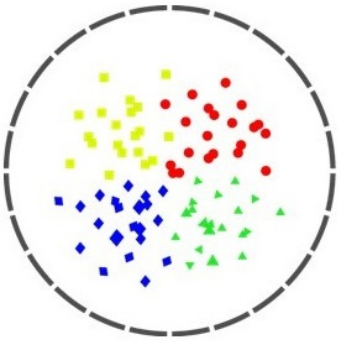
Online monitoring of H2S scavenging reactions in aqueous phase using Raman spectroscopy

Romero Logrono, Iveth Alexandra; Kucheryavskiy, Sergey; Maschietti, Marco

Publication date:
2021

The Danish Chemometric Society welcomes you to



SSC17

September 6 to 9, 2021
Comwell Hvide Hus, Aalborg, Denmark

info@ssc17.org

jmp Statistical Discovery.™ From SAS.

FOSS

interline

SARTORIUS

CHR. HANSEN
Improving food & health

RAMCON
PRODUCTS PEOPLE SOLUTIONS

Table of Content

SSC17 Program 4

O2: Optimization of UPLC-MS/MS assay for clinical diagnosis and therapeutic drug monitoring in patients with APRT deficiency by design of experiments 7

O3: Intervention studies on gut microbiota: Can ASCA compete with methods that are specifically tailored for microbial data? 10

O4: ANOVA-PCA and PLS-DA for volatile metabolites chromatographic profile analysis as an alternative method for early, non-destructive and non-invasive detection of fungi species in Carica papaya (in vivo) 12

O5: Two-step authentication of overlapping classes 14

O6: Interplay of decision rules and parameter optimization strategies in SIMCA 16

O7: Inter and intra class discrimination based on multivariate analyses applied on bacterial SERS fingerprints 18

O8: Two distinct frameworks to adapt source spectral calibrations to unlabeled target samples: (1) local modelling by linking linear classification to regression and (2) transfer learning 19

O10: Unified framework for calibration transfer 21

O11: Towards calibration transfer with arbitrary standards 23

O12: Chemometrics extended to a parallel world of nondestructive Natural Chemical Computing 25

O13: Comparing calibration transfer approaches 27

O14: Rank Expansion (REX): a mathematical tunnel effect? 29

O15: Contextual Mixture of Partial Least Squares Experts: Integrating process specific characteristics into model structure 31

O16: Alternative approaches to untargeted LC/GC-MS data analysis 32

O18: Spatial-spectral analysis of NIR imaging data - A case study 33

O19: Pixels that matter in chemical imaging 35

O21: Towards a machine learning based produced for interpretation of mass spectra for better understanding of hydrate phenomena in oil systems 37

O22: On the possible benefits of deep learning for spectral pre-processing 38

O23: Validation of classification models in cancer studies using simulated spectral data 39

O24: Towards successful silver anniversary with Advanced Process Control 41

O25: Multiblock supervised analyses, should we really normalize blocks? 42

O26: N-CovSel, a new strategy for feature selection in N-way data 44

O27: Bitterness in beer – investigated by fluorescence spectroscopy and chemometrics 46

O28: Classification of cellulose textile fibers 48

O29: Water quality control based on the analysis of high-resolution phytoplankton data 50

O30: Time Domain Reflectometry (TDR) and classification algorithms to detect injection of different water solutions in fresh tuna 52

O32: Online monitoring of H_2S scavenging reactions in aqueous phase using Raman spectroscopy 54

O33: Chemical quality prediction by inversing dynamic PLSMAR: balancing interpretability and accuracy 56

O34: Process monitoring of a pesto production process through RGB Imaging and Near Infrared Spectroscopy 57

O35: Improved understanding of industrial process relationships through conditional path modelling with Process PLS 58

Posters SSC17 60

P01: Application of Raman spectroscopy for monitoring of hydrogen sulfide scavenging reactions using biomass-based chemicals 61

P02: PARAFAC handles inner filter effects and FRET in fluorescence spectra 62

P03: Class-modeling: Reviving old tools unjustly forgotten 64

P04: Targeted proteomics and multivariate data analysis for search of novel biomarkers for early breast cancer diagnosis 65

P05: Characterization of different waste wood for assessing the best reuse on the basis of their quality attributes 67

P06: Explorative and causal path modeling - limitations and synergies 69

P07: Novel spectrophotometric method to determine simultaneously hypophosphite and phosphite in electroless baths 71

P08: Comparing multivariate ANOVA methods in Multicolor Flow Cytometry 73

P09: When and how do artificial neural networks learn domain knowledge for near infrared food application 75

P10: pH measurement and phosphate determination in pharmaceutical eye drops for eye diseases by digital image analysis 77

P11: The stability of oat drinks assessed using low field NMR T2 relaxation 79

P12: Finding new chemometric tools for SERS spectra cluster analysis and predictive modelling 80

P13: Modelling of scattering signal for direct PARAFAC decompositions of excitation-emission matrices 81

P14: Stochastic optimisation as a straightforward strategy for laser-induced calibration-free breakdown spectroscopy 83

P15: Interpolation of scattering signal before PARAFAC processing of EEM-fluorescence spectra 85

SSC17 Program

Monday 06.09.2021

0845	Henrik Toft	Introduction to Chemometrics in Python
1130		LUNCH and Registration
1230	Per Waaben Hansen	Welcome to SSC17
	ASCA/ DoE	*Chairman: Lars Houmøller*
1240	Ingrid Måge	Experimental design: the ultimate ChemomeTrick (O1)
1315	Margrét Thorsteinsdóttir	Optimization of UPLC-MS/MS assay for clinical diagnosis and therapeutic drug monitoring in patients with APRT deficiency by design of experiments (O2)
1340	Ingunn Berget	Intervention studies on gut microbiota: Can ASCA compete with methods that are specifically tailored for microbial data? (O3)
1405	Larissa R. Terra	ANOVA-PCA and PLS-DA for volatile metabolites chromatographic profile analysis as an alternative method for early non-destructive and non-invasive detection of fungi species in Carica papaya (in vivo) (O4)
1430		COFFEE
	Classification	*Chairman: Lars Houmøller*
1500	Zuzanna Małyjurek	Two-step authentication of overlapping classes (O5)
1525	Marina Cocchi	Interplay of decision rules and parameter optimization strategies in SIMCA (O6)
1550	Ana Maria Raluca Gherman	Inter and intra class discrimination based on multivariate analyses applied on bacterial SERS fingerprints (O7)
1615	John H. Kalivas	Two distinct frameworks to adapt source spectral calibrations to unlabeled target samples: (1) Local modeling by linking linear classification to regression and (2) transfer learning (O8)
1635		End of scientific presentations
1800		DINNER

Tuesday 07.09.2021

0700	John Holm & Pia Jørgensen	Morning run/ walk
	Theoretical	*Chairman: Åsmund Rinnan*
0845	Puneet Mishra	The hype in deep learning of spectral data, and when it really is useful (O9)
0920	Valeria Fonseca Diaz	Unified framework for calibration transfer (O10)
0945		COFFEE
	Theoretical	*Chairman: Åsmund Rinnan*
1015	Nikzad-Langerodi Ramin	Towards calibration transfer with arbitrary standards (O11)
1040	Lars Munck	Chemometrics extended to a parallel world of nondestructive Natural Chemical Computing (O12)
1105	Lars Erik Solberg	Comparing calibration transfer approaches (O13)
1130		LUNCH
1230		Walk'n'Talk and Online Networking
	Theoretical	*Chairman: Georg Rønsch*
1330	Carsten Ridder	Rank expansion (REX): A mathematical tunnel effect? (O14)
1355	Francisco Souza	Contextual Mixture of Partial Least Squares Experts: Integrating process specific characteristics into model structure (O15)
1420	Andrea Jr Carnoli	Alternative approaches to untargeted LC/GC-MS data analysis (O16)
1445		COFFEE
1515	Johan Trygg	Herman Wold medal
	Imaging	*Chairman: Georg Rønsch*
1545	Mohaman Ahmad	Spatial-spectral analysis of NIR imaging data - A case study (O18)
1610	Raffaele Vitale	Pixels that matter in chemical imaging (O19)
1630		End of scientific presentations
1830		Conference bubbles
1900		Conference dinner

Wednesday 08.09.2021

0700	John Holm & Pia Jørgensen	Morning run/ walk
	Machine learning	*Chairman: John Holm*
0845	Line Clemmensen	On deep learning for spectral data (O20)
0920	Elise Lunde Gjelsvik	Towards a machine learning based procedure for interpretation of mass spectra for better understanding of hydrate phenomena in oil systems (O21)
0945	Runar Helin	On the possible benefits of deep learning for spectral pre-processing (O22)
1010		COFFEE
	Applications	*Chairman: John Holm*
1040	Ekaterina Boichenko	Validation of classification models in cancer studies using simulated spectral data (O23)
1105	Anette Yde Holst	Towards successful silver anniversary with Advanced Process Control (O24)
1130		LUNCH
1230		Walk'n'Talk and Online Networking
	Multiway and multiblock	*Chairwoman: Pia Jørgensen*
1330	Hadrien Lorenzo	Multiblock supervised analyses should we really normalize blocks? (O25)
1355	Jean-Michel Roger	N-CovSel, a new strategy for feature selection in N-way data (O26)
1420	Helene Fog Froriep Halberg	Bitterness in beer – investigated by fluorescence spectroscopy and chemometrics (O27)
1445		COFFEE
	Applications	*Chairwoman: Pia Jørgensen*
1515	Mikko Mäkelä	Classification of cellulose textile fibres (O28)
1540	Gerjen H. Tinnevelt	Water quality control based on the analysis of high-resolution phytoplankton data (O29)
1605	Sonia Nieto-Ortega	Time domain reflectometry (TDR) and classification algorithms to detect injection of different water solutions in fresh tuna (O30)
1630		End of scientific presentations

1700		Poster session (P1-P15) & Dinner
1900	Pubcrawl	Jomfru Ane Gade

Thursday 09.09.2021

0700	John Holm & Pia Jørgensen	Morning run/ walk
	Process	*Chairwoman: Mette-Marie Løkke*
0830	Marco Reis	Incorporating expert knowledge and system structure in high-dimensional statistical process monitoring (O31)
0905	Iveth Romero	Online monitoring of H2S scavenging reactions in aqueous phase using Raman spectroscopy (O32)
0930		COFFEE
	Process	*Chairwoman: Mette-Marie Løkke*
1000	Sin Yong Teng	Chemical Quality Prediction by Inversing Dynamic PLSMAR: Balancing Interpretability and Accuracy (O33)
1025	Alessandro D'Alessandro	Process monitoring of a pesto production process through RGB Imaging and Near Infrared Spectroscopy (O34)
1050	Tim Offermans	Improved understanding of industrial process relationships through conditional path modelling with Process PLS (O35)
1115		Prizes and Closing of SSC17
1130		LUNCH & Farewell

O2: Optimization of UPLC-MS/MS assay for clinical diagnosis and therapeutic drug monitoring in patients with APRT deficiency by design of experiments

Unnur Arna Thorsteinsdóttir[1], Hrafnhildur L. Runolfsdottir[3], Vidar O. Edvardsson[1,3], Runolfur Palsson[1,3], Margrét Thorsteinsdóttir[1,2]

1. University of Iceland, Reykjavik, Iceland
2. ArcticMass, Reykjavik, Iceland
3. Landspitali – The National University Hospital of Iceland, Reykjavik, Iceland

e-mail: margreth@hi.is

Design of experiments (DoE) is an efficient tool for development and optimization of UPLC-MS/MS bioanalytical method which involves many experimental factors that need to be simultaneously optimized to obtain maximum sensitivity with adequate resolution at minimum retention time. Adenine phosphoribosyltransferase deficiency (APRTd) is an inborn error of adenine metabolism, characterized by excessive urinary excretion of poorly soluble 2,8-dihydroxyadenine (DHA), causing nephrolithiasis and chronic kidney disease (CKD) [1]. Treatment with the xanthine oxidoreductase (XOR) inhibitors allopurinol or febuxostat effectively reduces DHA excretion and prevents urinary stone formation and renal crystal deposition [2]. Currently, diagnosis and therapeutic drug monitoring (TDM) are performed by urine microscopy, which lacks specificity and is operator dependent. The aim of this study was to optimize a UPLC-MS/MS assay for clinical diagnosis and TDM of patients with APRTd utilizing DoE.

D-optimal design with several quantitative factors and multi-level qualitative factors was selected for experimental screening to reveal significant factors influencing the analysis of DHA, adenine, adenosine, 2-deoxyandenosine, inosine, 2-deoxyinosine and hypoxanthine in human urine by UPLC-MS/MS. Significant factors were studied via central composite face design and related to sensitivity, resolution and retention time utilizing PLS-regression. Urine samples from APRTd patients and healthy controls before and after treatment were analyzed with the optimized UPLC-MS/MS assay.

A sensitive UPLC-MS/MS assay for simultaneous quantification DHA and the main purine metabolites was successfully optimized utilizing DoE. There was a strong interaction effect between several variables, indicating that these variables cannot be independently controlled to obtain optimal conditions. DHA was detected in all urine samples from untreated APRTd patients but not in any specimens from healthy controls. Significant changes were observed in the urinary excretion of DHA and adenine with drug therapy and DHA excretion in APRTd patients decreased with conventional doses of both allopurinol and febuxostat. Today the UPLC-MS/MS assay is used for clinical diagnosis and TDM of patients with the rare kidney stone disorder APRTd.

This study demonstrates the utilization of DoE in ensuring that selected experiments contain maximum information and optimization is conducted efficiently. We believe the optimized clinical UPLC-MS/MS assay will greatly facilitate clinical diagnosis of patients with APRTd.

References

1. Runolfsdottir H.L.; Palson R.; Augustdottir I.M.; Indridason O.S.; Edvardsson V.O. Kidney disease in adenine phosphoribosyltransferase deficiency. Am. J. Kidney Dis. 2015, 67, 431-438.
2. Edvardsson V.O.; Runolfsdottir H.L.; Thorsteinsdottir U.A.; Augustdottir I.M.; Oddsdottir S.; Eiriksson F.; Goldfarb D.S.; Thorsteinsdottir M.; Palsson R. Comparison of the effect of allopurinol and febuxostat on urinary 2,8-dihydroxyadenine excretion in patients with Adenine phosphoribosyltransferase deficiency (APRTd): A clinical trial. Eur J Intern Med. 2018, 48, 75-79.

O3: Intervention studies on gut microbiota: Can ASCA compete with methods that are specifically tailored for microbial data?

Ingrid Måge[1], Maryia Khomich[2], Ida Rud[1], Ingunn Berget[1]

1. Nofima, Ås, Norway
2. University of Bergen, Norway

e-mail: ingrid.mage@nofima.no

Gut microbiome has recently gained considerable attention, and its composition and diversity has been linked to several aspects of health and disease. Intervention studies are often used to investigate how the microbiome is affected by external factors such as treatments and diets. Data from such trials need to be analysed by multivariate ANOVA-like methods.

Microbiome data have some special features. The raw data typically consist of millions of DNA reads, which are converted to taxa counts or abundances through advanced bioinformatic pipelines. The data is zero-inflated, and a high number of rare taxa are usually removed before further analysis. The statistical analysis can be performed on either sequence counts, abundances, transformed abundances or distances.

We have compared the chemometric-based ANOVA-Simultaneous Component Analysis (ASCA) [1] to a range of other ANOVA-like methods that are frequently used for analysing microbial data, including: PERMANOVA [2], ANOSIM [3], SIMPER [3], ALDEx2 [4], ANCOM [5], LEfSe [6] and 50-50 MANOVA [7].

Comparisons were done using simulated data and five real dietary intervention studies. We have evaluated the methods abilities to detect community-level (multivariate) effects, as well as their abilities to identify differentially abundant bacterial groups. We report on the overall agreement between the methods, to assess to what extent the choice of method affects the results.

References

1. Smilde, A. K. et al. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics 21, 3043-3048; (2005).
2. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26, 32-46 (2001).
3. Clarke, K. R. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18, 117-143 (1993).
4. Fernandes, A. D. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2, 15; (2014).
5. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Health Dis. 26, 27663; (2015).

6. Segata, N. et al. Metagenomic biomarker discovery and explanation. Genome Biol. 12, R60; (2011).
7. Langsrud, Ø. 50–50 multivariate analysis of variance for collinear responses. J. R. Stat. Soc. - Ser. D Stat. 51, 305317; (2002).

O4: ANOVA-PCA and PLS-DA for volatile metabolites chromatographic profile analysis as an alternative method for early, non-destructive and non-invasive detection of fungi species in Carica papaya (in vivo)

Larissa R. Terra[1,2], Sonia C. N. Queiroz[2], Daniel Terao[3], Marcia M. C. Ferreira[1]

1. Laboratory for Theoretical and Applied Chemometrics. Chemistry Institute. University of Campinas (Unicamp). Campinas. Brazil
2. Central de Resíduos e Contaminantes. Embrapa Meio Ambiente. Jaguariúna. Brazil
3. Laboratório de Microbiologia Ambiental. Embrapa Meio Ambiente. Jaguariúna. Brazil

e-mail: larissarochaterra@gmail.com

Carica papaya postharvest problems, such as diseases caused by fungi, generate huge economic losses for all those involved in an export chain. Thus, detection and identification of fungi species in the early stage are necessary and helpful for reducing the losses. Conventional methodologies are time-consuming, laborious, invasive, destructive, and can only be performed after the onset of symptoms in fruits [1]. It has been proposed recently in our laboratory an alternative method to uncover the metabolites produced by papaya's fungi species *in vitro* based on the volatile analysis by gas chromatography-mass spectrometry (GC-MS) and chemometrics. It was possible to determine some biomarkers that indicate the presence of fungi [2]. In this work, it is being proposed a non-invasive and non-destructive methodology, based on volatile metabolites analysis by GC-MS coupled to chemometric tools, for the in vivo early detection of three fungi species (*Alternaria alternata, Colletotrichum gloeosporioides, Lasiodiplodia theobromae*) frequently found in Brazilian papaya.

Fruits were previously inoculated by depositing 5-mm Potato Dextrose Agar medium (PDA) plug, containing mycelium of fungus in active growth, onto small wounds made on papaya surface. The inoculated and the control papayas (fruits only with small wounds) were placed in hermetically closed glass bottles. The system was allowed to stand before the analysis for the accumulation of volatile organic compounds (VOCs). The VOCs were collected by exposing an SPME fiber in the bottle headspace, and, then, they were analyzed by GC-MS. The analysis was performed in four replicates (four inoculated and four non-inoculated papayas) four times a week.

Conventional principal component analysis (PCA) and analysis of variance – principal component analysis (ANOVA-PCA) were used to perform an initial exploratory analysis. In the ANOVA-PCA, the influence of three factors on the data variability — class (inoculated and control papayas), "day", and "replicate" — and the interaction between two of them — class versus day— was investigated. Then, the PLS-DA method was used for the discrimination between the papayas inoculated with different fungi species and for the identification of the metabolites produced by each fungi species.

The distinction of the control and inoculated papayas was improved by the decomposition of the original matrix, according to the factors proposed in the experimental design, by ANOVA before

applying the PCA. Some metabolites as a primary alcohol with five carbons and diethyl phthalate were identified in infected papaya, and other metabolites such as phenylmethanol were only produced by healthy papaya.

The developed method has proven to be a potential alternative for the early diagnosis of fungi disease with small false negative and false positive rates, in addition to an accurate discrimination of the pathogenic fungal species in the fruits during postharvest storage.

References

1. Amorin L.; Bergamin A.; Rezende J.; Manual de Fitopatologia: Princípios e Conceitos. Ceres 2018, Volume 1.
2. Terra L. R..; Queiroz S. C. N.; Terao D.; Ferreira M. M. C. Detection and discrimination of Carica papaya fungi through the analysis of volatile metabolites by gas chromatography and analysis of variance-principal component analysis. J. Chemom. 2020, v. 34, n. 12, p. 1–13.

O5: Two-step authentication of overlapping classes

Zuzanna Małyjurek[1], Dalene de Beer[2,3], Elizabeth Joubert[2,3], Beata Walczak[1]

1. Institute of Chemistry. University of Silesia. Katowice. Poland
2. Plant Bioactives Group, Post-Harvest & Agro-Processing Technologies, Agricultural Research Council (ARC), Infruitec-Nietvoorbij. Stellenbosch. South Africa
3. Department of Food Science. Stellenbosch University. Stellenbosch. South Africa

e-mail: zuzanna.mitrega@op.pl

Class-modelling and discriminant methods are applied to construct mathematical models that are used to predict whether samples belong to the classes studied. Class-modelling methods, also known as one-class classification methods, are used for the construction of a class-model for the target class studied. The class-model is based on the similarities among samples of the target class. Whether a new sample belongs to the class is based on the similarity measures of this sample to the class modelled. If more than one target class are considered, then for each class an individual model is constructed, and the new sample is tested against each of them. The class-modelling is widely applied for e.g., food and drug authentication, product origin confirming, or process monitoring, since it enables rejecting a sample if it belongs to none of the classes studied, e.g., counterfeits, outliers, or samples of poor quality [1].
On the other hand, the discriminant model is based on the differences among the classes studied. The multivariate feature space is divided by the discriminant model into regions that correspond to the classes considered. A new sample is always predicted by the discriminant model as belonging to one of the classes accordingly to the region on which the sample is projected. The discrimination in its classical form cannot be used for authentication purposes since nontarget samples are always assigned as a member of one of the classes studied [1].
However, the class-modelling can lead to unsatisfactory results, when the goal is to authenticate classes which overlap, since samples from different classes are too similar. Thus, individual class models can incorrectly recognize similar samples as belonging to several target classes. In such situations, the discriminant model usually leads to better classification of the samples than individual class models, since it takes advantage of the differences between classes. However, the discrimination cannot be applied for authentication alone, thus we propose the two-step authentication of the overlapping classes that benefit from both class-modelling and discrimination [2]. The first step is the construction of the class-model for the training set consisting of samples from all authentic classes considered. The class-model is meant to identify samples that do not belong to any of the classes studied and can be regarded as potential counterfeits or samples of poor quality. The samples assigned by the class-model as belonging to one of the studied classes are in the second step discriminated into specific classes with a discriminant model. The discriminant model in the second step is constructed for the same training set as the class-model.
The performance of the two-step authentication approach is illustrated for three *Cyclopia* species, used for the production of honeybush tea. The two-step authentication approach enabled obtaining much higher classification results than in the case of class-models constructed for each of the *Cyclopia* species studied individually.

Acknowledgements: the authors acknowledge the financial support of the bilateral project PL-RPA2/04/DRHTeas/2019, co-financed by the National Research Foundation (NRF), South Africa,

(grant nr 118672 to DdB) and the National Centre for Research and Development (NCBR), Poland. Z. Małyjurek acknowledges the financial support from the project PIK, POWR.03.02.00-00-I010/17.

References

1. Oliveri P., Class-modelling in food analytical chemistry: Development, sampling, optimization and validation issues- A tutorial, Anal. Chim. Acta, 2017, 982, 9-19.
2. Rodionova O.Y.; Titova A.V.; Pomerantsev A.L.; Discriminant analysis is an inappropriate method of authentication, Trac. Trends Anal. Chem., 2016, 78, 17-22.

O6: Interplay of decision rules and parameter optimization strategies in SIMCA

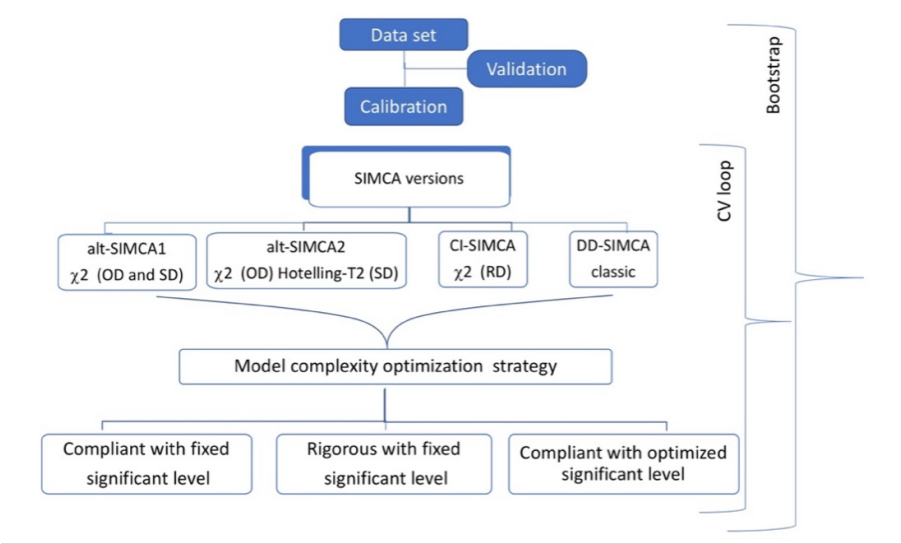
Raffaele Vitale[1], Valeria Carboni[2], Caterina Durante[2], Marina Cocchi[2]

1. Université de Lille, LASIR - Laboratoire de Spectrochimie Infrarouge et Raman, Lille (FR)
2. Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Modena (IT)

e-mail: marina.cocchi@unimore.it

SIMCA [1-2] is a well-established class modeling method based on building a disjoint principal component analysis model for each of the investigated classes. Its underlying classification rule is defined on the basis of the distance of every sample from (Orthogonal Distance) and within (Scores Distance) the model space of the concerned category. However, the way these distance measures are combined, and the distributional assumptions on which this classification rule is based lead to different implementations of the methodology. Although all over the years several works (one of the most recent being [3]) have surveyed the properties of such distinct implementations, far less studied is how they are affected by the optimization approach used to tune the SIMCA model parameters, i.e., the class subspace dimensionality/complexity and the significance level defining the distance boundary. For this reason, the main aim of this work is to assess the interplay between SIMCA versions (namely, two variants of the so-called *alternative* SIMCA – alt-SIMCA [2] – combined index-based SIMCA – CI-SIMCA [4] – and Data Driven SIMCA – DD-SIMCA [3]) and three different SIMCA model optimization strategies: i) significance level fixed at 95% and class model complexity optimized in cross-validation according to a "rigorous" criterion (i.e., by minimizing the difference with respect to the nominal classification sensitivity); ii) significance level fixed at 95% and class model complexity optimized in cross-validation according to a "compliant" criterion (i.e., by maximizing the classification efficiency) and iii) simultaneous significance level and model complexity tuning through the Receiver Operating Characteristic (ROC) curve-based procedure proposed in [5].

A flowchart of comparative assessment is shown below.



References

1. Wold, S. Pattern Recognition by Means of Disjoint Principal Components Models. Pattern Recogn. 1976, 8, 127-136.
2. SIMCA Model Builder GUI (http://wiki.eigenvector.com/index.php?title=SIMCA_Model_Builder_GUI)
3. Pomerantsev A.L., Rodionova O.Y. Popular decision rules in SIMCA: Critical review. J. Chemometrics 2020;34:e3250.
4. Joe Qin S. Statistical process monitoring: basics and beyond. J Chemometr. 2003;17, 480-502.
5. Vitale, R., Marini, F., Ruckebusch, C. Anal. Chem. 2018, 90, 10738−10747.

O7: Inter and intra class discrimination based on multivariate analyses applied on bacterial SERS fingerprints

Ana Maria Raluca Gherman[1], Nicoleta Elena Dina[1]

1. Department of Molecular and Biomolecular Physics, National Institute for R&D of Isotopic and Molecular Technologies, Donat 67-103, 400293 Cluj-Napoca, Romania

e-mail: raluca.gherman@itim-cj.ro

Overusing and misusing of bactericidal medication in the past years led to the rapid emergence of antibiotic resistance in bacteria. As a result, designing new antibiotics is a constant need for the medical sector in order to be able to control human infectious diseases caused by different pathogens which become more and more resistant to the classical medication. In order to overcome these needs, besides designing new medicine, one should be able to detect and identify the pathogens correctly before prescribing a treatment.
A first step that we took several years ago into the neverending marathon of antibiotic resistance was to develop a fast method for detection and identification of pathogens involved in human infectious diseases with the aid of Surface-Enhanced Raman Scattering (SERS) [1-4].
Most recently, part of our research is focused on designing statistical models able to discriminate between different classes and species of pathogens and further identify unknown samples by using these models. Here we present several multivariate analyses applied on database containing SERS fingerprints of both Gram-positive (*Staphylococcus aureus, Enterococcus faecalis*) and Gram-negative (*Pseudomonas aeruginosa*) bacteria by employing different chemometric methods such as principal component analysis (PCA), linear discriminant analysis (LDA) and PCA-LDA.

Acknowledgements: This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P1-1.1-TE-2019-0910, within PNCDI III.

References

1. Zhou H.; Yang D.; Ivleva N.P.; Mircescu N.E.; Niessner R.; Haisch C. SERS detection of bacteria in water by using in situ coating with Ag nanoparticles. Anal. Chem. 2014, 86, 3, 1525-1533.
2. Mircescu N.E.; Zhou H.; Leopold N.; Chiș V.; Ivleva N.; Niessner R.; Wieser A.; Haisch C. Towards a receptor-free immobilization and SERS detection of urinary tract infections causative pathogens. Anal. Bioanal. Chem. 2014, 406, 3051-3058.
3. Zhou H.; Yang D.; Ivleva N.P.; Mircescu N.E.; Schubert S.; Niessner R.; Wieser A.; Haisch C. Label-free in situ discrimination of live and dead bacteria by Surface-Enhanced Raman Scattering. Anal. Chem. 2015, 87, 13, 6553-6561.
4. Dina N.E.; Zhou H.; Colniță A.; Leopold N.; Szoke-Nagy T.; Comen C.; Haisch C. Rapid single-cell detection and identification of pathogens by using surface-enhanced Raman spectroscopy. Analyst. 2017, 142, 1782-1789.

O8: Two distinct frameworks to adapt source spectral calibrations to unlabeled target samples: (1) local modelling by linking linear classification to regression and (2) transfer learning

John H. Kalivas[1], Robert C. Spiers[1]

1. Idaho State University, Department of Chemistry. Pocatello, Idaho. USA

e-mail: kalijohn@isu.edu

Multivariate spectral calibration forms an accurate prediction model by correctly characterizing the relationship between sample spectral profiles and analyte concentration. Hampering real-time analysis is that sample spectra depend on measurement conditions such as humidity, temperature, instrument drift, manufacturer, etc., sample composition (analyte and other species amounts), and the physicochemical sample matrix effects from inter- and intra-molecular interactions. Thus, model performance degrades when target conditions are different from the original source calibration conditions. Needed are methods ensuring calibration and target samples are equally affected by these inherent hidden variables. This matching constraint is the crux of chemical analysis and two frameworks are presented to alleviate the problems. Both processes allow on demand modeling.

One approach is local modeling where it is presumed a subset of samples can be selected from a reference analyte library (encompassing a vast diversity of matrix effects) to form a linear model and predict a target sample. Current local modeling methods suffer because it is wrongly assumed that simple spectral similarity translates to the hidden matrix effect matching. The presented approach, termed local adaptive fusion regression (LAFR), solves the problem by considering local modeling as a classification issue where target samples are classified into linear calibration sets according to the respective hidden matrix effects. There are four stages to LAFR: (1) library searching by a fusion approach to decimate a large library into a reasonably-sized library spectrally similar to the target sample, (2) linear clustering of the smaller library using our indicator of system uniqueness (ISU) with another fusion process to form the calibration sets of distinctive hidden variables (matrix effects), (3) target sample classification into a calibration set by another ISU based fusion process using over a hundred similarity measures that are extended up to thousands using a novel cross-modeling technique, and (4) analyte prediction of the target sample by the selected calibration set. All LAFR hyperparameters are self-optimizing. Results from multiple near infrared (NIR) datasets demonstrate effective identification of hidden variables and great improvement over global models in difficult massive soil libraries with over 100,000 samples.

The second approach is model updating (transfer learning) based that forms a model orthogonal to the spectral matrix effect differences between source and unlabeled target samples. The process is termed null augmented regression (NAR). An impediment to adapting a model without target analyte reference values has been model selection. Due to multiple tuning parameters, thousands of models are typically formed. Presented is an automatic model selection process by model diversity and prediction similarity (MDPS). The unlabeled target samples to be predicted are used twice; to

form updated models and again to select the final predicting models. Thus, the models formed and selected are specific to these particular target samples. If new target samples need to be predicted, then new models may need to be formed depending on the degree of difference between the previously predicted target samples and the new target samples. Results for several NIR data sets are evaluated showing that MDPS selects reliable updated NAR models outperforming or rivaling prediction errors from total recalibrations requiring target reference values.

O10: Unified framework for calibration transfer

Valeria Fonseca Diaz[1], Bart De Ketelaere[1], Wouter Saeys[1]

1. KU Leuven, Division of Mechatronics, Biostatistics and Sensors, Kasteelpark Arenberg 30, 3001 Leuven, Belgium

e-mail: valeria.fonsecadiaz@kuleuven.be

The success of transferring calibration models contributes to diminishing the costs and waste involved in building models for new instruments or environments. Several methods have been proposed in the last two decades to successfully transfer models between instruments[1][2]. However, in many applications, the transferred models using state- of-the-art methods did not render models with satisfactory performance or models with highly noisy regression coefficients. We have elaborated a unified framework for transferring multivariate calibration models, defining the problem as a combination of instrument transfer and model specification transfer. This framework allows to position state-of-the-art methods for calibration transfer such as (Piecewise) Direct Standardization[3], Orthogonalization[4], [5] or Joint PLS[6] with respect to each other in order to define the conditions under which they will provide a successful transfer. These findings are summarized in generalized guidelines for calibration transfer including the most suitable methods and required number of samples for a successful transfer.

This work is part of the contribution with data and software in Python for Chemometrics users and the development of the unified framework will be available in the coming months for Open Access. Public data and software can be accessed and contributed to these public repositories

https://gitlab.com/vfonsecad/chemometrics_data

https://gitlab.com/vfonsecad/chemometrics_software

References

1. C. Pasquini, “Near infrared spectroscopy: A mature analytical technique with new perspectives – A review,” Anal. Chim. Acta, vol. 1026, pp. 8–36, 2018, doi: 10.1016/j.aca.2018.04.004.
2. J. J. Workman, “A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy,” Appl. Spectrosc., vol. 72, no. 3, pp. 340–365, 2018, doi: 10.1177/0003702817736064.
3. Y. Wang, D. J. Veltkamp, and B. R. Kowalski, “Multivariate Instrument Standardization,” Anal. Chem., vol. 63, no. 23, pp. 2750–2756, 1991, doi: 10.1021/ac00023a016.
4. A. Andrew and T. Fearn, “Transfer by orthogonal projection: Making near-infrared calibrations robust to between-instrument variation,” Chemom. Intell. Lab. Syst., vol. 72, no. 1, pp. 51–56, 2004, doi: 10.1016/j.chemolab.2004.02.004.
5. J. M. Roger, F. Chauchard, and V. Bellon-Maurel, “EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits,” Chemom. Intell. Lab. Syst., vol. 66, no. 2, pp. 191–204, 2003, doi: 10.1016/S0169-7439(03)00051-0.

6. A. Folch-Fortuny, R. Vitale, O. E. de Noord, and A. Ferrer, “Calibration transfer between NIR spectrometers: New proposals and a comparative study,” J. Chemom., vol. 31, no. 3, pp. 1–11, 2017, doi: 10.1002/cem.2874.

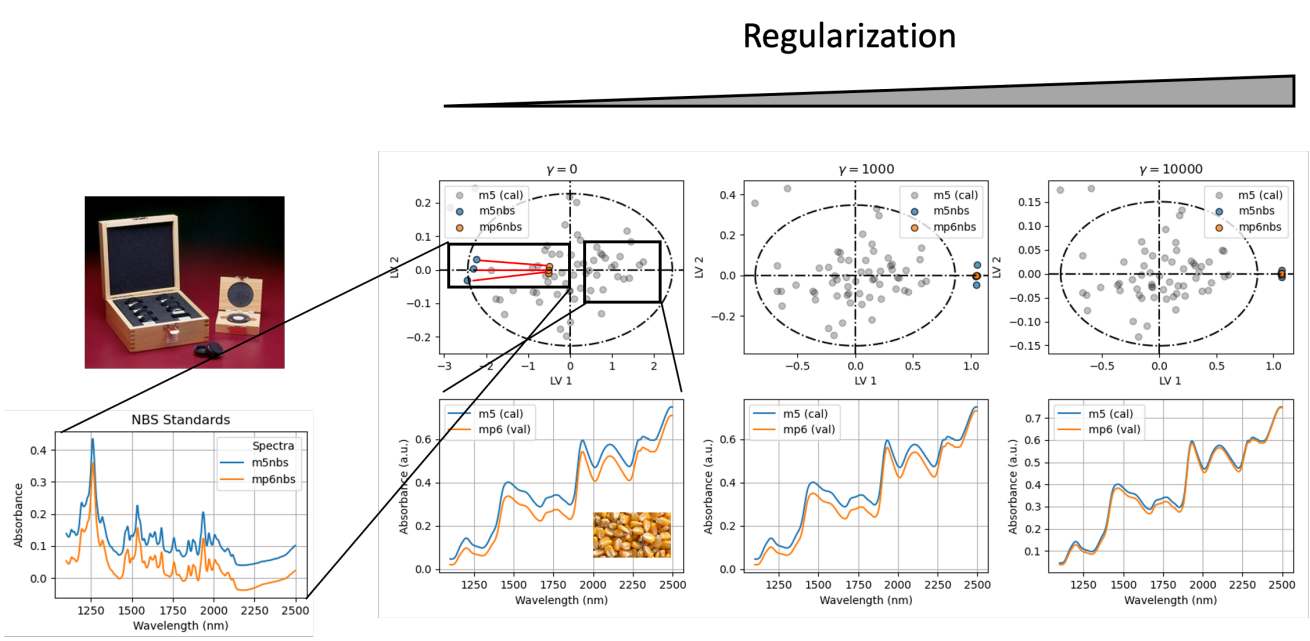
O11: Towards calibration transfer with arbitrary standards

Nikzad-Langerodi Ramin[1], Florian Sobieczky[1]

1. Software Competence Center Hagenberg. Hagenberg. Austria

e-mail: ramin.nikzad-langerodi@scch.at

Current state-of-the-art methods for calibration transfer (CT) require that the samples (i.e. the calibration standards), that are used to standardize the instruments (e.g. spectrometers) between which a calibration needs to be transferred and the samples (i.e. the calibration samples) for which the calibration is valid, have similar (spectral) features. In most studies on CT, that have appeared over the past decades, a (carefully selected) subset of the calibration samples are used as calibration standards. However, the CT methods that perform well in this setting are of limited use in a large number of real-world scenarios, e.g. if the calibration samples are chemically unstable. Towards enabling CT with "arbitrary" standards, we thus propose a Laplacian regularization scheme for partial least squares (PLS) regression, which allows building the primary calibration model under the constraint that the matched calibration standards, measured on the primary and the secondary device, have (nearly) invariant projections in the primary model's LV space [1]. To this end, we first derive the Laplacian of a bipartite graph over the matched standards and then construct a LV space (using the calibration samples) that trades-off preservation of the topology of this graph and predictiveness with respect to the response. Using the Corn benchmark data set, we empirically show that our approach allows transferring near infrared (NIR) calibrations on corn samples between similar instruments using glass standards from the national institute of standards (NIST). We further discuss some figures of merit that can be used to assess if CT using one type of samples and another type of standards is feasible for different samples/standards - pairs.



References

1. Nikzad-Langerodi, R, Sobieczky, F. Graph-based calibration transfer. Journal of Chemometrics. 2021;e3319. https://doi.org/10.1002/cem.3319

The research reported in this work has been funded from the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Federal Ministry for Digital and Economic A_airs (BMDW), and the Province of Upper Austria in the frame of the COMET - Competence Centers for Excellent Technologies programme managed by the Austrian Research Promotion Agency FFG, the COMET Center CHASE and the FFG project Interpretable and Interactive Transfer Learning in Process Analytical Technology (Grant No. 883856).

O12: Chemometrics extended to a parallel world of nondestructive Natural Chemical Computing

Lars Munck[1]

1. Department of Food Science, University of Copenhagen, Copenhagen, Denmark

e-mail: lmu@food.ku.dk

Scope: Welcome to a new Chemometrics extended by uncompressed *coherent* soft calculated chemical fingerprints analyzed by NIRS. They are independently calculated by Natural Computing (NC) within the organism (seed) visualizing the data structure behind a Principal Component (PC) analysis in biology. We are here referring to the key paper on coherence 2021 in Trends in Plant Science [1]. The local chemical fingerprints communicate as a global unit in a virtual *coherent* network of chemical bonds instructed by genotype and environment. If one significant local fingerprint changes e. g. NIRS all the others e.g., chemical-metabolic NC-patterns follows.

Results and data: Fingerprints are used as selectors for calibration with minimal mathematics in a recombinant material to breed for an optimal chemical composition in seeds. Each biological individual is created within one unique deterministic stochastic fertilization event [1: Figs III, 3], that has to be evaluated separately in pairs by NC calibration. Chemometric interpretation of *covariance by PC´s* includes a statistical error by biological ignorance. In-stead in a deterministic event the ignorance gap was filled by pattern descriptors suggested by Harald Martens [1:ref.25,Table 2] as with phenotypic *coherence* by NC described as a Linnaean differential two sample linearization of NIRS chemical patterns [1:Fig. 2A] far from population statistics. Physical mathematics was prestigious in technology because of *less coherence* in in-animate non-living matter. This was touted as a worth-while example of exact science in molecular biology. It is serious that *coherence* became ignored by the narrow anthropocentric molecular-causal SNP-gene marker definition of single traits at odds with NC and mathematics resulting in **genome chaos** [1:ref.18] with too many genes. The huge coordinative persistent power of coherence is visually demonstrated by genotype specific NIR-spectra. [1: Fig.1A] from 104 normal N barley 2280-2360nm seed samples grown in different environments (field light-/pots dark green) with protein varying from 9.7 to 21.0 %. They have a spectral std. of stunningly low 1.3%. representing the deterministic force of coherence. Chemical fingerprint calibration represents a deep theoretical understanding by Natural computing on how compressive chemometric PCA classification works when multigene/trait chemical composition is moved in one operation.

Implications: How can chemometrics rescue molecular genetics from conceptual chaos?[1]. As S. Wold and M. Sjöström warned in 1998 "*we must be careful not to separate chemometrics from chemistry*". There is no Science in megavariate chemometric (AI) machine learning apps *per se*. They work perfectly statistically in populations as a fast preliminary analysis but without the informative precision of deterministic visual two sample comparative fingerprint calibration by NC. AI do not promote a conceptual language and theory on biological meaning. Design of Experiment - DoE -should be introduced early to reveal data structure and "*visualize the effects and possibilities*" that Albert Einstein thought is more important than mathematics. "*Natural Computing*" is slow but precise, visually refuting man made "genome chaos" by deterministic fingerprint calibrators, descriptors, and selectors securing maximal conservation of information and a conceptual

theoretical language beyond chemometrics. *Casually directed Molecular geneticists do not accept soft chemo metrics. Natural computing does not compete with mathematical evaluation - instead NC visualizes by patterns to geneticists what is behind successful advanced chemometrics [1.ref 21]*

Acknowledgements: I am grateful to our institute and my faithful coworkers in allowing me generous research time and facilities since I left administration in 2001 for an interdisciplinary extension of chemometrics into the uncatered territory of real Natural Computing.

References

1. Munck L, Rinnan Å, Khakimov B, Møller Jespersen B, Balling Englesen S. Physiological Genetics Reformed: Bridging the Genome-to-Phenome Gap by Coherent Chemical Fingerprints-the Global Coordinator. TIPS, 2021, Vol. 26(4),325-337.DOI:https://doi.org/10.1016/j.tplants.2020.12.014

O13: Comparing calibration transfer approaches

Lars Erik Solberg[1], Tormod Næs[1], Ulf Indahl[1]

1. Nofima. Ås. Norway.
2. Norwegian University of Life Sciences. Ås. Norway

e-mail: lars.erik.solberg@nofima.no

Calibration transfer denotes several situations that have in common an existing calibration model and a new situation – a new instrument, a new recipe, drift over time – for which one would like to avoid the need for a full re-calibration. The primary motivation is to avoid the associated costs – monetary but also in terms of time and other resources. This problem has been addressed since the early '90s, and new approaches are constantly proposed in the literature. While the introduction of each new technique typically compares its performance with reference techniques on a couple of datasets, few broader and independent comparisons are known to the authors with the notable exception in Malli et al. [2], focusing on scenarios without standardization samples.

In our research efforts, we aim at describing why methods may work and therefore when they are applicable. In this presentation, we will restrict our focus on the simpler issue of comparing a selection of methods on simulated data.

One of the real scenarios that are of prime interest to the authors is the transfer of models from the *at-line* (laboratory) to an *in-line* (process) situation: when process monitoring uses models based on samples taken in-line, but analyzed at-line. In such scenarios, calibration models tend to result in poorer performance when used in-line, in spite of best efforts to ascertain comparable measurement situations. We will further assume the case where there exists a "standardization set": when samples have been measured both in-line *and* at-line. Therefore, the question is how methods perform when both calibration data at-line and standardization data are available for building models for use in-line.

The methods we will consider are: Dual Domain Transfer using Orthogonal Projections [4], Tikhonov regularization [5], Trimmed Scores Regression [6], Principal Components Canonical Correlation Analysis [7] and finally Piecewise Direct Standardization [3] as a reference method. These approaches are all relevant for the chosen scenario, and they span a breadth with regards to how the problem is addressed.

We hope our comparison strategy will provide some general indications on the choice of methods.

References

1. (Relevant review, Fearn, 2001 – old but good. Alternatively, Woodman, 2018 – but it is actually far inferior in my opinion) Author N.; Author N.; Author N.N.; etc. Title of Article. Journal Abbreviation Year, Volume, Inclusive Pagination.
2. Malli B, Birlutiu A, Natschläger T. Standard-free calibration transfer-An evaluation of different techniques. Chemometrics and Intelligent Laboratory Systems, 2017, 161, 49-60.

3. Bouveresse E, Massart DL. Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration. Chemometrics and intelligent laboratory systems, 1996, 32, 201-13.
4. Poerio DV, Brown SD. Dual-domain calibration transfer using orthogonal projection. Appl. spectrosc., 2018 72, 378-91.
5. Kalivas JH, Siano GG, Andries E, Goicoechea HC. Calibration maintenance and transfer using Tikhonov regularization approaches. Appl. spectrosc., 2009, 63, 800-9.
6. Folch-Fortuny A, Vitale R, De Noord OE, Ferrer A. Calibration transfer between NIR spectrometers: New proposals and a comparative study. J. Chemom., 2017, 31, e2874.
7. Fan X, Lu H, Zhang Z. Direct calibration transfer to principal components via canonical correlation analysis. Chemom. Intell. Lab. Syst., 2018, 181, 21-8.

O14: Rank Expansion (REX): a mathematical tunnel effect?

Carsten Ridder[1]

1. ERA Data Science ApS, Karlby, Denmark

e-mail:cr@carstenridder.dk

An absorbance spectrum **x** of a sample measured on a spectrometer contains signals from all chemical components absorbing radiation in the given frequency range. Thus, both the analyte **a** of interest and any interferences **b** present along with **a**, contributes to the sample spectrum **x**. Mathematically it can be formulated as $\mathbf{x} = c\mathbf{a} + \mathbf{b}$, where c is the concentration of the analyte in the sample. The purpose of analytical chemistry is to find a value of c as close as possible to the 'true' value. The spectrum **a** is assumed to be of unity concentration and **b** is the sum of all interferences present: $\mathbf{b} = \Sigma k_n * \boldsymbol{\beta}_n$.

If the spectrum has J wavelengths the problem can be written as a system of linear equations:

$$(x_1, x_2, ..., x_J)^T = c\ (a_1, a_2, ..., a_J)^T + (b_1, b_2, ..., b_J)^T$$

Given only **x** and **a**, the system is underdetermined having J+1 unknowns (c and **b**), but only J equations. Thus, infinitely many solutions exist for c, as $\mathbf{b} = \mathbf{x} - c\mathbf{a}$ satisfies the equations. The common solution is to build multivariate calibration models based on samples representing independent variations of the analyte and (all possible) interferences. I will here present an algorithm (*Rank Expansion*) that - in many, but not all - cases gives the unique and correct value c based on only **a** and **x**.

It is well-known, that second-order data, e.g. arising from fluorescence spectrometry, possesses the so-called second-order advantage. This implies that access to the excitation/emission-landscape of a sample containing the analyte alone (**A**) enables quantification of this analyte in the landscape measured on a sample (**X**). Mathematical we have $\mathbf{X} = c\mathbf{A} + \mathbf{B}$, where **X**, **A** and **B** are now matrices instead of vectors. The scalar c is, as before, the analyte concentration sought for. If we have a two-component system of analyte and one interferent, we can write (using fluorescence spectrometry as an example):

$$\mathbf{X} = c\ (a_1, a_2, ..., a_J)_{excitation}^T (a_1, a_2, ..., a_J)_{emission} + (\beta_1, \beta_2, ..., \beta_J)_{excitation}^T (\beta_1, \beta_2, ..., \beta_J)_{emission}$$

The rank of matrix **X** equals the number of components in the system (here two) and the rank increases one-by-one with the number of interferences in the sample. The second-order method Rank Annihilation Factor Analysis (RAFA) is based on the calculation of the rank of the reduced matrix ($\mathbf{X} - c_{guess}\mathbf{A}$). As the rank equals the number of components in the sample, the rank will drop with exactly one, when c_{guess} equals the correct value c. In practice, one examines when the f^{th} eigenvalue of the reduced matrix drops to a minimum, f being the number of components in the sample.

The REX-algorithm i) uses a sub-algorithm to transform the first-order data into second-order data: W(**z**) = **Z** and ii) investigates the eigenvalues of the matrices W(**x**) - c_{guess}W(**a**). Despite the fact that all matrices involved have full rank, including W(**x**) - cW(**a**), significant and distinct minima are nevertheless observed in all eigenvalues. For reasons still unknown these minima are - in many cases - found exactly at the correct analyte concentration *c*, and e.g. the median of the REX-estimates for the first nine eigenvalues gives from good to perfect fit in these cases.

I use the phrase *a mathematical tunnel effect*, because at least two 'classical' mathematical laws are obviously violated. I hope that 'crowd research' will reveal *why* REX works and - as important - *when* it works and when it does *not* work.

O15: Contextual Mixture of Partial Least Squares Experts: Integrating process specific characteristics into model structure

Francisco Souza[1], Michiel Theelen[1], Tim Offermans[1], Sin Yong Teng[1], Geert Postma[1], Jeroen Jansen[1]

1. Radboud University, Institute for Molecules and Materials, Analytical Chemistry & Chemometrics, Heyendaalseweg 135 6525 AJ Nijmegen, The Netherlands

e-mail: f.souza@science.ru.nl

There is more need than ever for the industrial digitalization towards a more sustainable and greener world. The artificial intelligence (AI) is at the front of the 4th industrial revolution, by redefining the decision making at the operational, technical and strategical levels, allowing a faster, data-driven, and whenever possible, automatic decisions along the value chain. This can reduce costs, the impact on the environment while increasing the process efficiency. In that sense, there is an increase demand on AI models that are explainable, or which at least can give valuable insights on the process to be modeled, instead of the pure black-box modeling in which the objective is only on predictive performance. The partial least square (PLS) model go in that direction, as it is robust to collinearity, noise, while being interpretable. In this work, we expand the power of PLS model through a new method called contextual mixture of partial least squares experts. This new approach integrates the process specific characteristics into the model structure, allowing the PLS to model the process more accuratelly, while providing the interpretability. In this approach, the process specific characteristics are assigned into distinct regions governed by an expert model. The contextual mixture of partial least squares experts overcomes the limitation of traditional modeling, in which the relation between the input-output data is mapped from a global perspective. This approach is very flexible in terms of modeling, and has show promising results in modeling industrial data.

O16: Alternative approaches to untargeted LC/GC-MS data analysis

Andrea Jr Carnoli[1,2], Geert Postma[1], Jeroen Jansen[1]

1. Department of Analytical Chemistry/Chemometrics. Radboud University.Nijmegen. Netherlands
2. Teijin. Arnhem. Netherlands

e-mail: andrea.carnoli@ru.nl

Data obtained by untargeted LC/GC-MS are characterized by high dimensionality, collinearity and noise. Therefore, exhaustive data analysis is required to retrieve information regarding similarity among samples and feature importance. One of the most common strategies is to process the data, quantify the signal and reduce the dimensionality using preprocessing algorithms and linear models such as Principal Component Analysis. However, PCA has the strong implicit assumption that the response within the measurement is linear with respect to the concentration of the data. For spectroscopic data this assumption is in many cases valid, but Mass Spectrometric data certainly not. Therefore, we explore alternative multivariate approaches that assume less stringent patterns than linearity, like presence/absence of certain ions as a fully qualitative approach to the data and Nonparametric Multidimensional Scaling which reduces the assumption of linearity to one of monotonicity in a multidimensional context. We critically analyze and compare the approaches on the results of the different case studies with several complementary Mass Spectrometric datasets.

O18: Spatial-spectral analysis of NIR imaging data - A case study

M. Ahmad[1,2], R. Vitale[1], C. Ruckebusch[1], M. Cocchi[2]

1. Université de Lille, LASIRE CNRS, Lille, France
2. Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Modena, Italy

e-mail:m.ahmad@live.nl.

Hyperspectral imaging (HSI) is used in many fields of science and industry for its powerful ability to capture information related to both the spatial and spectral domain, with applications ranging from cell imaging to remote sensing. However, in most cases, the spatial correlation structure encoded in hyperspectral images is disregarded by chemometric analyses that solely focus on the pixelwise unfolded data. Sometimes, though, it is difficult (if not counterintuitive) to ignore the interplay between spatial and spectral information, i.e., the so-called spatial-spectral correlation. A very clear example is NIR imaging of highly-scattering and complex samples with spatial non-homogeneities such as morphological/textural variation, object edges or fibers.

We present a case study of samples consisting of a cotton fabric on which semen stains have dried, analyzed with NIR imaging [1], where these spatial-spectral correlation structures are evident. There are significant scattering effects visible, with complete spatial and significant spectral overlap between the semen and cotton contributions. The data are shown in figure 1 (left side). The mean image clearly shows the fibers of the cotton fabric, and no clear information on where semen is present. When looking at an image for a single spectral channel (at 1714 nm), where the semen stain is the most identifiable, both cotton and semen show significant contributions. Due to the significant entanglement of the spatial and spectral information, the extraction of components spatially/spectrally becomes incredibly difficult for methods where only the spectral dimension is considered.

For this reason, we introduce here a novel methodological approach [2] that considers the spatial-spectral interactions to extract distinct spatial components underlying the hyperspectral imaging arrays, while simultaneously identifying their spectral contributions. The methodology utilizes wavelet decomposition applied to all the scanned wavelength channels followed by image encoding and multivariate analysis, to highlight distinct spatial features, while simultaneously correlating it to specific spectral features, allowing for the characterization of the physicochemical information associated to these images. In figure 1 (right side), the extracted images show the two spatially distinct components with their respective spectral contributions.

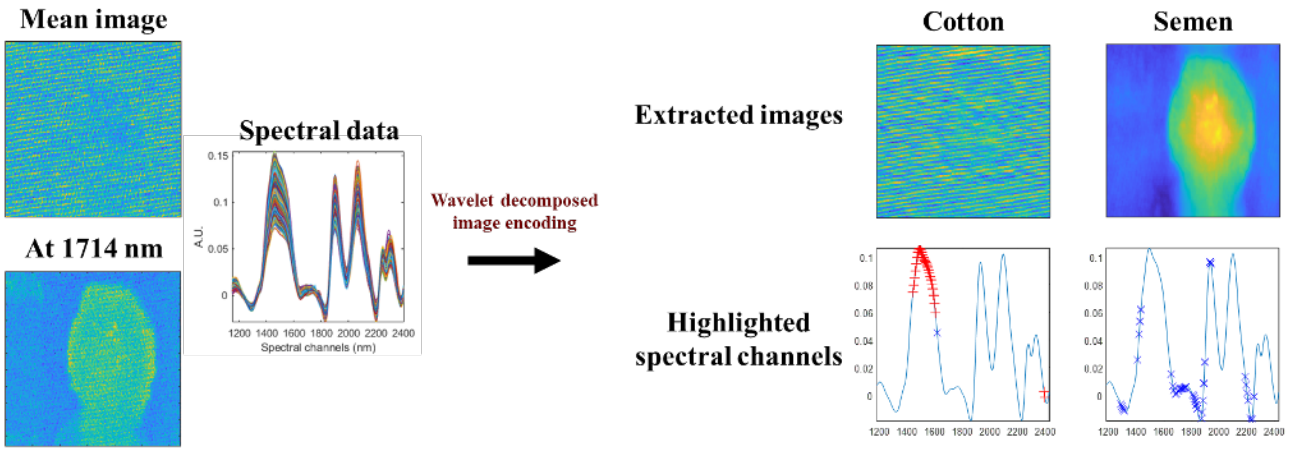


Figure 1: Results of the methodology; Semen stain on cotton fabric; NIR imaging data.

References

1. Silva C.S.; Pimentel M.F.; Amigo J.M.; Honorato R.S.; Pasquini C. Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models. Tr. in Anal. Chem. 2017, 95, p: 23-35.
2. Ahmad M.; Vitale R.; Silva C.; Ruckebusch C.; Cocchi M. Exploring local spatial features in hyperspectral images. J. of Chem. 2020, Volume 34, Issue 10.

O19: Pixels that matter in chemical imaging

Raffaele Vitale[1], Olivier Devos[1], Michel Sliwa[1], Cyril Ruckebusch[1]

1. Dynamics, Nanoscopy and Chemometrics (DyNaChem) Group. Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement (LASIRE CNRS – UMR 8516). Université de Lille. F-59000 Lille. France

e-mail: raffaele.vitale@univ-lille.fr

In statistics and unsupervised learning, by the term *archetypes* one refers to the most linearly dissimilar observations of a multivariate dataset which, geometrically, correspond to the points supporting its multidimensional convex hull [1]. Archetypes share an important mathematical property: all the other objects of the dataset, in fact, can be expressed as convex linear combinations of their archetypes' measurement vectors. This presentation aims at shedding light on how this aspect can have a tremendous impact when it comes to multilinear unmixing of chemical images, where the principal objective is unravelling the purest signal contributions ideally associated to individual compounds or species active within the inspected field-of-view.
Indeed, most, if not all, chemical imaging experiments (such as Raman hyperspectral microscopy or Fluorescence Lifetime IMaging – FLIM) leads to the generation of extremely redundant data, i.e., all scanned pixels are underlain by linear combinations of the aforementioned purest signal contributions. Thus, analysing their whole ensemble is not strictly necessary in the light of the final multilinear resolution. In such contexts, if users' attention were focused only on processing *essential* (archetypal) pixels, i) a dramatic decrease of the data to be handled would be achieved and ii) given the aforementioned mathematical property, the outcomes yielded by any multilinear unmixing approach applied to these reduced sets of recordings would converge to those ideally obtained when coping with entire imaging arrays. In other words, from a spectroscopic perspective, the same level of physico-chemical understanding of the investigated systems would be attained much faster and with far less intensive computational operations [2].
In this work, a recently developed methodology for essential pixel selection will be reviewed. In a nutshell, the main idea behind it is determining the convex hull of the data cloud spanned by the multiwavelength/multichannel pixels of a chemical image [3-5]. Examples of its application in challenging scenarios encompassing different analytical techniques and various decomposition approaches (ranging from multi-exponential fitting to Multivariate Curve Resolution – MCR-ALS [6, 7] – or PARAllel FACtor analysis – PARAFAC [8]) will be given.

References

1. Cutler A.; Breiman L. Archetypal analysis. Technometrics. 1994, 36, 338-347.
2. Ruckebusch C.; Vitale R.; Ghaffari M.; Hugelier S.; Omidikia N. Perspective on essential information in multivariate curve resolution. Trend. Anal. Chem. 2020, 132, article number 116044.
3. Ghaffari M.; Omidikia N.; Ruckebusch C. Essential spectral pixels for multivariate curve resolution of chemical images. Anal. Chem. 2019, 91, 10943-10948.
4. Ghaffari M.; Omidikia N.; Ruckebusch C. Joint selection of essential pixels and essential variables across hyperspectral images. Anal. Chim. Acta. 2021, 1141, 36-46.

5. Coïc L.; Sacré P.Y.; Dispas A.; De Bleye C.; Fillet M.; Ruckebusch C.; Hubert P.; Ziemons E. Pixel-based hyperspectral identification of complex pharmaceutical formulations. Anal. Chim. Acta. 2021, 1155, article number 338361.
6. Tauler R.; Smilde A.; Kowalski B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. J. Chemometr. 1995, 9, 31-58.
7. Jaumot J.; de Juan A.; Tauler R. MCR-ALS GUI 2.0: new features and applications. Chemometr. Intell. Lab. 2015, 140, 1-12.
8. Bro R. PARAFAC. Tutorial and applications. Chemometr. Intell. Lab. 1997, 38, 149-171.

O21: Towards a machine learning based produced for interpretation of mass spectra for better understanding of hydrate phenomena in oil systems

Elise Lunde Gjelsvik[1], Martin Fossen[1], Anders Brunsvik[2], Kristin Tøndel[1]

1. Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway.
2. SINTEF AS, Trondheim, Norway

e-mail: elise.lunde.gjelsvik@nmbu.no

Gas hydrates are crystalline structures consisting of gas molecules entrapped in cages formed by water molecules. Petroleum gas hydrates are formed from light hydrocarbons at low temperatures and elevated pressures, and can cause problems by plugging pipes and process units. It has been shown that some crude oils form hydrates that do not agglomerate but remain as "dry" particles that are transportable as a slurry in the oil phase. This self-inhibiting property is commonly accepted to be due to naturally occurring "hydrate active" components in the crude oil. Despite decades of investigation, their exact structures have not yet been determined. FT-ICR-MS (Fourier Transform Ion Cyclotron Resonance Mass Spectroscopy) offers a high mass accuracy and enables more detailed analysis of crude oil constitution. Correlations between hydrate properties and oil composition may be utilized as a parameter base for improved hydrate management strategies, better decision support tools and pipe flow simulations.

In this study, multivariate analysis and machine learning methods were used to extract relationships between the mass spectra and experimental data of the hydrate properties for various crude oils, with the ultimate aim of identifying the components responsible for natural hydrate anti-agglomeration. Latent variable-based methods (Principal Component Analysis (PCA), Partial Least Squares Regression (PLSR) and two extensions of PLSR - Hierarchical Cluster-based PLSR (HC-PLSR) and Sparse-PLSR) were used to decompose the data set into lower-dimensional subspaces for effective exploration of the co-variance patterns in the data and variable selection. Furthermore, Convolutional Neural Networks (CNNs) were used to extract higher level features from the data by using multiple neural layers. Commonly used in pattern recognition, CNNs are good feature extractors by learning the most important features by themselves. A combination of CNNs with latent variable-based methods such as PLSR may increase the robustness of the data analysis and improve the possibility of identification of the hydrate active components.

The aim of this study was to develop new data science methods that can correlate complex spectra from FT-ICR-MS to crude oil properties important for gas hydrate formation/inhibition. Our results indicate the potential to identify components associated to the gas hydrates, but more experimental data and method development is required to develop proper tools to interpret such data. Creation of robust methods to identify components that inhibit agglomeration of gas hydrates would provide new flow assurance and hydrate management strategies for transportation of crude oils with gas hydrates present.

O22: On the possible benefits of deep learning for spectral pre-processing

Runar Helin[1], Ulf Geir Indahl[1], Oliver Tomic[1], Kristian Hovde Liland[1]

1. Norwegian University of Life Sciences. Ås. Norway

e-mail: runar.helin@nmbu.no

Pre-processing is a mandatory step in most types of spectroscopy and spectrometry. The choice of pre-processing method depends on the data being analysed, and to get the pre-processing right, domain knowledge or trial and error is required. Given the recent success of deep learning-based methods in numerous applications and their ability to automatically detect patterns in data, we aimed at exploring the possibilities of using such methods for pre-processing. Two types of deep learningbased pre-processing techniques were considered. The first performed denoising and baseline correction and was based on the idea from Dong et al. [1]. The second is a novel implementation of an adaptive EMSC within the body of the neural network. Our study considered a flexible but systematic investigation of spectroscopic pre-processing methods (classical and deep learning-based) combined with predictive modelling, including both traditional linear modelling and artificial neural network based modelling.

The main ambition of the present work was to assess if the advantages of deep learning-based methods in spectral pre-processing are sufficient to justify the additional efforts in model setup and -training and the possible losses of interpretability and transparency. Using data from different vibrational spectroscopy techniques, we demonstrated that deep learning-based pre-processing successfully increased the predictive performance of our models but that classical pre-processing still is a good alternative, or even the best one in some cases. A significant increase in effort was required when using deep learning-based pre-processing together with linear model prediction. Compared with classical pre-processing techniques, deep learning-based pre-processing decreased the transparency and showed only modest improvements of the prediction performance of linear models. Our conclusion is that deep learning-based pre-processing is best suited when integrated in neural network predictions.

References

1. Jialin Dong; Mingjian Hong; Yi Xu; Xiangquan Zheng. A practical convolutional neural network model for discriminating Raman spectra of human and animal blood. J Chemometr. 2019, 33.

O23: Validation of classification models in cancer studies using simulated spectral data

Ekaterina Boichenko[1], Andrey Panchenko[2], Dmitry Kirsanov[3]

1. Laboratory of artificial sensory systems. ITMO University. Saint-Petersburg. Russia
2. Laboratory of carcinogenesis and aging, FSBI «N.N. Petrov National Medical Research Center of Oncology» St. Petersburg, Russia
3. Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia

e-mail: ekaterina.boichenko@inbox.ru

Spectroscopy has become a popular method in research devoted to cancer diagnostics, therapy, and surgery – anywhere we need to detect tumor cells surrounded by non-cancerous ones. Tumor cells show specific spectral features because of reprogrammed metabolism: higher oxygen saturation and glucose level, etc. Usually, different chemometrics or machine learning methods are applied to classify cancerous and non-cancerous sites, so most of the published studies can be described as "some spectroscopic method + some cancer + some data processing algorithm". As soon as a classification model appears in the study, a question is begged: how should we validate it? Unfortunately, there is no well-developed and, which is more important, concerted validation strategy for this type of the cancer studies. It makes the direct comparison of the reported results incorrect; besides, these models are often overfitted and the results are often over optimistic and sometimes misleading. The review of Sattlecker et al. reveals a significant problem that leave-one-out cross-validation (LOOCV) is often used as the only way to test multivariate models for cancer diagnosis [1], which seems to be insufficient in this case. Independent test sets are not always an option, because the collection of large number of samples can be difficult at the preliminary steps of research.

In this study, we suggest to use real data for simulation of new spectral sets with varying characteristics (size, distribution of classes). Simulated data can be used for testing validation algorithms, as long as they reproduce the real data properties, relevant for a classifier, adequately. It could be helpful when measured data set is limited and we cannot create different scenarios to challenge a classification model, e.g. when one class in underrepresented in a test set. Simulating the data, we can create a number of such data sets– an analog of "sandbox" used in software development – and to test how the model generalizes to unseen data.

Near-infrared spectra (939-1796 nm) measured from breast tumors and healthy tissues of mice (152 spectra) were used for simulation of spectral data sets of different size (100, 300, 500 spectra). We used a simple simulation method taking spectral baselines from the real data and adding Gaussian peaks, corresponding to the real peaks in near-infrared spectra, with randomly distributed noise. Reproducibility of the real spectral features was verified by Principal Component Analysis of real and simulated data sets and Tucker's coefficient. Several algorithms of train and test set selection have been applied to the simulated data (Kennard-Stone, DUPLEX, random, Monte-Carlo cross-validation), and corresponding Support Vector Machines classification models have been trained, optimized, and validated by using a series of test sets with varying "normal : tumor" classes distribution (1:1,3:1,1:3) and size (10%, 30%, and 50% of the training data set). Performance of the

classification models, expressed in values of accuracy, sensitivity, and selectivity, has been compared, and a validation strategy has been proposed.

The reported study was funded by RFBR, project number 20-38-90214.

References

1. M. Sattlecker, N. Stone, C. Bessant; Current trends in machine-learning methods applied to spectroscopic cancer diagnosis, Trends Anal Chem 2014, 59, 17–25.

O24: Towards successful silver anniversary with Advanced Process Control

Anette Yde Holst[1], Niels Anton Lang Sørensen[2], Julie Lundtoft Andersen[1]

1. Manufacturing Intelligence & Technology, Supply Chain Arla Foods, Aarhus, Denmark
2. Akafa, Svenstrup Skolevej 2, Svenstrup, Denmark

e-mail: Anette.yde@arlafoods.com

The backbone of Advanced Process Control (APC) systems consists of complex algorithms, hardcore mathematics, and sophisticated IT communication; however, the successful use of APC in production on daily basis requires a lot of non-complex maintenance as well. Arla has been happily "married" to APC systems since the early 2000's focused on spray dryers and the honeymoon is long gone. Production changes, new equipment, or new recipes are all factors that can affect the performance of the APC system. If these are not taken into consideration, it does result in decreased trust in the APC system and this again will lead to low utilization.

Low utilization is not only sorrowful for those who implemented APC, but is also a loss of earning opportunities. Higher and better utility results in decreased moisture variation, increased moisture content, and better exploitation of the raw milk.

We have found that one of the key solutions to increase utilization, is to stop treating the APC systems as standalone systems. They need to be better integrated with the already existing production IT and e.g. retrieve information from the Recipe Management System (RMS). Furthermore, established "ways of working", Responsibility assignment Matrix (RACI) as well as alignment and benchmarking across sites has paved the way for an increased utilization.

In addition, we have initiated APC v.2.0 as we believe as much information as possible is a way to optimize even further. Chemometric approaches help us to investigate new opportunities and introduce new information in the systems. For instance, information on the raw material allows the towers to run as fast as possible without compromising the risk of lump formation.

This presentation will show how we have improved the performance on certain parameters by introducing more information than standard APC systems resulting in increased throughput with consistently high quality milk powder.

Keywords: APC, continuous improvements

O25: Multiblock supervised analyses, should we really normalize blocks?

Hadrien Lorenzo[1], Rodolphe Thiébaut[2], Jérôme Saracco[1], Olivier Cloarec[3]

1. ASTRAL, INRIA BSO, 200 Avenue de la Vieille Tour, 33405 Talence, France
2. SISTM, INRIA BSO, 200 Avenue de la Vieille Tour, 33405 Talence, France
3. Corporate Research Advanced Data Analytics, Sartorius, Zone Industrielle les Paluds, Avenue de Jouques CS 71058,13781 Aubagne Cedex, France

e-mail: hadrien.lorenzo@inria.fr

In recent years, data analysis methods have had to deal with new type of heterogeneous data sets. Multi-omics studies are perfect examples of cases where such heterogeneous data sets are obtained. While these technologies are improving in terms of accuracy, the number of variables measured simultaneously for each observation is also rising tremendously. However, these measurements are also very often carried out on very small number of observations n compared to the number of variables. A block is a matrix of size $(n \times p_k)$ where p_k is the total number of variables in block $k \in [\![1, K]\!]$ and K is the total number of blocks. It is then common to have to deal with problems associated with data sets where some blocks are several tens or even hundreds of thousands of variables wide ($p_k \propto 10^{(4,5,6,...)} >> n$) which is denoted as the high-dimensional field. These high-dimensional data-sets are often treated assuming a latent variable model, meaning that a smaller number of variables are hidden from the users but can be estimated looking at empirical relationships between the different blocks. In the scope of this work, we deal with linear supervised analyses, and we focus on the Partial Least Squares (PLS) method and its sparse adaptation approaches, which allow to deal with high dimensional settings. Moreover, the latter have been adapted to multiblock analyses by reducing it to single block analyses where the covariate block $\boldsymbol{x}$ results from the concatenation of the different blocks divided by the square root value of the number of associated variables [1] such as

$$\boldsymbol{x} = (\boldsymbol{x}'_1/\sqrt{p_1},...,\boldsymbol{x}'_K/\sqrt{p_K})'.$$

Then, it was proposed to simply concatenate the different blocks of variables with no normalization [2] such as

$$\boldsymbol{x} = (\boldsymbol{x}'_1,...,\boldsymbol{x}'_K)'.$$

Later, authors from the *mbpls* **R**-implementation (accessible in the **ade4 R**-package [3]) focused on this solution but also gave the option to divide each block by its "total inertia" (meaning the square Frobenius norm) before concatenation. For interpretation, this solution is equivalent to assume that all the blocks must have the same influence in the final regression model while the "non-weighting" approach assumes that all the variables of all the blocks have the same influence.

What is the best solution?

We propose here to provide elements to answer this question by assessing different PLS-based methods, integrating variable selection, or not, in order to manage the large dimension of the data. We are going to show that the sparse PLS approaches provide different perspectives on how to answer this question. This study is going to be performed using simulations and real dataset applications are going to be presented.

References

1. Westerhuis J.A.; Kourti T. and Macgregor J.F.. Analysis of multiblock and hierarchical PCA and PLS models. Journal of Chemometrics 1998, vol. 12(5), p. 301-321.
2. Westerhuis J.A. and Smilde A.K.. Deflation in multiblock PLS (short communication). Journal of Chemometrics 2001, vol. 15(5), p. 485-493.
3. Bougeard S.; Dray S.. Supervised Multiblock Analysis in R with the ade4 Package. Journal of Statistical Software 2018, vol. 86(1), p. 1-17.

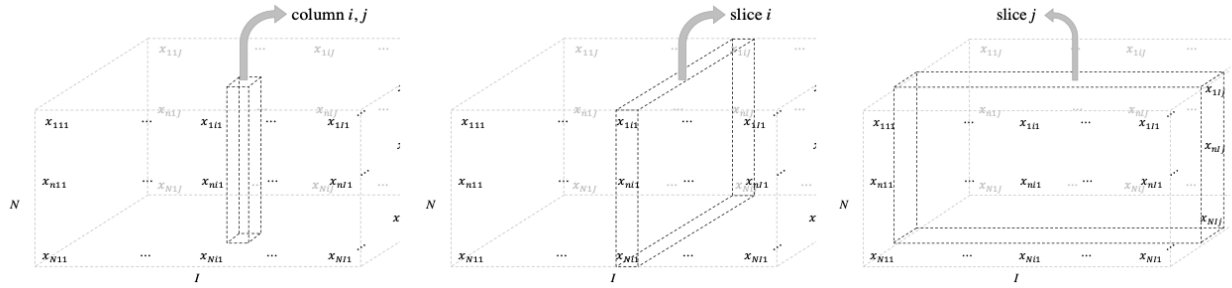
O26: N-CovSel, a new strategy for feature selection in N-way data

Alessandra Biancolillo[1], Federico Marini[2], Jean-Michel Roger[3]

1. University of L'Aquila, Via Vetoio, 67100, Coppito, L'Aquila, Italy
2. University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185, Rome, Italy
3. ITAP, Inrae, Montpellier SupAgro, University of Montpellier, Montpellier, France

e-mail: jean-michel.roger@inrae.fr

In data analysis, how to select meaningful variables is a hot and wide-debated topic and several variable selection (or feature reduction) approaches have been proposed into the literature. These methods aim at different purposes; they can be used to reduce the number of total variables and restrict it to the most significant ones for the problem under consideration, or simply for interpretative purposes, in order to understand which variables contribute the most to the investigated system.
In general, variable selection strategies are divided into three main categories: filter, wrapper and embedded methods. In addition to these three categories, a further meta-category, presenting intermediate characteristics between filter and embedded methods, can be identified. In fact, some feature selection approaches, like Covariance Selection (CovSel) [1], provide a filter selection based on model parameters embedded in the model building. CovSel is conceived to select variables in regression and discrimination contexts, and it assesses the features' relevancy based on their covariance with the response(s). Although variable selection methods are numerous and they have been quite widely debated into the literature, most of them refer to contexts in which data are collected in matrices, and not in higher order structures. How to assess the relevancy of variables in a multi- way context has not been extensively discussed yet. To the best of our knowledge, only Cocchi and collaborators developed a variable selection approach for multi-way data, extending the application of VIP analysis to high-order structures [2].
The present contribution, named N-CovSel, proposes to extend the CovSel principle to the N-Way structures, by selecting features in place of variables. Three main questions are addressed to achieve this: (i) How to define a feature in a N-Way array (Figure 1); (ii) How to define the covariance between a feature and a response Y; (iii) How to deflate a N-Way array with regard to a selected feature.
The complete algorithm of N-CovSel will be presented and its theoretical properties discussed. Two applications on 3 way real data will be presented, illustrating that the proposed method can be differently used, depending on the final purpose of the analysis. In fact, on one side, it represents a suitable option for the interpretation of N-way data sets, but, on the other, it can be applied prior to any regression or classification model in order to perform the analysis on a reduced, highly informative, sub-set of features.



References

1. J. M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, Chemom. Intell. Lab. Syst. 2011, 106, 216.
2. S. Favilla, C. Durante, M. L. Vigni, M. Cocchi, Chemom. Intell. Lab. Syst. 2013, 129, 76.

O27: Bitterness in beer – investigated by fluorescence spectroscopy and chemometrics

Helene Fog Froriep Halberg[1], Åsmund Rinnan[1]

1. Department of Food Science, University of Copenhagen, Copenhagen, Denmark

e-mail: helene.halberg@food.ku.dk

Beer is an old alcoholic beverage traditionally brewed by four ingredients: barley, water, hops and yeast [1]. The hops are, amongst other things, responsible for bringing bitterness to the beer. The lupulin glands of the hops contain α-acids which can isomerize to the bitter iso-α-acids during the wort boiling [1, 2]. The majority of the perceived bitterness in beer is caused by the iso-α-acids [3]. The beer bitterness is measured in International Bitterness Units (IBU) through an acidified iso-octane extraction where the absorbance is measured at 275 nm and multiplied by 50 [4, 5].

In order to give a proper estimate of the perceived bitterness one must be able to separate the iso-α-acids from the other beer constituents. This has already been achieved by High Performance Liquid Chromatography [5, 6]. The aim of this study was to investigate the beer bitter acids by front-face fluorescence spectroscopy in combination with PARAllel FACtor analysis (PARAFAC) [7], and for elucidation of whether this technique appears a suitable, rapid and non-destructive alternative for assessing the perceived bitterness in beer.

A dilution series for 34 beer samples and an isohumulone solution (iso-α-acid dissolved in water) were measured using front-face fluorescence. PARAFAC proved useful for elucidation of iso-α-acids fluorescence in beer which was observed in the same area where protein exhibit fluorescence. The fluorescence landscape displayed only one peak in the area of λ_{ex} = 280 nm and λ_{em} = 350 nm where the PARAFAC excitation and emission loadings revealed the presence of two components. This study, furthermore, found that using front-face geometry was in itself not enough to eliminate inner filter effects. Diluting the samples was needed to reveal the presence of iso-α-acids.

Altogether, front-face fluorescence spectroscopy in combination with PARAFAC appears promising in separating the bitter iso-α-acids from other beer components.

References

1. De Keukeleire, D. Fundamentals of Beer and Hop Chemistry. Quim. Nova, 2000, 23(1):108–112
2. Wang, G., Tian, L., Aziz, N., Broun, P., Dai, X., He, J., King, A., Zhao, P. X., and Dixon, R. A. Terpene biosynthesis in glandular trichomes of hop. Plant Physiol., 2008, 148(3):1254–1266
3. Keukeleire, D. D., Vindevogel, J., Szücs, R., and Sandra, P. The history and analytical chemistry of beer bitter acids. TrAC, Trends Anal. Chem., 1992, 11(8):275–280.
4. Analytica-EBC (2004). 9.8 Bitterness of Beer (IM)
5. Technical Committe, A. (2011). Beer Bitterness. In ASBC Methods of Analysis, pages 1–10. American Society of Brewing Chemists

6. Maye, J. P. and Smith, R. Dry Hopping and Its Effects on the International Bitterness Unit Test and Beer Bitterness. MBAA TQ, 2016, 53(3):134–136.
7. Bro, R. PARAFAC. Tutorial and applications. Chemom. Intell. Lab. Syst., 1997, 38:149–171.

O28: Classification of cellulose textile fibers

Mikko Mäkelä[1], Marja Rissanen[2], Herbert Sixta[2]

1. VTT Technical Research Centre of Finland, Ltd. Espoo. Finland
2. Aalto University, Department of Bioproducts and Biosystems. Espoo. Finland

e-mail: mikko.makela@vtt.fi

Classification of cellulose fibers plays an important role in chemical recycling of textiles. Cotton and regenerated cellulose fibers differ in cellulose structure and polymer chain length, which governs their viscosity after chemical dissolution and the ease in which the dissolved fibers can be spun again into regenerated textile fibers. Textile recycling is a timely topic as the global fiber production for textiles has doubled to over 100 million tons since the year 2000. Increasing textiles production and consumption are associated with decreasing average garment-use times, a trend which will inevitably increase the quantity of generated textiles waste. The need for textiles collection has been acknowledged by the European Commission, which requires EU member states to organize separate collection of household textiles waste by 2025. Chemical recycling of used textiles into man-made cellulose fibers enables converting waste materials into new fiber products with even improved mechanical properties while decreasing our future dependence on primary fiber production.

Here, we focus on the use of near infrared (NIR) imaging spectroscopy and chemometrics for supervised classification of cotton, viscose and lyocell fibers. These three fibers currently cover approximately 30% of annual textile fiber production and are challenging to identify quickly and accurately. Most recent studies on optical textile identification have reported classification of synthetic and natural fibers. These previous results are important for developing automated textile identification for efficient separation and sorting once the upcoming EU regulation on textile collection will be enforced. Chemical recycling of cellulose fibers, however, requires more detailed information on fiber composition and properties. We trained our binary classification models on pure and blended fiber consumer textiles and determined model performance on an independent pixel test of specifically made pure cellulose fabrics of known compositions. Our results showed that 9 out of 10 test set pixels were on average classified correctly using cross-validated PLS-DA models coupled with subset variable selection. These results enable estimating the potential of NIR imaging for cellulose fiber identification using only a limited number of pixel spectra, which is important for combining spectral imaging sensors with cameras operating in the visible range for faster image segmentation. This work continues our recent efforts on evaluating the potential of NIR imaging in determining textile properties with the aim of developing machine vision tools for chemical recycling of textile fibers.

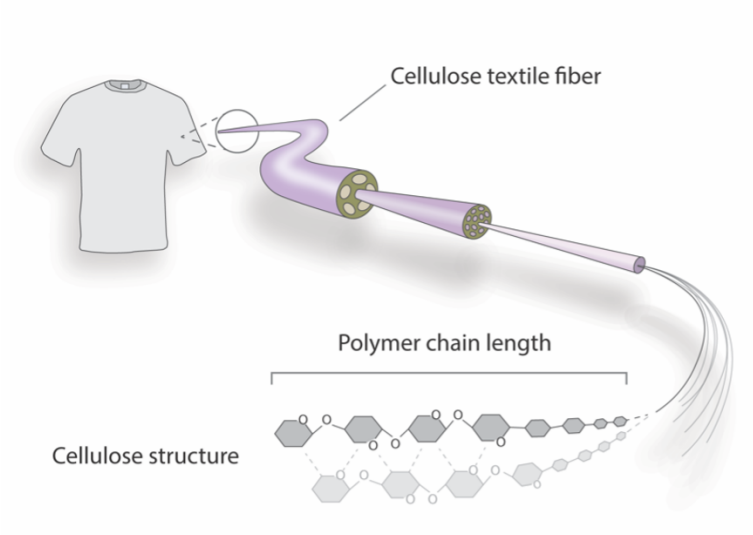


Figure: graphical abstract.

O29: Water quality control based on the analysis of high-resolution phytoplankton data

Gerjen H. Tinnevelt[1,2], Olga Lushchikova[1,2], Mathijs Lochs[1,2], Dillen Augustijn[1,2], Rinze W. Geertsma[3], Machteld Rijkeboer[3], Harrie Kools[4], George Dubelaar[4], Arnold Veen[3], Lutgarde M.C. Buydens[1], Jeroen J. Jansen[1]

1. Radboud University, Institute for Molecules and Materials, (Analytical Chemistry), Nijmegen, The Netherlands
2. TI-COAST, Amsterdam, The Netherlands
3. Laboratory for Hydrobiological Analysis, Rijkswaterstaat (RWS), Lelystad, The Netherlands
4. CytoBuoy bv, Woerden, The Netherlands

e-mail: gtinnevelt@science.ru.nl

River water is an important source for Dutch drinking water. For this reason, continuous monitoring of river water quality is needed. However, comprehensive chemical analyses with high resolution mass spectrometry (GC-MS/LC-MS) are quite tedious and time consuming, making them poorly fit for routine water quality monitoring and therefore many pollution events are missed. Phytoplankton are highly sensitive and responsive to toxicity, which makes them highly usable for effect-based water quality monitoring. Flow cytometry can measure the optical properties of phytoplankton every hour, generating a large amount of information-rich data in one year. This however requires chemometrics, as the resulting fingerprints need to be processed into information about abnormal phytoplankton behavior. We developed the Discriminant Analysis of Multi-Aspect CYtometry (DAMACY) to model the "normal condition" of the phytoplankton community imposed by diurnal, meteorological and other exogenous influences, see Figure 1. DAMACY first describes the cellular variability and distribution of phytoplankton in each measurement using PCA, and then aims to find subtle differences in these phytoplankton distributions that predict normal environmental conditions using (O)-PLS-DA. Deviations from these normal environmental conditions indicated abnormal phytoplankton behavior that happened alongside pollution events measured with the GC/MS and LC/MS systems. Thus, our results demonstrate that flow cytometry in combination with chemometrics may be used for an automated hourly assessment of river water quality and as a near real-time early warning for detecting harmful (un)known contaminants. Additionally we use automatic updating of the model to account for year-to-year variance. We are currently working on implementing this warning system such that drinking water companies can temporary stop pumping water whenever abnormal phytoplankton behavior is detected. In the case of prolonged abnormal phytoplankton behavior, comprehensive chemical analysis can still be used to identify the (un)known chemical compound, its origin and toxicity.

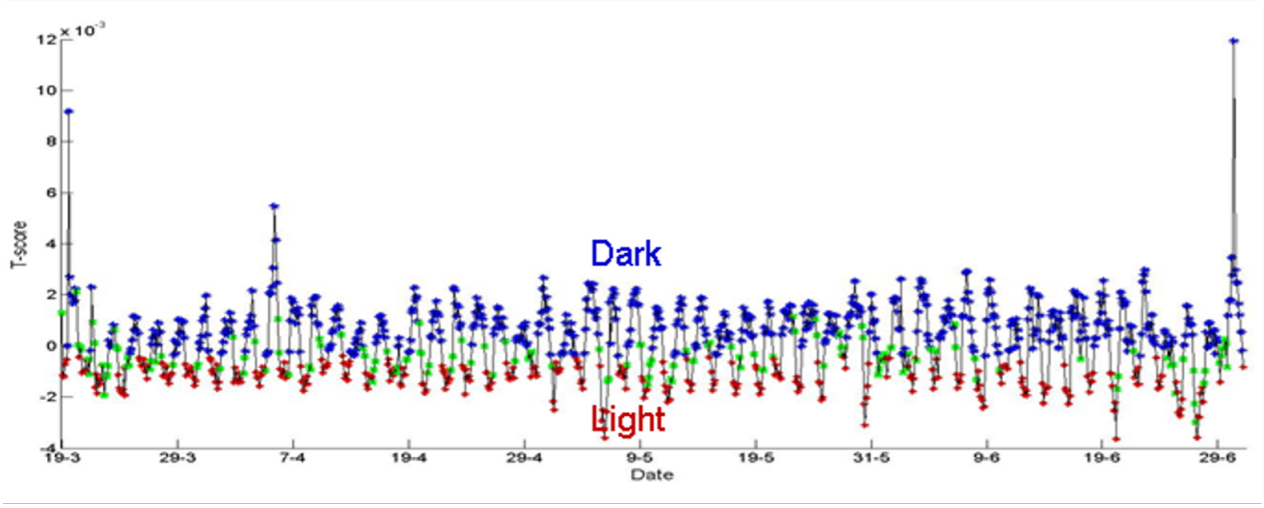


Figure 1: DAMACY on flow cytometry data of phytoplankton can successfully predict diurnal rhythms.

O30: Time Domain Reflectometry (TDR) and classification algorithms to detect injection of different water solutions in fresh tuna

Sonia Nieto-Ortega[1], Ángela Melado-Herreros[1], Idoia Olabarrieta[1], Giuseppe Foti[1], Graciela Ramilo-Fernández[2], Carmen G. Sotelo[2], Bárbara Teixeira[3], Amaya Velasco[2] and Rogério Mendes[3]

1. AZTI, Bizkaia Technology Park, Astondo bidea 609. Derio. Spain
2. Instituto de Investigaciones Marinas, CSIC, Eduardo Cabello 6. Vigo. Spain
3. Portuguese Institute for the Sea and Atmosphere (IPMA), Department for the Sea and Marine Resources. Lisbon. Portugal

e-mail: snieto@azti.es

Fish is susceptible to be affected by practices that can lead to mislabelling [1], as the non-reported water addition during harvest, processing, and storage to compensate moisture losses [2]. Though this water addition must be included in the label if it makes up more than 5 % of the weight of the product [3], this declaration is not always added, leading to mislabelling problems.

Non-destructive technologies, such as Time Domain Reflectometry (TDR), in combination with chemometrics, have emerged as powerful tools for the assessment of food quality parameters. TDR can probe the dynamics of dipolar molecules like water subjected to an alternating electric field, allowing a precise characterization of the dielectric properties of the sample. Changes in the amount of water in the muscle are expected to be detected by this technology [4].

In this work, a TDR sensor, (RFQ-SCAN®, Sequid company), coupled with two different classifiers (PLS-DA and SIMCA), have been used to detect the injection of different water solutions in fresh bigeye tuna (*Thunnus obesus*). A total of 11 dorsal and ventral loins were purchased in July 2018 and July 2019 (7 loins in 2018 and 4 loins in 2019). They were sectioned in portions of 500 g weight, and randomly distributed for the control and different injection treatments. The different treatments consisted of the injection of water with different additives: A. 3 % salt; B. 3 % salt + 3 % polyphosphates; C. 3 % salt + 5 % polyphosphates; D. 3 % salt + 5 % hydrolysate prepared from four-spot megrim *Lepidorhombus boscii*; E. a polyphosphate commercial solution.

TDR signals were pre-processed using several techniques: SNV with and without detrend, Savitzky-Golay 1st and 2º derivatives and different combinations. All of them were also mean centred. Then, data were divided into a calibration and a validation dataset using the Duplex algorithm [5]. Both classifiers separated correctly control samples, with a precision of 0.94 in PLS-DA and 0.91 in SIMCA. However, they failed in the differentiation of some of the treatments, such as A and C (0.00 and 0.25 of precision with PLS-DA and 0.33 and 0.57 in SIMCA respectively). The precision in separation of treatments B, D and E gave good results with PLS-DA (0.75, 0.80 and 0.75 respectively) but poorer results with SIMCA (0.67, 0.67 and 0.50).

As a conclusion, both classifiers work properly to separate control from the injected samples, but PLS-DA worked better to differentiate some of the injected treatments.

References

1. Pardo, M.A.; Jiménez, E. and Pérez-Villareal, B. Misdescription incidents in seafood sector. Food Control 2016, 62, 277-283.
2. van Ruth, S. M.; Brouwer, E.; Koot, A. and Wijtten, M. Seafood and Water Management. Foods 2014, 3, 622-631.
3. EU. Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers.
4. Jha, S. N., Narsaiah, K., Basediya, A. L., Sharma, R., Jaiswal, P., Kumar, R., & Bhardwaj, R. Measurement techniques and application of electrical properties for nondestructive quality evaluation of foods—a review. J. Food Sci. Technol. 2011, 48, 387-411
5. Snee, R. D. Validation of regression models. Methods and examples. Technometrics, 1977, 19, 415-428.

O32: Online monitoring of H_2S scavenging reactions in aqueous phase using Raman spectroscopy

Iveth Romero[1], Sergey Kucheryavskiy[1], Marco Maschietti[1]

1. Aalborg University. Esbjerg. Denmark

e-mail: iarl@bio.aau.dk

In offshore oil and gas industry, the removal of hydrogen sulfide from produced gas is typically carried out by injecting basic aqueous solutions of triazine-based scavengers. The method has been widely used due to relative simplicity of implementation and high efficiency, with 1,3,5-tri-(2hydroxyethyl)-hexahydro-*S*-triazine, also known as MEA-triazine, being the most common H_2S scavenger. Even though the reaction scheme of bisulfide ion (HS^-) and MEA-triazine is deemed to be well established, very little is known about the reaction kinetics. This lack of knowledge causes the operators to dose the scavenger in an empirical fashion, leading to a large excess of MEA-triazine typically used. This is undesirable, as it increases the cost for the operator and leads to the discharge of unreacted MEA-triazine into the sea. The availability of the reaction kinetics would allow to model the scavenging process, leading to a more efficient use of the scavenger at field operating conditions, thus decreasing significantly both the operational expenditures and the environmental impact.

One of the main obstacles in deriving a kinetic model for the scavenging reaction is the very high values of the reaction rate. Usually, substantial HS^- removal happens within seconds to minutes and no practical methods are available to sample and stop the reaction, which would allow off-line analytical laboratory methods. However, in a previous work from this research group [1], it was shown that Raman Spectroscopy is able to detect the main species and it is fast enough to capture several spectra during the reaction, thus enabling to follow trends.

In this work, we applied Raman spectroscopy for monitoring the aqueous phase reaction between MEA-triazine and HS^-. Building upon existing literature [1,2], Raman Identifier Bands were identified and proved to be suitable for tracking the evolution of the main reactants and products. The concentration of HS^- over time was predicted using a PLS-regression model. The model was calibrated based on samples containing only HS^- (in the range 10 to 100 mM) and acetonitrile as internal standard, using pre-selected characteristic peaks, which made possible to use the model for prediction of HS^- in a system containing additional chemical components (the scavenger and the reaction products). The model was used to study the reaction for different initial concentrations of the reactants and different initial pH values. All experiments were carried out at room temperature.

The presentation covers in detail the experimental methodology and the procedure for the data analysis. The main results are presented, showing the effect of the reactant concentrations and pH on the reaction rate.

References

1. Leah Johansen; Lykke Kloster; Anders Andreasen; Sergey Kucheryavskiy; Rudi P. Nielsen; Marco Maschietti. Raman Spectroscopy for Monitoring Aqueous Phase Hydrogen Sulfide Scavenging Reactions with Triazine: A Feasibility Study. Chem. Eng. Trans. 2019, 74, 541-546.
2. Rolando Perez Pineiro; Craig A. Peeples; Honathan Hendry; Jody Hoshowski; Gabriel Hanna; Alyn Jenkins. Ind. Eng. Chem. Res. 2021, 60, 5394-5402.

O33: Chemical quality prediction by inversing dynamic PLSMAR: balancing interpretability and accuracy

Sin Yong Teng[1], Tim Offerman[1], Francisco Souza[1], Geert Postma[1], Jeroen Jansen[1]

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

e-mail: sinyong.teng@ru.nl

Process variable predictions are critical building blocks for models constructed with the purpose of process optimization, process control, and other process improvement purposes to enhance production consistency, economic profitability and environmental sustainability. Recent black-box approaches for chemical quality predictions can achieve high accuracy predictions. Contrarily, conventional statistical methods may compensate for reduced model accuracy for increased explainability. Within the context of sustainable Industry 4.0, the implementation of effective algorithms requires both excellent accuracies and interpretability. In this work, we proposed a dynamic PLSMAR (Partial Least Square Multivariate Adaptive Regression Spline) model which can provide linear interpretability while maintaining considerably high accuracies for chemical quality prediction. We demonstrate that our model can be optimized to achieve competitive results against current state-of-art methods for two benchmark industrial debutanizer and desulphurization oil and gas refinery case studies. Concisely, the proposed PLSMAR method is compared with conventional partial least squares (PLS), support vector machines (SVM) and multi-layer perceptron (MLP) neural networks. The proposed method shows potential in modelling nonlinear process dynamics in the latent space while maintaining linear interpretability. Furthermore, the latent space of dynamic PLSMAR can be conveniently inversed into the variable space, giving insights for dynamic process optimization strategies.

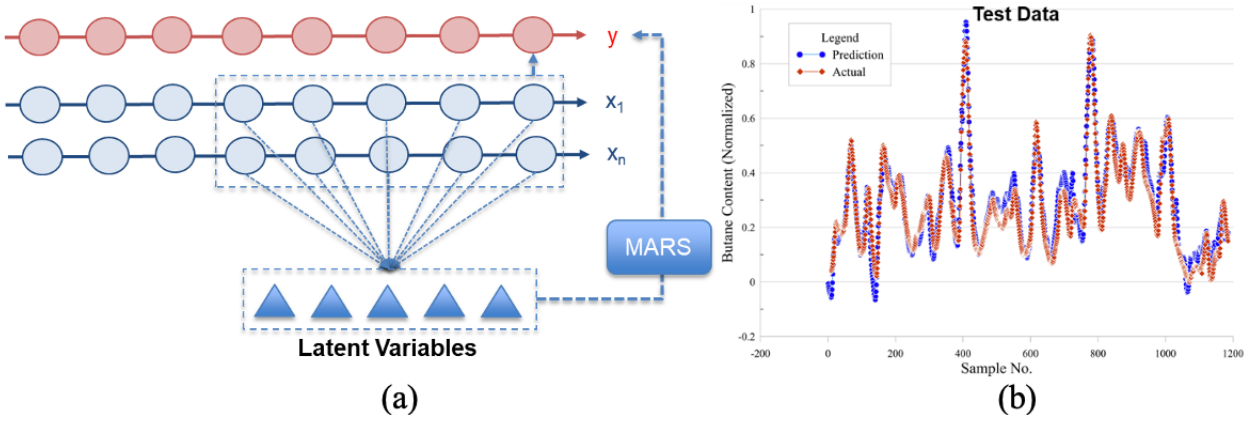


Figure 1: (a) Conceptual structure of the dynamic PLSMAR model (b) Prediction versus actual data from dynamic PLSMAR.

O34: Process monitoring of a pesto production process through RGB Imaging and Near Infrared Spectroscopy

Lorenzo Strani[1], Alessandro D'Alessandro[1,2], Caterina Durante[1], Marina Cocchi[1]

1. Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy
2. Barilla G. e R. Fratelli, Via Mantova 166, 43122 Parma, Italy

e-mail: alessandro.dalessandro@barilla.com

Barilla is one of the most important Italian food companies, well known worldwide not only for its pasta production, but also for other food product, such as ready sauces. Among them, pesto, a traditional Genovese sauce made by basil, extra-virgin olive oil, garlic, parmesan cheese and other minor ingredients, is one of the most popular product; consequently, its production process needs to be meticulously monitored, in order to provide the consumers with a high quality final product, reducing at the same time wastes and avoiding failures. The most important pesto sauce ingredient is Basil, as its composition strongly influences the final sensorial features of the product [1]. Therefore, it is crucial to control the first steps of pesto production process, where basil plants enter in the production line before being blended and mixed with the other ingredients. With this aim, a camera was installed above a conveyor belt based in the Barilla's production plant, collecting information about leaves color and defects, characteristics that strongly affect the quality of pesto sauce [2]. Furthermore, in order to have a better understanding of process progression, two Near Infrared (NIR) probes were placed in two different advanced steps of the process, acquiring spectra on blended basil leaves mixed with the other ingredients.
The aim of the work was to evaluate the quality of basil plants during 2020 basil harvesting season and among different basil suppliers, developing Multivariate Statistical Process Control (MSPC) charts for possible faults detection and for the real time monitoring of the process, using both Imaging and NIR data. Different strategies for data integration were compared, such as mid-level data fusion (DF) and multiblock low-level DF. Quality measurements on the final product were also modeled and used as reference data to assess normal operating conditions (NOC).

References

1. Bertoli, A.; Lucchesini, M.; Mensuali-Sodi, A.; Leonardi, M.; Doveri, S.; Magnabosco, A.; Pistelli, L. Aroma characterization and UV elicitation of purple basil from different plant tissue cultures. Food Chem. 2013, 141, Issue 2, 776-787.
2. Ciriello, M.; Formisano, L.; El-Nakhel, C.; Kyriacou, M.C.; Soteriou, G.A.; Pizzolongo, F.; Romano, R.; De Pascale, S.; Rouphael, Y. Genotype and Successive Harvests Interaction Affects Phenolic Acids and Aroma Profile of Genovese Basilfor Pesto Sauce Production. Foods 2021, 10(2), 278.

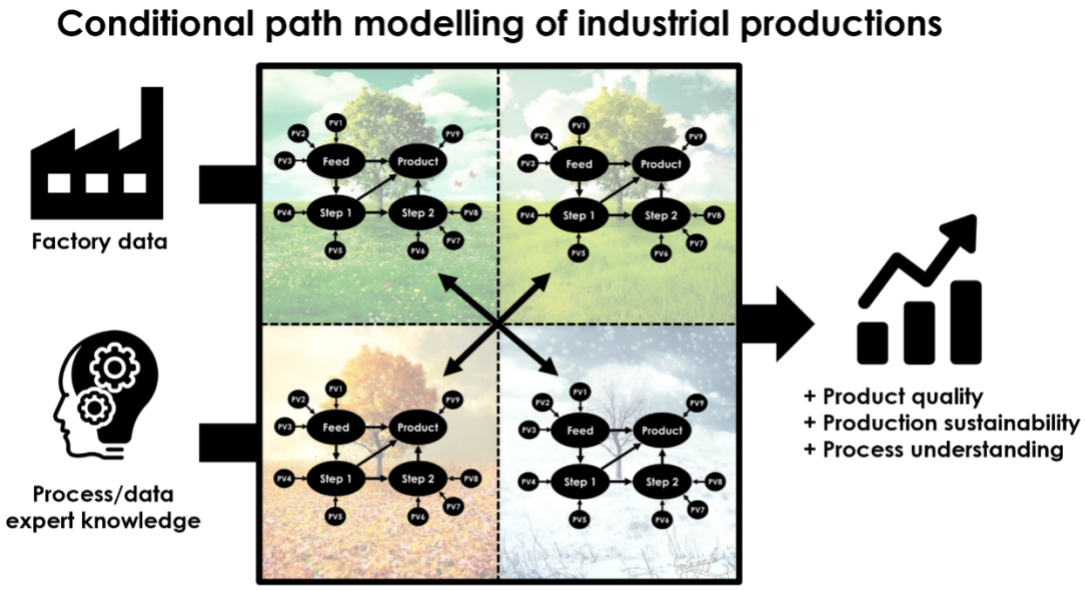
O35: Improved understanding of industrial process relationships through conditional path modelling with Process PLS

Tim Offermans[1], Geert van Kollenburg[2], Ewa Szymańska[2], Jeroen Jansen[1]

1. Radboud University, Nijmegen (the Netherlands)
2. FrieslandCampina, Amersfoort (the Netherlands)

e-mail: t.offermans@science.ru.nl

Understanding how different units of an industrial production plant are operationally related is key to improving production quality and sustainability. Chemometrics has proven indispensable in obtaining such understanding from vast amounts of historical process data. Path modelling is a valuable statistical tool to obtain such information from historical production data, and allows for a high level of structural knowledge on the plant and data to be incorporated in the analysis [1-3]. Investigating how relationships within a process are affected by multiple production conditions and their interactions can however an even deeper understanding of the plant's daily operation. We therefore propose conditional path modelling as an approach to obtain such improved understanding, demonstrated for a milk protein powder production plant. For this plant we studied how the relationships between different production units and steps are dependent on factors like production line, different seasons and product quality range. We show how the interaction of such factors can be quantified and interpreted in context of daily plant operation. Process PLS, which is a path modelling method recently developed to be optimally suited for analyzing industrial data, is used for this study [4]. Our analysis revealed an augmented insight into the process that can be readily placed in the context of the plant's structure and behavior. Such insights can be vital to identify and improve upon shortcomings in current plant-wide monitoring and control routines.



References

1. Van Kollenburg, G; et al. Understanding chemical production processes by using PLS path model parameters as soft sensors. Computers & Chemical Engineering 2020, 139
2. Höskuldsson, A; et al. Path modeling and process control. Chemometrics & Intelligent Laboratory Systems 2007, 88, 84-99
3. Bauer, M; Thornhill, N. A practical method for identifying the propagation path of plant-wide disturbances. Journal of Process Control 2008, 18, 707-719
4. Van Kollenburg, G; et al. Process PLS: Incorporating substantive knowledge into the predictive modelling of multiblock, multistep, multidimensional and multicollinear process data. Submitted to Computers & Chemical Engineering

Posters SSC17

P01	Fernando Montero	Application of Raman spectroscopy for monitoring of hydrogen sulfide scavenging reactions using biomass-based chemicals
P02	Isabelle M. A. Viegas	PARAFAC handles inner filter effects and FRET in fluorescence spectra
P03	Federico Marini	Class-modeling: Reviving old tools unjustly forgotten
P04	Kristrún Ýr Holm	Targeted proteomics and multivariate data analysis for search of novel biomarkers for early breast cancer diagnosis
P05	Manuela Mancini	Characterization of different waste wood for assessing the best reuse on the basis of their quality attributes
P06	Christian Bernhard Holth Thorjussen	Explorative and causal path modeling - limitations and synergies
P07	Gorka Albizu	Novel spectrophotometric method to determine simultaneously hypophosphite and phosphite in electroless baths
P08	Carlo G. Bertinetto	Comparing multivariate ANOVA methods in multicolor flow cytometry
P09	K. Magnus Åberg	When and how do artificial neural networks learn domain knowledge for near infrared food applications
P10	Irati Berasarte	pH measurement and phosphate determination in pharmaceutical eye drops for eye diseases by digital image analysis
P11	Tiffany Patra	The stability of oat drinks assessed using low field NMR T_2 relaxation
P12	Nicoleta Elena Dina	Finding new chemometric tools for SERS spectra cluster analysis and predictive modelling
P13	Ivan Krylov	Modelling of scattering signal for direct PARAFAC decomposition of excitation-emission matrices
P14	Ivan Krylov	Stochastic optimisation as a straightforward strategy for laser-induced calibration-free breakdown spectroscopy
P15	Ivan Krylov	Interpolation of scattering signal before PARAFAC processing of EEM-fluorescence spectra

P01: Application of Raman spectroscopy for monitoring of hydrogen sulfide scavenging reactions using biomass-based chemicals

Fernando Montero[1], Sergey Kucheryavskiy[1], Marco Maschietti[1]

1. Aalborg University. Esbjerg

e-mail: frmr@bio.aau.dk

Hydrogen sulfide (H_2S) is a toxic and corrosive species which is brought to the surface in oil and gas production and ends up in the produced oil, gas and water posing serious operating, health, safety and environmental problems. In offshore oil and gas production, H_2S is removed by direct injection of chemicals, called H_2S scavengers, which convert H_2S into by far less harmful species. The process is however not unproblematic, as H_2S scavengers are typically used in large excess than the stoichiometric requirement, with the excess often ending up in the water discharge into the sea thus increasing the environmental impact factor of offshore oil and gas production. For this reason, in recent years research on H_2S scavengers for offshore oil and gas industry has been directed towards the development of novel environmental-friendly chemicals, not posing risk for the aquatic life. One of the most promising chemicals of such type, currently under active investigation, is obtained by functionalization of sugars derived from biomass. The purpose of this work is to assess the feasibility of a method for monitoring the aqueous phase reaction of H_2S with one type of biomass-based scavengers, provided with a nitrile functional group capable of reacting with H_2S and generating a thioamide group. The monitoring method gives the possibility to make a quantitative assessment of the scavenging reactions, allowing to investigate how different factors (e.g. temperature, concentration of the chemicals) influence the conversion and the rate of the reactions. The method is based on Raman Spectroscopy, which is a fast and non-invasive analytical technique, which has been proved to be sensitive to changes in concentration of both H_2S as well as the biomass-based scavengers of interest. The method can be used both for qualitative assessment via monitoring of time developing of characteristic peaks as well as for quantitative estimation of concentration of the chemicals by using chemometric methods (PLS regression and MCR-ALS curve resolution). The presentation covers the details of the experimental set-up, data analytical procedure and presents preliminary results of on-line monitoring of the aqueous phase scavenging reaction at room temperature, as well as for off-line measurements of samples obtained by bubbling H_2S into aqueous solutions containing the biomass-based scavengers at both room temperature and 80 °C.

P02: PARAFAC handles inner filter effects and FRET in fluorescence spectra

Isabelle M. A. Viegas[1], Simon I. Andersen[1], Åsmund Rinnan[2]

1. The Danish Hydrocarbon Research and Technology Centre. Center for Oil and Gas, Technical University of Denmark. Kgs. Lyngby. Denmark
2. Chemometrics and Analytical Technology. Department of Food Science, Faculty of Science, University of Copenhagen. Frederiksberg C. Denmark

e-mail: iviegas@dtu.dk

Quantification and analysis by fluorescence spectroscopy of complex mixtures containing polyaromatic hydrocarbons such as biochar and crude oil are notoriously affected by inner filter effects and fluorescence resonance energy transfer (FRET) that lead to red-shift and quenching of fluorescence signal and therefore poor quantification [1,2]. In this context, our goal is to verify whether PARAFAC can properly handle fluorescence excitation-emission matrix (EEM) measurements with inner filter effects and FRET of oil by a straightforward methodology exempt of sample treatment or additional steps before the fluorescence analysis. Fluorescence EEMs of 11 oil (5 to 500 ppm) in toluene solutions were measured with excitation wavelengths from 290 to 450 nm (20 nm intervals) and 300 to 600 nm (0.1 nm resolution) for emission. Both right-angle (RA) and front-face (FF) cuvette geometries were measured, composing two 3-mode datasets with dimensions 11 × 3001 × 9 (samples × emission × excitation) that further were preprocessed to remove Rayleigh and Raman scatterings. Then, PARAFAC models with up to five components were fit and models with three components for both RA and FF were selected based on criteria such as total explained variance, split-half analysis, physic-chemical meaning of recovered profiles, core consistency and algorithmic indications [3]. The recovered profiles, shown in Fig. 1, were quite similar in RA and FF datasets, especially in emission and excitation modes. One component in each model (represented by yellow in Fig.1-(a) and red in Fig.1-(b)) had relative concentrations (scores) highly correlated with actual concentrations ($r = 0.9991$ and 0.9957 for RA and FF, respectively), in comparison to the conventional methodology that only considers the maximum emission intensity at a single excitation wavelength, which led to $r = 0.9362$ and 0.9742 for RA and FF, respectively, and lower sensitivity (slopes 30 and 29 times smaller, respectively for RA and FF) compared to the linear fitting with PARAFAC scores. Therefore, PARAFAC was proven to provide excellent fittings of EEM to actual concentrations of crude oil, which supports its ability to handle energy transfer effects in fluorescence of oil in toluene samples.

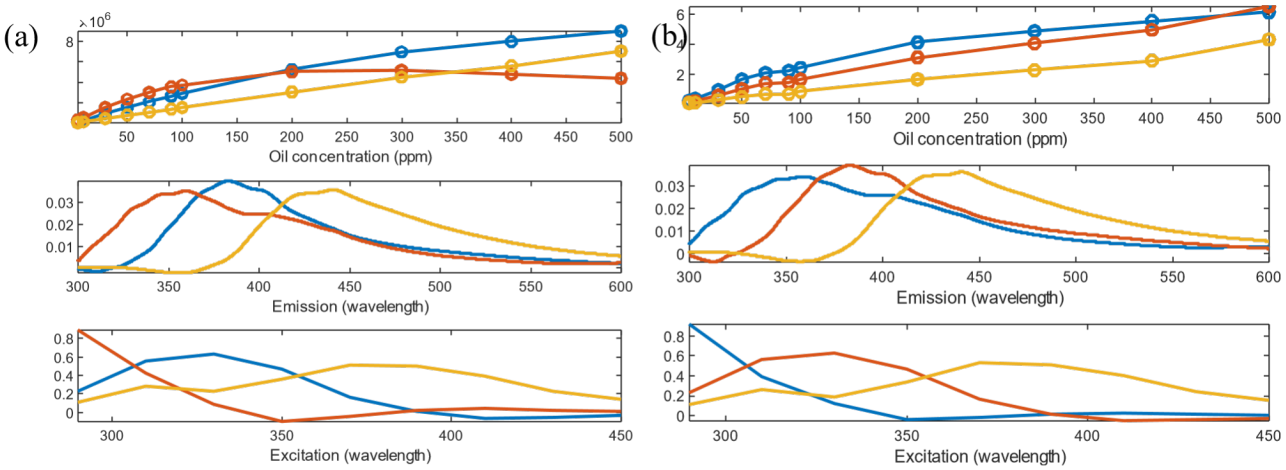


Figure 1 – Profiles recovered by PARAFAC for (a) right-angle and (b) front-face excitation-emission matrices of oil in toluene samples.

References

1. Zhang H.T.; Li R.; Yang Z.; Yin C.X.; Gray M.R.; Bohne C. Evaluating steady-state and time-resolved fluorescence as a tool to study the behaviour of asphaltene in toluene. Photochem. Photobiol. Sci. 2014, 13, 917-928.
2. Gao J.; Shi Z.; Wu H.; Lv J. Fluorescent characteristics of dissolved organic matter released from biochar and paddy oil incorporated with biochar. RSC Adv. 2020, 10, 5758-5793.
3. Murphy K.R..; Stedmon C.A.; Graeber D.; Bro R. Fluorescence spectroscopy and multi-way techniques. PARAFAC. Anal. Methods 2013, 5, 6557-6566.

P03: Class-modeling: Reviving old tools unjustly forgotten

Daniele Tanzilli[1], Federico Marini[1]

1. Dept. of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy

e-mail: federico.marini@uniroma1.it

The growing interest in chemometrics towards class-modeling tools, particularly triggered by problems where one-class/asymmetric problems are involved (e.g., food authentication), implies the need to evaluate alternative approaches to the more widely used UNEQ and SIMCA methods. In this work, a non-parametric class modeling method, originally proposed by M.P.Derde and co-authors in 1989 [1] and, since then, almost never used in real-world problems, was considered. In this method, the acceptance rule is based solely on the definition of a distance to the model (which can be that of the unknown sample to its closest or farthest element from the category, or a mean/median distance), which is then compared to a limit distance, identified as a threshold (e.g., based on the IQR, the fourth spread or the 95th percentile of the non-parametric distribution. The use of different datasets, of different dimensionality, highlights the classification efficiency of the method also for data characterized by many variables, both by considering them in the original hyperspace or after PCA projection.
Finally, to verify the validity of the method under study, the results obtained were compared with those of the SIMCA and, in this context, the results obtained showed how the predictive capacity of the two approaches and, in particular the sensitivity and specificity values, were comparable and indeed, as in some cases, the method evaluated, provided superior performance.

References

1. M.P. Derde, L. Kaufman, D.L. Massart, A non parametric class modeling technique, J. Chemometr. 3 (1989) 375-395.

P04: Targeted proteomics and multivariate data analysis for search of novel biomarkers for early breast cancer diagnosis

Kristrún Ýr Holm[1,2],Selma Dögg Magnúsdóttir[1], Finnur Eiríksson[1,3], Sigríður Klara Bodvarsdottir[2], Margrét Thorsteinsdóttir[1,2,3]

1. Faculty of Pharmaceutical Sciences, University of Iceland, Reykjavík, Iceland
2. BioMedical Center, University of Iceland, Reykjavik, Iceland
3. ArcticMass, Reykjavik, Iceland

e-mail: kyh4@hi.is

Breast cancer (BC) is the most common cancer among women in Western societies and the second leading cause of cancer deaths. Early detection of breast cancer is crucial for increasing survival rates and screening methods play therefore a vital role. The x-ray mammography is the most common screening method for early BC detection. Unfortunately, often in the early stages of the BC development the tumor is not visible on these mammographs and therefore insufficient for early detection. The importance to detect and treat BC early is critical in order to increase chances of survival. An improved performance of screening tests for early detection of BC is needed that is more reliable than x-ray mammography. By using targeted proteomics, we hope to identify novel early-stage biomarkers in plasma that can be used as an early diagnostic tool for BC.

In this study 100 biobank-based plasma samples, thereof 50 from BC cases and 50 controls, from a well-defined Icelandic BC study cohort were analyzed using MRM proteomics PeptiQuant™ 125-protein human plasma assay kit with UPLC-MRM-MS/MS analysis. Prior to analysis the plasma samples were proteolytically cleaved with trypsin and were concentrated by solid-phase extraction. Data analysis was conducted using Skyline Quantitative Analysis software and R. Multivariate data analysis (MVA) methods such as principal component analysis (PCA) and orthogonal projections to latent structures (OPLS), were performed using SIMCA Pro 17 to evaluate variation in protein content between controls and BC cases.

The MRM assay was successfully implemented for quantification of 125 proteins in human plasma. The proteins were measured in 50 BC patient samples and in 50 samples from healthy controls in two separate UPLC-MRM-MS/MS runs. Each batch contained 25 samples from BC patients and 25 samples from healthy controls. However, the sample handling of the two batches did not follow the same protocol. One of the batches contained higher concentration of internal standard then recommended in the protocol. A PCA score scatter plot revealed a distinct difference between the batches. Due to this difference the focus was set on the batch that contained recommended concentration of the internal standard. In that batch 112 proteins out of the 125 proteins were successfully quantified in all the plasma samples with acceptable precision and accuracy. PCA score scatter plot revealed great variability in the data. Furthermore, t-test was performed to observe if there was any significant difference in concentration of the proteins between cases and controls. The t-test revealed no significant difference between cases and controls in respect of protein concentrations. This insignificance might potentially be due to small study cohort size. However, these preliminary data suggest that some proteins are up- or downregulated in the plasma samples

from the cases compared to controls. These preliminary results indicate a difference in protein concentrations between cases and controls. Further investigation with a larger study cohort could possibly yield more conclusive results.

P05: Characterization of different waste wood for assessing the best reuse on the basis of their quality attributes

Manuela Mancini[1], Åsmund Rinnan[1]

1. Department of Food Science, Faculty of Life Sciences, University of Copenhagen. Rolighedsvej 30, DK-1958 Frederiksberg C. Denmark

e-mail: manuela@food.ku.dk

In the last decades, we have witnessed an expansion of waste wood (WW) trade because of the increasing demand for WW by the panel industry [1]. Furthermore, the great interest in WW is also related to the possibility to use it as a bioenergy feedstock, mitigating greenhouse gas emissions and contributing to meet the European energy targets. However, energy production is not always recommended because of the low-combustion quality of some materials, and related high pollutant emissions during the combustion process, that may cause environmental and/or combustion problems [2]. Therefore, suitable wood-based materials could be also used for the production of composites. This is in line with the Waste Framework Directive (2008/98/EC, European Parliament 2008) that favors the reuse of the material as a secondary resource instead of a waste to dispose of.

European countries have defined the quality indicators for the classification of WW in several categories on the basis of their quality and related applications. Consequently, the waste wood management is slightly different among the different countries and depends mainly on their policies. The material could be used for particle board production, recycled for wood composite or bioenergy production, based on the different WW categories established [3]. As an example, glued or slightly treated waste wood materials can be used as biofuel in some European countries while in other countries they can be used to produce, and aggregate value to, products such as composites (panels) [4]. In general, wood with preservatives, paintings or other chemical substances cannot be reused and should be sent to the disposal.

In this study more than 100 waste wood samples have been collected in different locations in Italy and Denmark. The WW material has been collected from static lots and as large pieces of wood in their original form such as items of furniture, fiber board or pallet. This allowed us to guarantee the actual source of the material, assign the most appropriate WW category and define the possible end-users and applications. Before the NIR analysis the particle size of the material and the sample size has been reduced to about 5 cm. All the samples have been analyzed by means of NIR spectroscopy and different SIMCA classification models have been developed with the aim to separate the material in three main categories: energy production, panel board production or landfill. The identification of waste wood characteristics is relevant for choosing the best-suited application. This will help in promoting the Circular Economy with consequent economic and environmental advantages and avoiding expensive landfills.

Acknowledgements: The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 838560.

References

1. Bergeron FC.; Energy and climate impact assessment of waste wood recovery in Switzerland. Biomass and Bioenergy 2016;94:245–57. https://doi.org/10.1016/J.BIOMBIOE.2016.09.009.
2. Tsai W-T; Wu P-H.; Environmental Concerns About Carcinogenic Air Toxics Produced from Waste Woods as Alternative Energy Sources. Energy Sources, Part A Recover Util Environ Eff 2013;35:725–32. https://doi.org/10.1080/15567036.2010.514594.
3. Hossain M; Poon C.; Comparative LCA of wood waste management strategies generated from building construction activities. J Clean Prod 177 387-397 2018:387–97.
4. Huron M; Oukala S; Lardière J; Giraud N; Dupont C.; An extensive characterization of various treated waste wood for assessment of suitability with combustion process. Fuel 2017;202:118–28. https://doi.org/10.1016/J.FUEL.2017.04.025.

P06: Explorative and causal path modeling - limitations and synergies

Christian Bernhard Holth Thorjussen[1], Ingrid Måge[1], Lars Erik Solberg[1], Kristian Hovde Liland[2]

1. Nofima AS. Ås. Norway
2. Norwegian University of Life Sciences/Faculty of Science and Technology. Ås. Norway

e-mail: christian.thorjussen@nofima.no

Historically, structural equation modeling (SEM), also known as path modeling (PM), is a set of methods for confirmatory statistical modeling (i.e., hypothesis testing) by estimating a set of linear regressions. In applied research, there are two main methodological frameworks for PM. Covariance-based SEM (CB-SEM / CB-PM) and partial least squares PM (PLS-SEM / PLS-PM). The former is prevalent in many academic disciplines, and the basis for estimation is minimizing the difference between the model implied variance-covariance matrix and the estimated population variance-covariance matrix [1]. PLS-PM is more established in chemometrics and marketing research, and estimation is done by iteratively fitting least squares regression equations, known as the PLS-PM algorithm. Both frameworks are mainly for confirmatory research, although PLS-PM is a framework that lends itself to some exploratory modeling [2]. In general, the practice for applied PM research is first to establish a theory, specify a model, collect data, and finally estimate (and re-estimate) the model with the data. Both frameworks have limited exploratory capabilities.

A new explorative approach to structural equation modeling has emerged in recent years — Sequential Orthogonalized Partial Least Square Path Modeling (SO-PLS-PM) [3]. The method estimates forward direct and indirect effects, in the form of explained variance, between blocks of data. The underlying dimensionality in each block can vary. Translated to path modeling jargon, SO-PLS-PM finds the latent variables within each block while estimating every structural path in a recursive model. The criterion for estimation is maximizing explained variance. The explorative nature of SO-PLS-PM comes from the fact that the numbers of components from each block is flexible, it can detect new and unknown paths, and traditional visualization tools such as scatter plots of scores and loadings can be used to interpret the contributions.

Using an empirical example, we will explore the potential synergies between SO-PLS-PM and established PM methods. We will investigate a stepwise workflow for structural equation modeling, where the first phase is an exploratory analysis using SO-PLS-PM, and the second phase uses confirmatory methods such as CB-SEM. The benefit of such a workflow can be understood as first controlling the risk of type II errors, with the exploratory approach, then controlling the risk of type I error by the latter confirmatory approach.

References

1. Bollen, Kenneth A. Structural Equations with Latent Variables. John Wiley & Sons, 1989.

2. Rigdon, Edward E; Sarsted, Marko; Ringle, Christian M. On Comparing Results from CB-SEM and PLS-SEM: Five Perspectives and Fice Recommendations. Marketing 2017, 39, pp. 4-16.
3. Næs, Tormod; Romano, Roasria; Tomic, Oliver; Måge, Ingrid; Smilde, Age; Liland, Kristian H. Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. Journal of Chemometrics. 2020. Special Issue

P07: Novel spectrophotometric method to determine simultaneously hypophosphite and phosphite in electroless baths

Gorka Albizu[1], Mikel Gutierrez[1], Miren Ostra[1], Maider Vidal[1]

1. Department of Applied Chemistry, Faculty of Chemistry, University of the Basque Country (UPV/EHU), 20018, Donostia-San Sebastian

e-mail:gorka.albizu@ehu.eus

Electroless deposition is one of the most widely used alternatives in the coatings industry, being nickel the most common metal and sodium hypophosphite the most widely used reducing agent, leading to Ni-P coatings [1]. The parameters that are most frequently controlled in this type of bath are temperature, pH and the concentrations of nickel, hypophosphite and phosphite, the latter being formed when hypophosphite is oxidized.

In industry, hypophosphite and phosphite concentrations are generally determined by iodometric back-titrations [2]. These methods are time consuming; each of them taking a minimum of 30 minutes and with the consumption of large amounts of sample and reagents. In this work, a colorimetric method using molybdenum is proposed to determine both anions simultaneously by UV-Vis spectrophotometry. Hypophosphite reacts with molybdenum giving rise to a blue complex with an absorption maximum at 752 nm. Although phosphite does not react, hypophosphite band is red shifted because of its presence, making the use of chemometrics necessary (Figure 1). This method only takes 5 minutes and consumes 5 μL of sample.

PLS models were built for each anion, and errors of 4.4%, 4.8% and 5.6% were obtained for hypophosphite and 8.7%, 11.3% and 7.0% for phosphite for calibration, cross-validation and external validation respectively. Ion chromatography was used as a reference technique in real samples and the relative errors were calculated. An average relative error of 4.2% was obtained for hypophosphite and 7.1% for phosphite and the method is to be implemented in an electroless laboratory-scale bath.

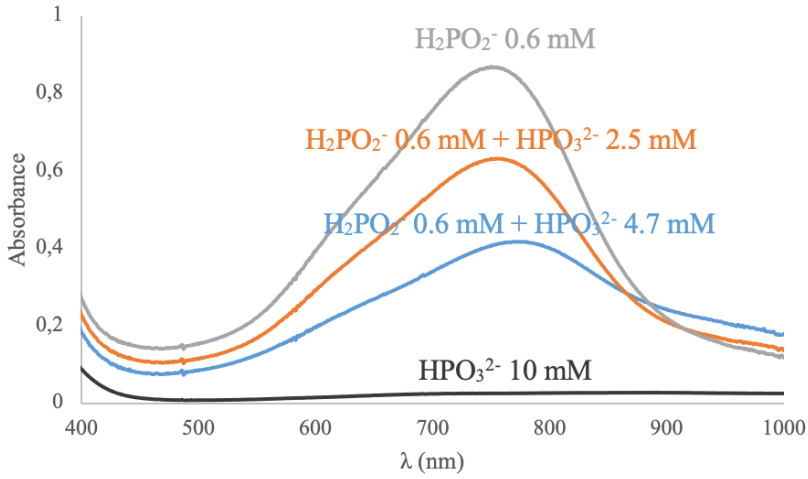


Figure 1. Effect of the presence of phosphite (HPO_3^{2-}) in the spectrum for the same concentration of hypophosphite ($H_2PO_2^-$) and the spectrum obtained only with phosphite.

References

1. Sudagar J.; Lian J.; Sha W. Electroless nickel, alloy, composite and nano coatings – A critical review. J. Alloys Compd. 2013, 571, 183-204.
2. Rossman J.; Portala F.; Kirner G.; Steinbach. Monitoring of Nickel Sulfate, Hypophosphite and Alkalinity in Electroless Nickel Plating Baths. The ProcessLab Analysis. 70771.

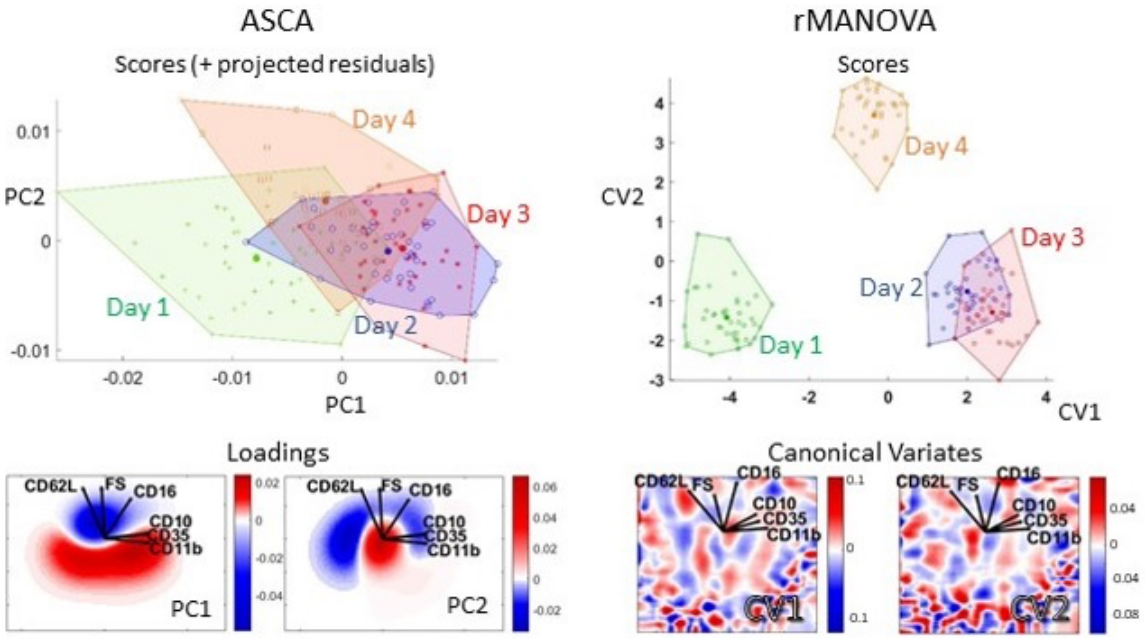
P08: Comparing multivariate ANOVA methods in Multicolor Flow Cytometry

Carlo G. Bertinetto[1], Jasper Engel[2], Jeroen J. Jansen[1]

1. Department of Analytical Chemistry, Institute of Molecular Materials, Radboud University, Nijmegen, The Netherlands
2. Biometris, Wageningen University & Research, Wageningen, The Netherlands

e-mail: c.bertinetto@science.ru.nl

Multicolor Flow Cytometry (MFC) produces highly complex data of marker (co)expressions. In multifactor experiments it can be particularly challenging to unravel the associations between marker expression and an experimental factor. One option is to transform the data into multidimensional histograms of their PCA scores [1], then model these histograms using a multivariate ANOVA approach. This latter step can be realized in several ways, of which this study intends to explore a few, including ASCA [2], rMANOVA [3], ANOVA-TP [4] and LiMM-PCA [5].
The data consists of cohorts of participants to a 4-Day long-distance walking event held in Nijmegen,The Netherlands. The results suggest that the different methods have advantages and disadvantages in terms of statistical power, interpretability and applicability for MFC analysis. An illustrative example is provided in the figure below: ASCA gives low separation but easily interpretable loadings, whereas rMANOVA achieves high separation (exploiting the correlations among the bins of the distribution maps) but its loadings (canonical variates) are much less clear. All the employed methods are compared and critically discussed with respect to the peculiarities of MFC data. This study should help understand which method is most suitable depending on the data and scope of the analysis, as well as possible ways to improve the results further.



References

1. Tinnevelt G.H. et al. Novel data analysis method for multicolour flow cytometry links variability of multiple markers on single cells to a clinical phenotype. Sci. Rep. 2017, 7, 1–11.
2. Jansen J.J. et al. ASCA: Analysis of multivariate data obtained from an experimental design, J. Chemom. 2005, 19, 469–481.
3. Engel J. et al. Regularized MANOVA (rMANOVA) in untargeted metabolomics, Anal. Chim. Acta. 2015, 899, 1–12.
4. Marini F. et al. Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach, J. Chromatogr. A. 2015, 1405, 94–102.
5. Martin M.; Govaerts B. LiMM-PCA: Combining ASCA+ and linear mixed models to analyse high-dimensional designed data, J. Chemom. 2020, e3232.

P09: When and how do artificial neural networks learn domain knowledge for near infrared food application

K. Magnus Åberg[1], Martin Lagerholm[1]

1. PerkinElmer, Stockholm, Sweden

e-mail: magnus.aberg@perkinelmer.com

Artificial neural networks (ANNs) have been used for decades in food applications with near infrared spectroscopic measurements. Even so, PLS models dominate among calibrations. There is a consensus in that ANNs require much more samples than PLS models and that ANNs are often worse at extrapolating compared to PLS [1]. Here, a case study where protein content in wheat determined by NIR, is presented where ANN models are compared to PLS-models with different pretreatments, number of factors and number of samples. The purpose of pretreatments applied to the spectra is to enhance and linearize the information that is related to the Y-parameter and suppress irrelevant information [2]. Choosing the right pretreatment requires domain knowledge and depend on the nature of the parameter. The more samples, the more irrelevant variation the models can handle, given that the samples span the irrelevant information independently of Y. Consequently, with more samples there is less need for pretreatments because the model can learn the domain knowledge. This is especially true for ANNs compared to PLS models. Most pretreatments destroy or alter part of the spectral information because the result of pretreatments as well as calibration equations are susceptible to noise and spectral artifacts. A pretreatment that introduces non-linearity in the data will be difficult for PLS to compensate for and we hypothesize that this is the reason for levelling out differently depending on the pretreatment. This is likely the reason why ANNs, being inherently non-linear, produce more accurate multivariate models than PLS when the number of samples is large enough.

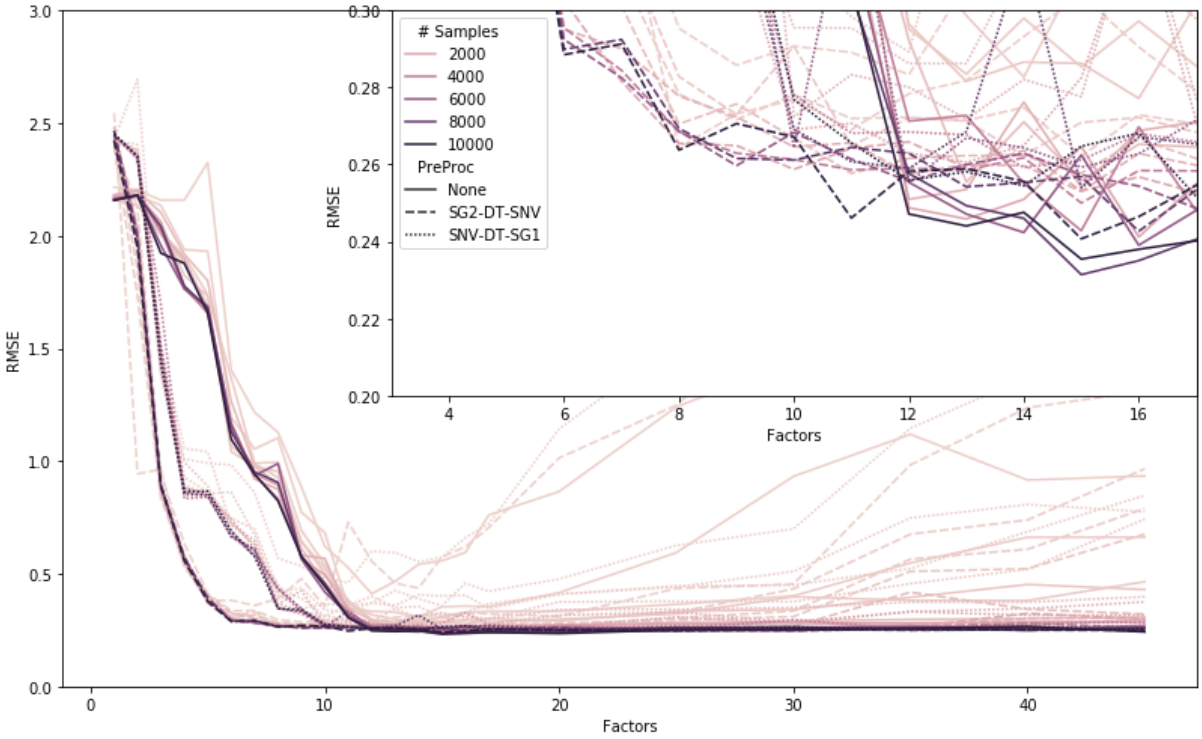


Figure 1. Performance of ANN models depending on the number of samples, pretreatment and PCA factors used in data reduction. Zoomed in part shows that no pretreatment is the best option with sufficient number of samples (> 4000) and PCA factors in the data compression.

References

1. Lagerholm M. A practical approach to ANN. Talk at 18th International Diffuse Reflectance Conference, 2016, Chambersburg, PA, USA
2. Thennadil S.N.; Martin E.B. Empirical preprocessing methods and their impact on NIR calibrations: a simulation study, J Chemometr. 2005, 19: 77-89.

P10: pH measurement and phosphate determination in pharmaceutical eye drops for eye diseases by digital image analysis

Irati Berasarte[1], Ane Bordagaray[1], Rosa Garcia-Arrona[1], Miren Ostra[1], Maider Vidal[1]

1. Department of Applied Chemistry, University of Basque Country (EHU/UPV), 20018 Donostia-San Sebastian, Spain

e-mail: irati.berasarte@ehu.eus

Eye drops are widely used in ocular diseases due to its good tolerance among patients. These drugs comprise pharmacologically active substance, preservatives and excipients in their composition. Excipients have to imitate the natural pH in tears (between 7.3 and 7.4) and although different buffers can be used, phosphate buffers are the most common ones. According to Directive 2001/83/EC, effective from November 2019, all excipients in ocular pharmaceutical products must appear on the labeling [1].

However, the increase of pH and phosphate concentration favors the formation of hydroxyapatite crystals and calcifications in the cornea. In this work, phosphate concentration and pH were determined by digital image analysis (DIA) and a microplate for sample support [2]. Phosphates were quantified by measuring the color of the blue complex formed with molybdenum, linearly relating the values of the R channel of the RGB system with the concentration. For pH determination, the acid-base indicator bromothymol blue was used. A color change can be appreciated around the pKa value of 7.1: yellow below pH 6, blue above 7.6 and different shades of green between them. In consequence, the color change cannot be modelled by linear regression and a more complex model such as Partial Least Square Regression (PLS) is necessary, using the whole histogram of each well as signal.

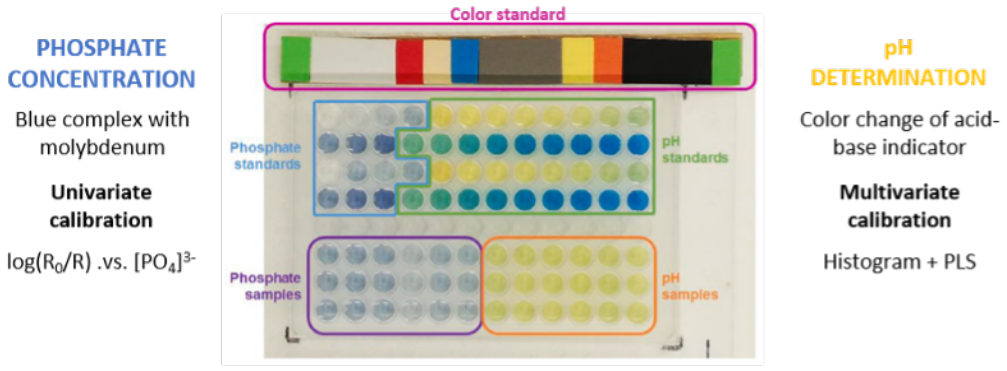


Figure 1. Image of a microplate containing the standards and samples for phosphate and pH determination. A color standard was added for illumination control. Left and right, explanations for both measurements.

The optimum PLS model was obtained with four latent variables. This model shows a Root Mean Square Error of Calibration (RMSEC) of 0.10 and a Root Mean Square Error of Validation (RMSEV) of 0.46. It was observed that minimum and maximum pH values tend to deviate from linearity, but eliminating these standards did not improve the model, so they were not removed.

Ten different commercial ophthalmic eye drop samples were analyzed, five of them used in the treatment of glaucoma and the other five in the treatment of dry-eye disease. The values obtained by DIA were compared with the ones obtained by reference methods, ion chromatography for phosphate determination and potentiometry for pH. Relative errors between 0.10 and 13.4% were obtained for phosphate and between 0.10 and 11.4% for pH except for one sample (18.4%) and no significant differences were found between DIA and reference methods.

References

1. DIRECTIVE 2001/83/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 November 2001 on the Community code relating to medicinal products for human use.
2. Berasarte I.; Bordagaray A.; Garcia-Arrona R.; Ostra M.; Vidal M. pH measurement and phosphate determination in pharmaceutical eye drops for eye diseases by digital image analysis. Microchem J 2021, 162, 105854.

P11: The stability of oat drinks assessed using low field NMR T2 relaxation

Tiffany Patra[1], Karsten Olsen[1], Åsmund Rinnan[1]

1. Department of Food Science. University of Copenhagen. Copenhagen. Denmark

e-mail: aar@food.ku.dk

Oat-based drink (OBD) is a heterogenous colloid dispersion, which tends to have poor stability. Characterization of the stability of the product is currently done mainly using conventional quality measurements, which is time-consuming (storage experiments) or need matrix disruption (particle size, zeta potential). On the other hand, low field NMR relaxation provides rapid in-situ measurement of the product's water mobility and can potentially describe the separation phenomena in OBD. This study aims to characterize the stability of oat drinks using the low field NMR relaxation technique compared to traditional stability methods and investigates the best curve resolution techniques for such datasets. OBD with varied separation characteristics were made using a 2^3 factorial design with oat source, β-glucanase level, and product pH as variables, all at two levels (respectively to the variables, wholemeal oat flour or oat bran, with or without β-glucanase, and pH 4.2 or 6.4). The stability of the products was examined with separation tendency measurement through 14 days storage test at 20 °C and rheology tests. Three different types of curve resolution techniques are used to assess the relaxation curves; discrete exponential fitting, power slicing, and double slicing. The best resulting T_2 values and proton concentrations are then compared to the stability profile. The OBD tested in this study showed phase separation as the primary separation phenomena with minor creaming, and a unique complex-coacervation showed only by the acidified oat bran products with 100% β-glucanase. The viscosity was only noticeable at intact β-glucan condition (0% β-glucanase). The best curve resolution technique for OBD is the double slicing approach, with discrete exponential fitting performed relative close or even better for only products that showed complex formation. The power slicing model was unable to facilitate the vast range of product characteristics tested and sometimes generated poor non-exponential loadings. The relaxation profiles overall can be explained using a two components system, though some random replicates might be described using three components system. The $T_{2,1}$ varies greatly from 207-1259 ms, while the $T_{2,2}$ ranges from 45-168 ms. The proton concentration range for the $T_{2,1}$ was 72-89%. The $T_{2,1}$ and $T_{2,2}$ profile can be grouped into viscous products (207-486 ms and 45-68 ms), thin products (757-1259 ms and 71-100 ms), and product that form complex coacervates (averaging at 1077, and 168 ms).

P12: Finding new chemometric tools for SERS spectra cluster analysis and predictive modelling

Nicoleta Elena Dina[1], Ana Maria Raluca Gherman[1], Costel Sârbu[2]

1. National Institute for Research & Development of Isotopic and Molecular Technologies, Donat 67-103, 400293 Cluj-Napoca, Romania
2. Faculty of Chemistry and Chemical Engineering, Babeş-Bolyai University, 11 Arany Janos, 400028 Cluj-Napoca, Romania

e-mail: nicoleta.mircescu@itim-cj.ro

Nowadays, spectral acquisition times have been reduced to maximum several minutes for thousands of spectra as a result of technological progress. Thus, performing fast and intelligent data analysis for classification of spectra databases and bacterial spectral fingerprinting is necessary and still a challenge, being an emerging research branch. An innovative detection and molecular diagnosis tool are now embodied in the combination of Raman spectroscopy with nanotechnology, namely **surface-enhanced Raman scattering (SERS)**, assisted by powerful **chemometrics**.

Actually, chemometrics has already been integrated with spectroscopic laboratory and process instrumentation as a standard and proved effective in facilitating spectral data analysis by employing **Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA)** or cluster analysis.

Considering the label-free aspect and high sensitivity, SERS-based detection will represent a more valuable tool when reducing the post-processing steps of data analysis. The SERS spectral profiles are usually examined for changes in intensity of individual bands. However, spectra of biological samples often exhibit variations originating from changes of measurement or cultivation conditions. Such cases make a classification extremely challenging, since a conventional classifier is prone to the intragroup variations and can fail to learn the patterns that help separate different groups. Our recently proposed clustering model based on the **fuzzy set theory** is more robust and can help to comprehensively distinguish bacteria at strain level from single-cell SERS spectra [1-2].

Acknowledgements: This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P1-1.1-PD-2016-0475, within PNCDI III.

References

1. Gherman AMR., Dina NE.., Yeast cell wall – Silver nanoparticles interaction: A synergistic approach between SERS and computational spectroscopy tools. Spectrochim. Acta A 2019, 222, 117223.
2. Dina NE., Gherman AMR. Characterization of Clinically Relevant Fungi via SERS Fingerprinting Assisted by Novel Chemometric Models. Analytical Chemistry 2018, 90, 2484-2492.

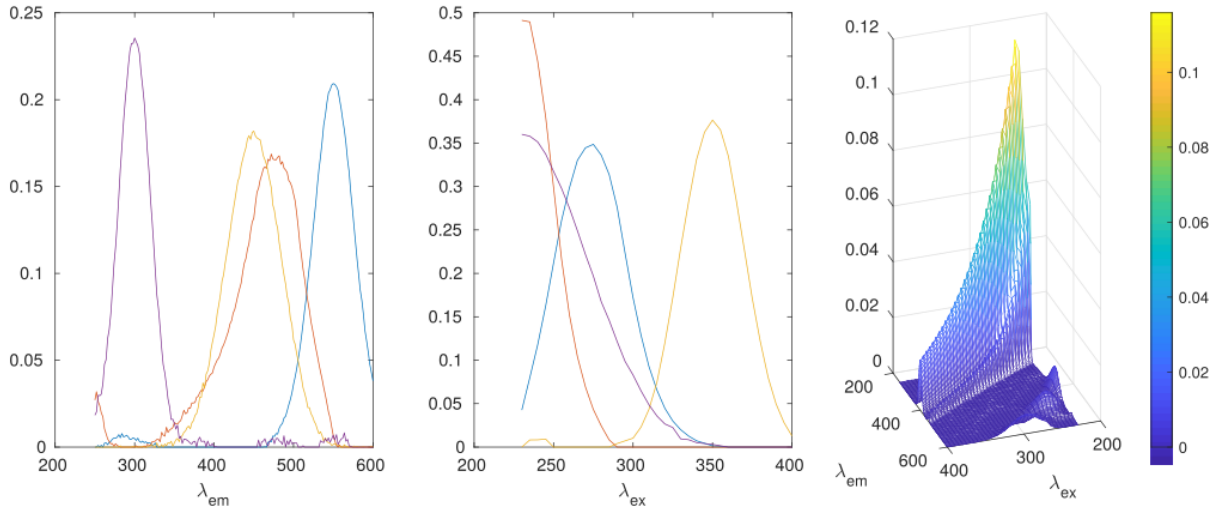
P13: Modelling of scattering signal for direct PARAFAC decompositions of excitation-emission matrices

Ivan Krylov[1], Timur Labutin[1], Åsmund Rinnan[2], Rasmus Bro[2]

1. Department of Chemistry, Lomonosov Moscow State University. Moscow. Russia
2. Department of Food Science, University of Copenhagen, Denmark

e-mail: ikrylov@laser.chem.msu.ru

The PARAFAC decomposition is widely used in the analysis of fluorescence data, owing to its direct correspondence to the underlying physical processes. Unfortunately, excitation-emission matrices (EEMs) typically contain scattering signals in addition to fluorescence, making it a requirement to handle the scattering signal either before or as part of the PARAFAC decomposition. Interpolation of the areas affected by the scattering signal [1] makes it possible to perform PARAFAC decomposition of EEMs while avoiding the local minima, but it may introduce artefacts in the shape of the resulting loadings, or even hide a component if it happens to fully overlap with a second order scattering band.
In this work, an approach based similar to multivariate curve resolution (MCR) with trilinear constraints [2] is applied to the task of modelling both the fluorescence and the scattering signal. On every iteration of the algorithm, PARAFAC and MCR fit each other's residuals, converging towards fluorescence being described by the PARAFAC model and the scattering signal in the MCR model. Typical limitations of this approach are nonlinearities arising from the detector being close to saturation when measuring scattering signal. Constraints fixing the values of the MCR components outside the scattering bands to zeros, use of multiple MCR components, and strategic positioning of missing data may be required to deal with these limitations. The approach has been tested on various EEM datasets, including fully synthetic, amino acid mixtures, sugar process data, and seawater DOM. Example results of the suggested approach – emission and excitation PARAFAC loadings and the scattering signal loadings – are presented in the Figure. We can also conclude that the suggested approach provides better level of precision in reconstruction of original components.



The reported study was funded by RFBR, project number 20-33-90280.

References

1. Bahram M.; Bro R.; Stedmon C.; Afkhami A.; Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation, Journal of Chemometrics. 2006, 20, 99–105.
2. Tauler R.; Marqués I.; Casassas E.; Multivariate curve resolution applied to three-way trilinear data: Study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths, Journal of Chemometrics, 1998, 12, 55–75.

P14: Stochastic optimisation as a straightforward strategy for laser-induced calibration-free breakdown spectroscopy

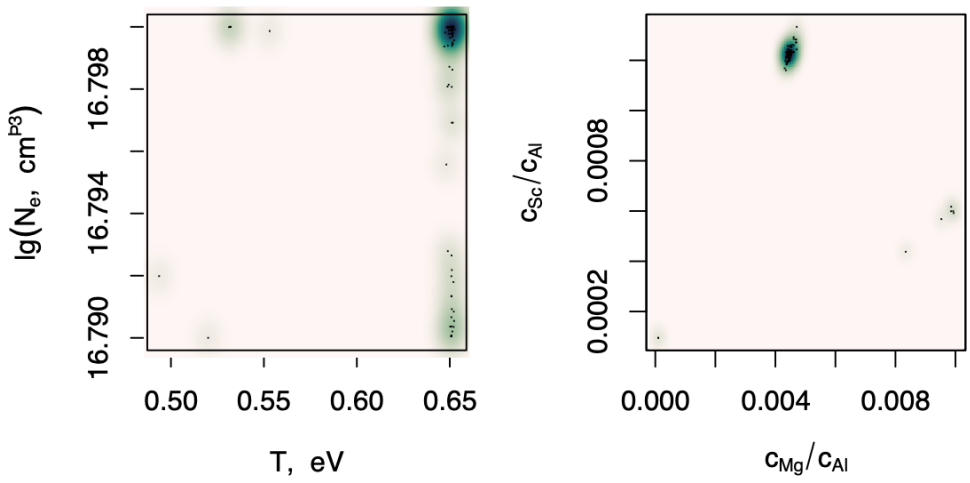
Ivan Krylov[1], Sergey Zaytsev[1], Timur Labutin[1]

1. Department of Chemistry, Lomonosov Moscow State University. Moscow. Russia

e-mail: ikrylov@laser.chem.msu.ru

Laser-induced breakdown spectroscopy (LIBS) is an atomic emission technique which contactless quantifies elemental composition of solids, liquids or gases using the emission of plasma produced by a high-power Q-switched laser on the surface or in a volume of a sample. Thus LIBS can be used in unique applications such as analysis of Mars surface, ocean floor, or process control of molten metal and dross. In most of those cases it is extremely difficult or even impossible to obtain certified reference materials for calibration. Common approaches to calibration-free (CF) LIBS are based on Boltzmann plots and, therefore, tend to impose constraints on the experiment, namely homogeneity of the source, local thermodynamic equilibrium in plasma and absence of self-absorption of involved atomic lines. Fulfilling these requirements thorough experimental optimisation and preliminary studies may prove impossible in space missions or in deep sea research.

We have earlier developed an algorithm based on a stationary model of laser-induced plasma [1] to simulate emission spectra for a given sample composition. In this work, we applied optimisation techniques to fit experimental and model spectra, in which relative elemental concentrations served as variables. For relative simplicity of an implementation, a gradient-free method has been chosen to minimise the sum of squared differences between scaled model and experimental spectra. Various shapes of the loss function are also considered, making it possible to account for analytical lines of very different intensity. We have tested the accuracy by 100 consequent runs of the algorithm with random initialisation (Figure 1). It has been shown that the homogeneous plasma model is suitable for narrow spectral regions, while a multi-zone model with different temperatures and electron densities provides the best results for the full spectral region, although in the latter case the accuracy is lower.



The reported study was funded by Grant of the President of the Russian Federation (No. MK-5513.2021.6).

References

1. Zaytsev S.M.; Popov A.M.; Labutin T.A.; Stationary model of laser-induced plasma: Critical evaluation and applications, Spectrochimica Acta Part B: Atomic Spectroscopy, 2019, 158, 105632.

P15: Interpolation of scattering signal before PARAFAC processing of EEM-fluorescence spectra

Ivan Krylov[1], Timur Labutin[1]

1. Department of Chemistry, Lomonosov Moscow State University. Moscow. Russia

e-mail: : ikrylov@laser.chem.msu.ru

PARAFAC is an attractive data exploration method in fluorescence spectroscopy because its mathematical formulation closely matches the underlying physical model for the fluorescence signal intensity in an excitation-emission matrix (EEM). Unfortunately, EEMs typically contain a scattering component, which doesn't adhere to PARAFAC assumptions and therefore must be taken care of before the analysis. Interpolation of the wavelength region containing scattering signal [1] is typically used, owing to the low possibility of resulting local minima.
This work takes a multitude of surface interpolation methods and compares their performance on model datasets consisting of simulated fluorescence and scattering signals. Compared performance metrics include root-mean-squared error of recovered intensity values and the Tucker's congruence coefficient between loadings estimated by PARAFAC and their a priori values. The methods compared include monotone piecewise bicubic interpolation of each row or column of the EEM, multilevel B-splines, Whittaker interpolation with different combinations of difference orders, the LOESS algorithm and Kriging. The model datasets vary in correlation coefficients between columns of ground truth scores.
The trends in recovery of the original fluorescence signal and in the PARAFAC estimation of the ground truth loadings are basically the same. The results (Figure 1) show that Whittaker interpolation, LOESS and Kriging provide the best results, subject to the need to cross-validate for optimal parameter values (Whittaker & LOESS) and high computational complexity (Kriging).

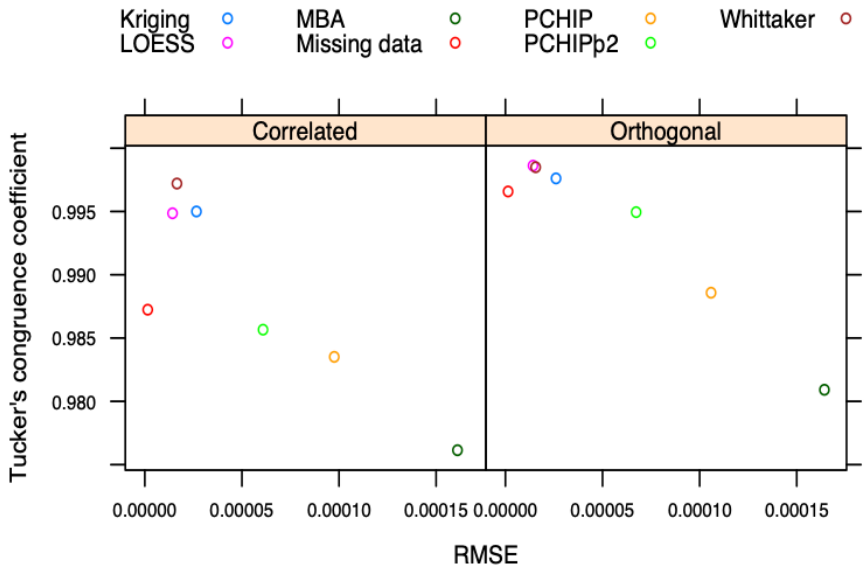


Figure 1. Fluorescence signal reconstruction error (RMSE) and Tucker's congruence coefficient between estimated and ground truth PARAFAC loadings after processing the datasets with various interpolation algorithms.

The reported study was supported by the Russian Science Foundation (project number 21-77-10064).

References

1. Elcoroaristizabal S.; Bro R.; García J.A.; Alonso L.; PARAFAC models of fluorescence data with scattering: A comparative study; Chemometrics and Intelligent Laboratory Systems; 2015, 142, 124–130.