ability of carrying out fast and relatively cheap analyses. In these cases, a complete spectrum (usually vibrational techniques such as NIR or Raman spectroscopies are used) is collected for each pixel location of the sample. Recently a new promising technology has been developed as it is now possible to collect the mass spectrum for each pixel of a considered sample obtaining also both chemical information and spatial distribution of each analyte detected. However, there are major drawbacks inherent to the mass spectrometry imaging (MSI) related to the high complexity of real samples and the extremely huge amount of data generated.

These problems require the use of chemometric tools in different steps of the analysis process, from data compression to the resolution of the different components present at each pixel. In this work, examples of preliminary analysis of biomedical MSI datasets (i.e. images of mouse lung or bran) will be presented, showing the potential of using the chemometric tools.

Acknowledgements
The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 32073

P72

COMPARATIVE STUDY OF VARIABLE SELECTION METHODS BASED ON VARIABLES IMPORTANCE IN THE PROJECTION AND SELECTIVITY RATIO FOR DIFFERENT KIND OF DATASETS

Mireia Farrés[1], Stefan Y. Platikanov[1], Stefan L. Tsakovski[2], and Romà Tauler[1]

[1]Environmental Chemistry,Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), c/ Jordi Girona 18, 08034 Barcelona, Spain

[2]Department of Analytical Chemistry, Faculty of Chemistry, Sofia University, James Bourchier Blvd, 1164 Sofia, Bulgaria

mireia.farres@idaea.csic.es

Several approaches for variable selection have been proposed in Partial Least Squares (PLS) literature [1], such as Variable Importance in the Projection (VIPs) and Selectivity Ratio. However, when these approaches are applied to the same data sets, distinct results can be obtained.

In this work, VIPs and SR variable selection methods are compared for the analysis of three different case studies. Selection of the best approach is performed according to the optimal description of the underlying experimental phenomena and the best interpretation of the results. The three datasets included in this work are: 1) sensorial data related with physicochemical water quality parameters[2]; 2) gas chromatography/mass spectrometry chemical profiles of fossil organic compounds related with sea surface temperature changes[3]; 3) transcriptome data from *Daphnia magna* related to total offspring production under two types of stress[4].

References:

[1] C.M. Andersen, R. Bro, Journal of Chemometrics, 24 (2010) 728-737.
[2] S. Platikanov, V. Garcia, I. Fonseca, E. Rullán, R. Devesa, R. Tauler, Water Research, 47 (2013) 693-704.
[3] M. Farrés, B. Martrat, J.O. Grimalt, R. Tauler, (2014).
[4] B. Campos, N. Garcia-Reyero, C. Rivetti, L. Escalon, T. Habib, R. Tauler, S. Tsakovski, B. Piña, C. Barata, Environmental Science & Technology, 47 (2013) 9434-9443.

Acknowledgement: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 32073

P73

STUDY OF HUMAN PROSTATE CANCER CELL LINES UNDER POLLUTANT STRESS BY RAMAN HYPERSPECTRAL IMAGING AND CHEMOMETRIC TECHNIQUES

Víctor Olmos[1], Carmen Bedia[2], Romà Tauler[2], and Anna de Juan[1]

[a] *Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Diagonal 645, Barcelona, Spain*

[b] *Environmental Chemometrics Group, Department of Environmental Chemistry, Institute of Environmental Assessment and Water Diagnostic (IDAEA-CSIC), Barcelona, Spain*

Chemical pollution can produce changes in the genome, proteome and metabonome of living organisms. The study of these effects provides information about the possible biological and environmental response to these contaminants. The aim of this work is to study the effect of some chemical pollutants on human prostate cancer cells (DU145) using Raman hyperspectral imaging and chemometric techniques.

Prostate cancer cell line DU145 will be exposed acutely or chronically to different doses of selected pollutants, some of them epidemiologically or biologically related to prostate cancer induction and progression, such as aldrin or chlorpyrifos. To identify spectral biomarkers associated with different pollutant scenarios, a previous exploratory analysis using conventional Raman spectroscopy will be performed on individual spectra coming from control and treated cell populations. Once the most relevant spectral features to distinguish control from contaminated cells are identified, Raman hyperspectral images from DU145 individual cells on the spectral ranges of interest will be recorded.

Multivariate curve resolution-alternating least squares (MCR-ALS) will be used to analyze hyperspectral images. Results from the analysis of multiset data including control and treated cells hyperspectral images and time course cell dynamics images will be presented. MCR resolved pure spectral signatures and distribution maps of the zones inside the cell with different behavior will be given. The advantage of using hyperspectral imaging is that they provide spatial and spectral information from cell changes. In this way, identification of which spectral variations and cell regions are the most affected by selected pollutants are obtained.

P74

THE INVESTIGATION OF THE DISSAPPEARANCE DYNAMICS OF THE REMAINDER OF PESTICIDES IN FRUIT AND LEAVES OF THE APPLE TREE

Barbara Debska and Agnieszka Rudy
Department of Computer Chemistry, Rzeszow University of Technology,12 Powstancow Warszawy Ave., 35-959 Rzeszow, Poland.
bjdebska@prz,edu.pl

The group that is most exposed to diseases caused by contaminated food or water includes infants and children below the age of ten. The possible health hazards due to a contact with inadmissible preparations are, among other things, immune system impairment, endocrine system disturbances, nervous system disorders and tumours. An EU directive (2003/13/EC) determines the highest allowable concentration level for the remains of plant protection chemical substances in foods for infants and children as lower than 0.01[mg/kg]. Fruits and vegetables are heavily contaminated by pesticides. The aim of the investigation was to fix the latest possible deadline for spraying plant protection substances so that apple crops could be used for the production of food for babies and small children.

The subject of the investigation was the dynamics of the disappearance of myclobutanil, kresoxim methyl and flusinasole, which are active substances in the preparations used to protect apple trees. A chromatographic analysis was conducted to determine the remains of the pesticides in the specimens of fruit and leaves. The level of the remaining pesticides was calculated using the StatSoft STATISTICA system.

The levels of the remains of those substances on the second day after the treatment were quite big. The reduction of the quantity of used substances to a half was obtained for: myclobutanil in fruit after 17 days, in leaves after 35 days; kresoxim methylin in fruit after 8 days, in leaves after 10 days; flusilazole in fruit after 17 days, in leaves after 14 days. The time after which the quantity of the detrimental substances in the fruit was estimated below 0,01 [mg/kg], was determined as follows: myclobutanil - 30 days, kresoxim methyl - 32 days, flusilazole - 43 days. The experimentally determined decay course of the pesticides suggests that, in the case of apples, the plant protection chemical substances should be used six weeks before harvest time at the latest.

P75

A NEW METHOD TO EVALUATE THE ROBUSTNESS OF PCA

Y.J. Liu[1,2], T. Tran[1], J. Jansen[1], G. Postma[1], H.L. Wu[2], L.M.C. Buydens[1]

[1]Radboud University Nijmegen, Institute for Molecules and Materials (IMM) Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands, [2]State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China.

yajuanliu1988@gmail.com

PCA is widely used in analytical chemistry not only for data reduction but also for interpretation purposes of the structure of the majority of the data. The robustness of the ordinary PCA is however seriously dependent on the actual data such as the data size and the availability of disturbance patterns, e.g., outliers or artifacts.

A lot of robust methods came up. Associated with them is another problem: how to evaluate the robustness of the resulting models as the robust model can give a proper fit for data with a possible presence of unhealthy patterns, for instance, due to outliers or incomplete samples.

The existing robust evaluation methods (e.g., influence function and breakdown point), focus on evaluation of a method's robustness in general as a function of outliers. In this study, robustness is evaluated in a different way. The model is viewed robust on a particular data set provided that we can get the similar principal component even with an incomplete data or small sample size in a complex mixture. The robustness is quantified by the Principal Component uncertainty assisted with a bootstrapping resampling statistical technique to reveal the underlying data structure for evaluating the uncertainty (robustness) of model parameters.

P76

STRATEGIES COMBINING RESOLUTION AND SEGMENTATION IN HYPERSPECTRAL IMAGE ANALYSIS OF BIOLOGICAL TISSUES

S. Piqueras[1,2], C. Krafft[3], C. Beleites[3], K. Egodage[3], F. von Eggeling[4], O. Guntinas Lichius[5], J. Popp[4,5], R.Tauler[2], and A. de Juan[1].

[1]Grup de Quimiometria. Dept. Química Analítica. Universitat de Barcelona. Barcelona, [2]Environmental Chemometrics Group. Department of Environmental Chemistry, Institute of Environmental Assessment and Water Diagnostic (IDEA-CSIC), Barcelona, [3]Leibniz Institute of Photonic Technology, Jena, Germany, [4]Institute of Physical Chemistry and Abbe Center of Photonics, University Jena, Germany, [5]Department of Otorhinolaryngology, University Hospital Jena, Germany.

piqueras.sara@gmail.com

Image segmentation is oriented to identify groups of similar pixels in an image; that is, pixels with similar spectra and, therefore, similar composition and chemical or biological properties are identified. Resolution (unmixing) methods, instead, focus on recovering concentration profiles (folded back into distribution maps) and pure spectra of image constituents under the assumption of a simple spectroscopic bilinear model. Both kinds of methods provide complementary information and can be differently combined depending on the purpose of the study. Using Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) as a resolution algorithm and K-means as segmentation method, two combinations of these methods are tested for the analysis of biological tissue images.

First, resolution followed by segmentation will be applied. In this way, basic spectral signatures and distribution maps of the most salient biological elements can be recovered. Later, MCR scores (concentration profiles) are used as compressed initial information for segmentation purposes. The results of segmentation analysis from MCR scores are chemically meaningful and can be easily associated with different tissue elements in biomedical images.

However, to see a more detailed differentiation of the different layers within the tissues, the opposite combination strategy can be applied. For this purpose, first the image is segmented into the main tissue zones and then MCR-ALS is applied to the zones of interest for diagnosis. Local MCR-ALS models will be required to identify finer spectral details that can distinguish better between healthy and affected tissue.

Both strategies combining resolution and segmentation analysis will be shown for the analysis of inflamed and non-inflamed tonsils by FT-IR micro imaging of whole sections and Raman-microscopic mapping.

P77

DRUG PHOTOSTABILITY STUDIES USING MCR-ALS SIMULTANEOUS ANALYSIS OF FUSED INCOMPLETE MULTISET DATA OBTAINED BY UV-VIS SPECTROPHOTOMETRIC TITRATION AND KINETIC STUDIES COUPLED TO HYPHENATED DAD/MS CHROMATOGRAPHIC ANALYSIS

Michele De Luca[1] , Giuseppina Ioele[1], Gaetano Ragno[1] and Roma Tauler[2]

[1]Department of Pharmacy, Health and Nutritional Sciences, University of Calabria, Rende (CS), Italy.

[2] Institute of Environmental Assessment and Water Diagnostic (IDEA-CSIC), Barcelona. Roma.Tauler@idaea.csic.es

An advanced and powerful chemometric approach is proposed for the simultaneous processing of fused incomplete data multiset. The procedure was applied to the comprehensive description of the amiloride (AML) drug photodegradation. The data multiset consisted of coupling the UV-Vis spectral measurements from kinetic and acid–base experiments with their hyphenated DAD/MS chromatographic data. These data sets were combined (fused) in a very large and incomplete column- and row-wise super augmented data matrix which was then processed by a new version of the multivariate curve resolution – alternating least squares algorithm (MCR-ALS)[1]. This version was developed specially for this purpose and proved to be able to perform complex mixed multilinear modeling. The method allowed a very powerful in-depth analysis of the AML system and the description of the kinetic pH dependent photodegradation pathway of the drug[2]. The proposed method can be easily extendable to other drugs and chemical compounds to define their simultaneous acid-base and kinetic reactions.

References

[1] M. Alier and R. Tauler. Chemom. Intell. Lab. Sys. 127 (2013) 17-28.

[2] M. De Luca, G. Ioele, S. Mas, R. Tauler, G. Ragno. Analyst, 2012, 137, 5428–5435.

P79

A CHEMOMETRIC APPROACH TO THE OPTIMIZATION OF BIO-INDUSTRIAL PROCESSES

Anna Klimkiewicz[1,2], Marianne B. Rousing[1] and Frans W.J. van den Berg[2]
[1]Novozymes A/S, Kalundborg, Denmark, [2]University of Copenhagen, Frederiksberg, Denmark
akcz@novozymes.com

In modern production, a massive number of diverse measurements, with a wide diversity in information content and quality, are stored in data historians. A tremendous amount of process signals is already collected throughout the different biomanufacturing steps - typically generated for specific and dedicated univariate monitoring and closed-loop control applications. This produces large amounts of data which are seldom used outside their direct scope. It this work we highlight the value of historical production data as a starting point of variance reduction and process optimization in biomanufacturing.

The necessary steps involved in *recycling* of the historical data from a full-scale downstream processing of industrial enzymes are briefly described. Depending on the extraction scheme and the final objective of the models different signal processing methods can be introduced to compensate for e.g. sensors with low signal to noise ratio (i.e. by smoothing) or drawbacks related to saving frequency. Another important step is alignment and synchronization of process data. This is particularly significant when looking at the relation between sequences of unit operations separated in time and even more so when working with continuous processes, generating the time-series data. For this application, the potential of auto- and cross-correlation analyses are explored. Finally, the identified, soft (*data-driven*) models that capture the principle behavior of the system can potentially be used for optimized control in future production runs.

P80

MONITORING ENZYME ACTIVITY DURING ULTRAFILTRATION PROCESS USING ON-LINE AND IN-LINE NIR MEASUREMENTS

Anna Klimkiewicz[1,2], Christian Bomholt Zachariassen[1], Peter Paasch Mortensen[1] and Frans W.J. van den Berg[2]

[1]Novozymes A/S, Kalundborg, Denmark, [2]University of Copenhagen, Frederiksberg, Denmark
akcz@novozymes.com

The key parameter during purification of enzymes is the strength of the intermediates and products. Pre-studies have proven a good relation between near infrared (NIR) spectra and enzyme activity in the retentate. We have compared the performance of two real-time setups (on-line v. in-line) of NIR transmission probes in the industrial scale purification plant. Four different types of industrial enzymes have been sampled over a period of ten months. Real-time output of the partial least squares regression models was used together with conventional process monitoring signals to evaluate the benefits of the in-line and on-line setup. Both arrangements deliver good results in monitoring the ultrafiltration process but problems (phase transition, fouling) have been encountered for the on-line setup.

P81

HIGH PERFORMANCE PLATFORM FOR MINING LARGE-SCALE MASS SPECTRAL DATASET

Zhi-Min Zhang, Xin-Bo Liu, Yi-Zeng Liang and Hong-Mei Lu

Institute of Chemometrics and Intelligent Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China

zhangzhimin.csu@gmail.com

High-resolution mass spectra (HRMS) play an important role in structure elucidation. However, the mining of discriminant markers from large-scale GC-MS or LC-MS dataset and the identification of them via HRMS are still difficult for most researchers. Presently the markers are often discovered by manual preprocessing and pattern recognition, and then identified by searching MS libraries. This procedure is time-consuming and subjective, and the spectra in MS libraries are limited. So some novel methods for preprocessing, pattern recognition and identification are needed urgently.

In this work, several techniques in computer science have been adopted to process the large-scale HRMS dataset effectively in acceptable time, including 64bit computing, GPGPU computing, Python+Cython and Julia programming language. We will implement baseline-fitting, peak detection, automatic deconvolution and alignment methods to construct 2D matrix for pattern recognition and corresponding HRMS for structure elucidation based on the above high performance computing techniques. Then random forests and sparse linear discriminant analysis have employed to discover the influential markers effectively. For the markers not including in the MS libraries, accurate m/z values, mass spectra calibration, isotopic abundance, PubChem database, and retention index and in silico fragmentation have been adopted for molecular formula and structure identification.

This study can provide a novel and systemic platform for analyzing and mining HRMS dataset of complex system effectively, which is meaningful to several research areas such as metabolomics, food safety, active compounds of herbal medicine and etc.

P82

CHEMOMETRICS IN COMPENDIAL APPLICATIONS

Lucy L. Botros, Jeffrey C. Moore, and Alan R. Potts
United States Pharmacopeial Convention, 12601 Twinbrook Parkway, Rockville, MD 20852-1790 USA
lyb@usp.org

The U.S. Pharmacopeial Convention (USP) is a scientific nonprofit organization that sets standards for the identity, quality, and purity of medicines, food ingredients, and dietary supplements by the application of the tests, procedures, and acceptance criteria set forth in relevant compendia. USP has recently applied chemometrics in the development of such test procedures using different analytical technologies (e.g. spectroscopy and chromatography), for different compendial purposes, including identification, adulteration detection, and properties prediction. Examples of such chemometric procedures will be presented, including non-targeted adulteration detection in food ingredients, and the identification of naturally sourced excipients. In addition, the need for a guidance document detailing scientifically sound practices for the chemometric interpretation and analysis of typical multivariate data from compendial and industrial applications will also be presented.

P83

3-WAY NETWORKS: NEW TOOLS FOR COMPARITIVE GENOMICS

Debbie Weighill and Dan Jacobson
Institute for Wine Biotechnology, Stellenbosch University, Stellenbosch, South Africa 700
jacobson@sun.ac.za

Evolution can occur on the scale of gene families through gene duplication and deletion events. Duplicate genes are an important driver of evolution as one member of the pair of duplicated genes can perform its original function while the other is free to diverge, possibly providing an evolutionary advantage to the organism. Thus, organisms build up families of related genes that collectively represent their proteomic/functional potential. Using TribeMCL, gene families were determined for 211 bacterial genomes and a gene family profile constructed for each species. These gene family profiles were compared to one another with similarity metrics resulting in pairwise comparisons of bacterial genomes and networks were constructed from these pairwise relationships. Networks are essentially a reductionist approach and, as such, are a useful tool for modelling such complex systems. However, modelling a system as purely pairwise relationships may miss more sophisticated patterns present in the data. To address this we have used an existing and developed a new 3-way similarity metric and created networks representing the resulting ternary relationships between species. These 3-way networks were applied to the phylogenomic profiles of the 211 bacterial genomes. Comparisons between 2-way and 3-way networks confirmed that 3-way networks revealed important patterns in the dataset that would have been missed otherwise. The most insightful result was achieved with a combination of 2-way and 3-way networks. The use of these more complex models has led to the discovery of patterns in bacterial genomes that would not have been discovered with simpler phylogenetic methods.

P84

DEVELOPMENT OF A FAST METHOD TO PREDICT LIGNIN CONTENT IN SUGARCANE USING NIR SPECTROSCOPY AND MULTIVARIATE CALIBRATION

Camila Assis[1], Lidiane A. Silva[2], Karla Gasparini[2], Rachel. S. Ramos[2], Thálisson S. Souza[1], Jussara V. Roque[1], Cecília. B. Rosado[1], Volmir Kist[2], Márcio H. P. Barbosa[2], and Reinaldo. F. Teófilo[1]

[1]Department of Chemistry, Universidade Federal deViçosa 36570-900 – Viçosa – MG-Brazil
[2]Department of Plant Science, Universidade Federal deViçosa 36570-900 – Viçosa – MG-Brasil
rteofilo@gmail.com

Multivariate calibration models were built for the determination of lignin in sugarcane using near infrared spectroscopy (NIR) and partial least squares (PLS). To perform the study, it was determined the content of lignin in 221 genotypes, using the Klason method. The NIR spectra were obtained from sugarcane stalk in two different areas: middle and top, without any sample treatment. Forty samples were selected for the prediction set. Furthermore different algorithms were compared for variable selection, such as: genetic algorithm (GA), interval PLS (*i*PLS) and ordered predictor selection (OPS).

RMSECV and RPD values obtained for the models (top) using the algorithms were, respectively: GA (0.83; 2.55), *i*PLS (2.10; 1.47) and OPS (0.89; 2.34). For the middle, the values were: GA (0.87; 2.17), *i*PLS (1.76; 1.38) and OPS (0.63; 3.24). The OPS algorithm selected a lower number of variables with a higher predictive capacity compared to the others. Moreover, the OPS algorithm performed the calculations in considerably less time compared to the GA. The PLS-OPS model showed high accuracy to predict the lignin content in sugarcane, reducing significantly the time spent with the analysis, the cost and the chemical reagents consumption.

Acknowledgements:
Supporting Agencies: CAPES, FAPEMIG, FUNARBE and CNPq.

P85

COMPARISON BETWEEN UV-MCR-ALS AND HPLC-DAD TO MONITORING THE ELECTRODEGRADATION OF ATRAZINE ON BORON-DOPED DIAMOND ANODE

Thálisson S. Souza[1], Camila Assis[1], Jussara V. Roque[1], Gilmare A. da Silva[2], Efraim Reis[1], Reinaldo F. Teófilo[1]

Departamento de Química, Universidade Federal de Viçosa, Viçosa/MG, 36570-900, Brazil.
Departamento de Química, Universidade Federal de Ouro Preto, Ouro Preto/MG - 35400-000, Brazil.
rteofilo@gmail.com

Pesticides such atrazine can be found in natural waters and their complete degradation becomes necessary; their presence in natural environmental is concerning by the possibility of implication of many compartments, including human health. Selective electrodegradation monitoring of such substance can be performed using chromatographic techniques, but they are expensive and often slow. Otherwise chemometric methods can be used to resolve and identify the species present in mixtures which have overlapping analytical signals such as UV/Vis spectra. In this way, the aim of this study was to perform atrazine electrodegradation on boron-doped diamond anode (BDD), monitoring the reaction using UV spectra continuously, solving the system using multivariate curve resolution–alternating least squares (MCR-ALS) and comparing the results with high performance liquid chromatography–diode array detector (HPLC-DAD), from aliquots (0, 100, 150, 200, 250 and 300 min). The MCR-ALS was performed using the augmented matrix technique. Five sources were found in both methods and three of them agreed by both methods. It was noticed the full degradation of atrazine and the formation of two byproducts which were partially degraded and remains after the end of degradation. The use of MCR-ALS showed to be simple, fast, fairly selective and efficient for monitoring this system.

Acknowledgements:
Supporting Agencies: CAPES, FAPEMIG, FUNARBE, RQ-MG and CNPq.

P87

COMPARISON BETWEEN FTIR-ATR AND NIR SPECTROSCOPIES FOR CAFFEINE DETERMINATION IN ENERGY DRINKS

Jussara V. Roque, Camila Assis, Thálisson S. Souza, Guilherme R. Pereira, Reinaldo F. Teófilo.
Departamento de Química. Universidade Federal de Viçosa. MG. 36570-900. Brazil.
rteofilo@gmail.com

The feasibility of Fourier transform mid infrared (FTIR) and near infrared (NIR) spectroscopies for the determination of caffeine in energy drinks was studied in order to present a fast and inexpensive quality control method. High performance liquid chromatography (HPLC) was used as reference method. Eighty energy drinks samples of different brands were used to build different multivariate calibration models using partial least squares (PLS) regression. Two methods of variable selection were applied, i.e., ordered predictors selection (OPS) and genetic algorithm (GA). The spectra were obtained directly from liquid without any sample treatment. Both data sets were mean centered, and a first derived was applied. Four multivariate calibration models were obtained, and the comparison between them was performed by analyzing statistical parameters of quality such as root mean square error prediction (RMSEP) and ratio of performance to deviation (RPD). The parameters obtained for FTIR and NIR with OPS were respectively: RMSEP 16.45; 12.46 and RPD 2.88; 2.75. For GA were obtained, respectively: RMSEP 18.88; 21.61 and RPD 2.96; 1.30. These results indicated that both spectroscopy techniques are reliable for predictions after PLS with variable selection, except for the NIR-PLS-GA model, which presented only for screening. Therefore the caffeine predictions in energy drinks showed possible without destroying the sample and in a few minutes. The best model found was the one obtained with NIR and OPS algorithm.

Acknowledgements:
Supporting Agencies: CAPES, FAPEMIG, FUNARBE and CNPq.