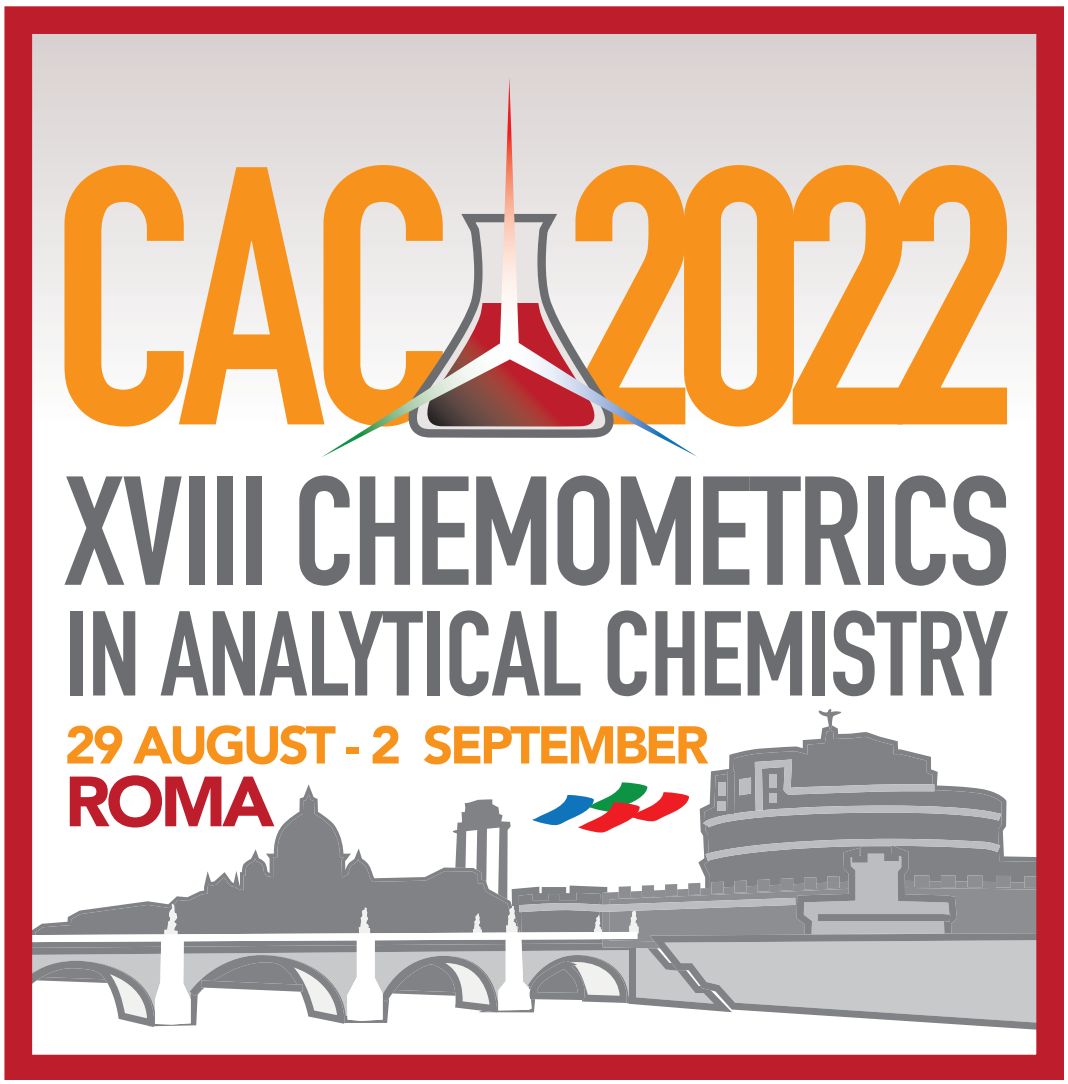


BOOK OF ABSTRACTS

CAC Permanent Committee

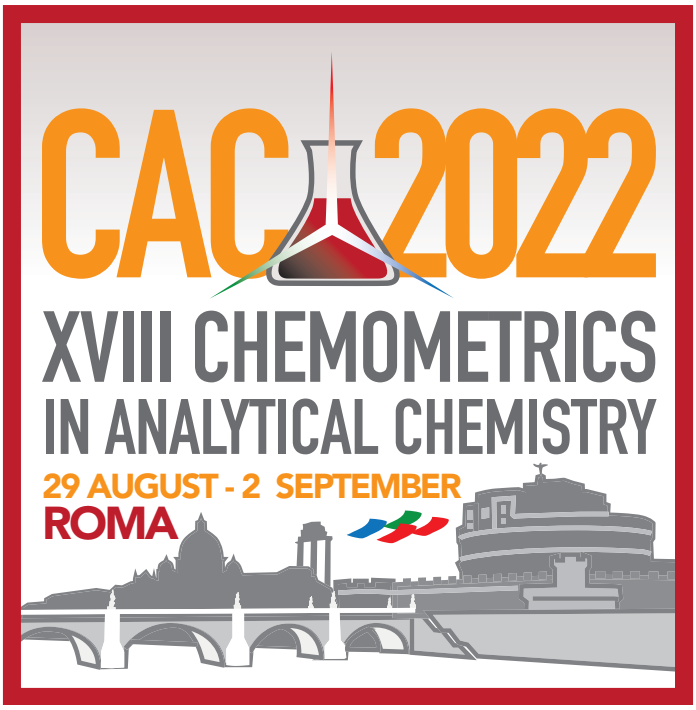
- **Ricard Boqué** *Universitat Rovira i Virgili (Spain)*
- **Lutgarde Buydens** *Radboud University Nijmegen (The Netherlands)*
- **Márcia M.C. Ferreira** *Universidade Estadual de Campinas (Brazil)*
- **Michele Forina** *Retiree University of Genova (Italy)*
- **Neal B. Gallagher** *Eigenvector Research (Inc. (United States)*
- **Philip Hopke** *Clarkson University (United States)*
- **Jose Cardoso Menezes** *Instituto Superior Tecnico (Portugal)*
- **Jean-Michel Roger** *INRAE (France)*
- **Sarah C. Rutan** *Virginia Commonwealth University (United States)*
- **Romà Tauler** *IDAEA – CSIC (Spain)*
- **Bernard Vandeginste** *Retiree Unilever Research and Development (The Netherlands)*
- **Yvan Vander Heyden** *Vrije Universiteit Brussel (Belgium)*
- **Pierre Van Espen** *Universiteit Antwerpen (Belgium)*
- **Peter Wentzell** *Dalhousie University (Canada)*
- **Barry M. Wise** *Eigenvector Research Inc. (United States)*

Scientific Committee

- **José Manuel Amigo** *University of Basque Country (Spain)*
- **Claudia Beleites** *Chemometrix GmbH (Germany)*
- **Anna De Juan** *University of Barcelona (Spain)*
- **Hector Goicoechea** *Universidad Nacional del Litoral (Argentina)*
- **Aoife Gowen** *University College Dublin (Ireland)*
- **Peter Harrington** *Ohio University (USA)*
- **John Kalivas** Idaho State University (USA)
- **Mohsen Kompany-Zareh** *IASBS (Iran)*
- **Hongmei Lu** *Central South University (China)*
- **Marena Manley** *Stellenbosch University (South Africa)*
- **Fernanda Pimentel** *Federal University of Pernambuco (Brazil)*
- **Åsmund Rinnan** *University of Copenhagen (Denmark)*
- **Oxana Rodionova** *Semenov Institute of Chemical Physics (Russia)*
- **Jean-Michel Roger** *INRAE (France)*
- **Serge Rudaz** *University of Geneva (Switzerland)*
- **Francesco Savorani** *Polytechnic University of Turin (Italy)*
- **Age Smilde** *University of Amsterdam (The Netherlands)*
- **Agnieszka Smolinska** *Maastricht University (The Netherlands)*
- **Beata Walczak** *University of Silesia (Poland)*
- **Peter Wentzell** *Dalhousie University (Canada)*

Organizing Committee

- **Federico Marini** *University of Rome La Sapienza (Italy)* Conference Co-Chair
- **Cyril Ruckebusch** *University of Lille (France)* Conference Co-Chair
- **Marina Cocchi** *University of Modena and Reggio Emilia (Italy)*
- **Ludovic Duponchel** *University of Lille (France)*
- **Raffaele Vitale** *University of Lille (France)*
- **Alessandra Biancolillo** *University of L'Aquila (Italy)*
- **Olivier Devos** *University of Lille (France)*
- **Paolo Oliveri** *University of Genova (Italy)*



CONFERENCE PROGRAM

PRELIMINARY PROGRAM

Monday 29 August 2022

10:00-19:00 **Registration of the participants**

14:30-15:00 **Conference opening**
15:00-16:00 **PL1: Harald Martens - Human-interpretable Machine Learning with an Eye for Causalities: *Making sense of modern measurement streams inside and outside chemistry***
16:00-16:30 **Coffee break & Poster session**
16:30-18:10 **Contributed Session I: Big Data and Machine Learning**
16:30-16:50 OL1: Davide Ballabio - Enhancing LC-MS/MS Spectral Searching With Multi-Task Neural Networks And Molecular Fingerprints
16:50-17:10 OL2: Gabriel Vivo-Truyols - On the use of Bayesian statistics for (big) data analysis: automation for both qualitative and quantitative chemometrics
17:10-17:30 OL3: Michael Sorochan Armstrong - Chemometrics with Amazon Web Services (AWS)
17:30-17:50 OL4: Jeroen Jansen - Process economy, efficiency and sustainability go hand in hand, how chemometrics can build a greener industry
17:50-18:10 OL5: Priyanka Kumari - QSRR for small pharmaceutical compounds in RPLC: A Machine learning approach
18:10-20:10 **Welcoming cocktail**

Tuesday 30 August 2022

09:00-10:00 **PL2: Roy Goodacre - Lessons from large-scale metabolic phenotyping**
10:00-11:00 **Contributed Session II: Omics/ASCA and related methods I**
10:00-10:20 OL6: Albert Menéndez Pedriza - Comparison of mid-level fusion strategies for the multi-omic analysis of toxicological data
10:20-10:40 OL7: Andrés Martínez Bilesio - Metabolomics-guided insights on Bariatric Surgery: a longitudinal chemometrics approach over 1H NMR spectra from serum samples
10:40-11:00 OL8: Sarah Malek - Evaluation of mid-infrared spectra of serum and synovial fluid in predicting early post-traumatic osteoarthritis in an equine model
11:00-11:30 **Coffee break & Poster session**
11:30-13:10 **Contributed Session III: Applications I**
11:30-11:50 OL9: Zuzana Małyjurek - Class-Modelling - Optimization, Validation and Application
11:50-12:10 OL10: Agnieszka Martyna - Likelihood ratio in forensic discrimination/classification tasks
12:10-12:30 OL11: Martina Foschi - Supervised and Unsupervised Chemometric Methods to deal with saffron aging and its Quality Control
12:30-12:50 OL12: Lorenzo Strani - Real time prediction of ABS properties through multiblock and local regression methods
12:50-13:10 OL13: Sebastian Orth - Spectral imaging - pre-harvest malting barley germination classification with sequential orthogonalised multiblock data fusion methodologies
13:10-14:30 **Lunch & Poster session**

14:30-15:00 **KN1: Mathias Sawall - On the ambiguity underlying the spectral recovery problem and its analysis by the area of feasible solutions**
15:00-16:00 **Contributed Session IV: Theory & algorithms I**
15:00-15:20 OL14: Sergey Kucheryavskiy - Procrustes cross-validation of multivariate regression models
15:20-15:40 OL15: Stephan Seifert - Opening the random forest black box with Surrogate Minimal Depth
15:40-16:00 OL16: Oxana Rodionova - Expansion of the DD-SIMCA concept
16:00-16:30 **Coffee break & Poster session**
16:30-18:10 **Contributed Session V: Multi-block/Multi-way/Multi-set I**
16:30-16:50 OL17: Paul-Albert Schneide - Speeding up PARAFAC2 dramatically
16:50-17:10 OL18: Isabelle Viegas - Coupled factorization of fluorescence data of proteins and quantum dots to assess their conjugation process
17:10-17:30 OL19: Oksana Mykhalevych - New tools for designing food ingredients structures
17:30-17:50 OL20: Maria Cairoli - Monitoring pollution pathways in river water by predictive path modelling using untargeted GC-MS measurements
17:50-18:10 OL21: Ivan Krylov - Fluorescence and scattering model estimation

Wednesday 31 August 2022

09:00-10:00 **PL3: Ingrid Måge - Industrial bioprocessing – an amusement park for chemometricians and analytical chemists**
10:00-11:00 **Contributed Session VI: Multi-block/Multi-way/Multi-set II**
10:00-10:20 OL22: Jean-Michel Roger - N-CovSel, a new strategy for feature selection in N-way data
10:20-10:40 OL23: Mahdiyeh Ghaffari - Using Multi-Block Non-Negative Matrix Factorization for Multi-layer Plastic Sorting
10:40-11:00 OL24: Paul Gemperline - Combining ASCA and Tucker3 models to explain high-dimensional data
11:00-11:30 **Coffee break & Poster session**
11:30-13:10 **Contributed Session VII: Applications II**
11:30-11:50 OL25: Sabina Licen - Data fusion based on self-organizing map algorithm for the integration of different source/frequency instrumental data and spot sampling contextualization for environmental monitoring
11:50-12:10 OL26: Vicky Caponigro - Application of different chemometric approaches for MALDI-MSI data set of heterogeneous tissues. Case study: parotid tumour
12:10-12:30 OL27: Ewa Szymanska - Comprehensive chemometric strategy for the high-throughput screening of in-line spectroscopic sensors for milk composition traits
12:30-12:50 OL28: Maxime Ryckewaert - Combining hyperspectral imaging data with climate data to predict physiological variables of grapevine plants
12:50-13:10 OL29: Eleni Ioannidi - Using ATR FT-IR and MCR as a method to understand the crystal state of chocolates tempered under different conditions
13:10-14:30 **Lunch & Poster session**

14:30-15:00 **KN2: Hadi Parastar - Integration of handheld spectrometers and chemometrics for food authentication**
15:00-16:00 **Contributed Session VIII: Spectroscopy & Imaging I**
15:00-15:20 OL30: Cristina Malegori - HSI-NIR and chemometrics for the quantification of collagen in bones: how chemical mapping can help in preserving archeological finds
15:20-15:40 OL31: Manuela Mancini - Spectroscopy and chemometrics for sorting waste wood material according to the best-suited application
15:40-20:30 **Social activity (tours will start from the agreed meeting points at 17:00)**

Thursday 1 September 2022

09:00-10:00 **PL4: Romà Tauler - Bilinear model factor decomposition: a general mixture analysis tool**
10:00-11:00 **Contributed Session IX: Curve Resolution**
10:00-10:20 OL32: Martina Beese - An active constraint approach to identify essential spectral information in noisy data
10:20-10:40 OL33: Laureen Coic - A phasor view of Multivariate Curve Resolution
10:40-11:00 OL34: Anna De Juan - Trilinearity in Multivariate Curve Resolution: hybrid modeling and missing data
11:00-11:30 **Coffee break & Poster session**
11:30-13:10 **Contributed Session X: Spectroscopy & Imaging II**
11:30-11:50 OL35: Florent Abdelghafour - Combining spectral and spatial features extracted from hyperspectral images: Application on the detection of scab disease
11:50-12:10 OL36: Rodrigo Rocha de Oliveira - 2-D wavelet image decomposition and Multivariate Statistical Process Control for blending end-point detection
12:10-12:30 OL37: Valeria Tafintseva - Modelling and preprocessing of sparse infrared spectra
12:30-12:50 OL38: Nicola Cavallini - Tracing the identity of mountain product Parmigiano Reggiano PDO cheese using 1H-NMR spectroscopy and multivariate data analysis
12:50-13:10 OL39: Paolo Oliveri - A combined chemometric strategy for a non-destructive age estimation of biological fluid stains
13:10-14:30 **Lunch & Poster session**
14:30-15:00 **KN3: Maria Cruz Ortiz - Analytical Quality by Design using a computational approach for the inversion of a PLS model**
15:00-16:00 **Contributed Session XI: Omics/ASCA and related methods II**
15:00-15:20 OL40: Miguel De Figuereido - Rebalanced ASCA (RASCA) to handle unbalanced multifactorial designs
15:20-15:40 OL41: Michel Thiel - LMWiRe: an R package for Linear Modeling of Wide Responses based on ASCA family of methods
15:40-16:00 OL42: Claudia Beleites - An Experimental Design Perspective on Cross-Validation
16:00-16:30 **Coffee break & Poster session**
16:30-17:50 **Contributed Session XII: Spectroscopy & Imaging III**
16:30-16:50 OL43: Siewert Hugelier - Quantifying the Tau protein aggregation degradation process by classification of super-resolution fluorescence microscopy localizations
16:50-17:10 OL44: Erik Tengstrand - Calibration transfer of Near-Infrared and Raman models without using transfer samples

17:10-17:30 OL45: Alisa Rudnitskaya - Characterization of microplastics from marine organisms using near infrared hyperspectral imaging
17:30-17:50 OL46: Jose Luis Aleixandre-Tudo - Spectral evaluation of fresh grapevine organs using self-organizing maps (SOM)
17:50-18:30 Awards Ceremony (Elsevier Chemometrics and Intelligent Laboratory Systems Award & Lifetime Achiement Award)
20:15-01:00 Social dinner

Friday 2 September 2022

09:00-11:00 Contributed Session XIII: Theory & algorithms II
09:00-09:20 OL47: Nematollah Omidikia - Infrared Ion Spectroscopy Peak Matching using Peak Annotation Technique
09:20-09:40 OL48: Sergio Oller Moreno - Peak matching across Gas Chromatography-Ion Mobility Spectrometry samples
09:40-10:00 OL49: Wouter Saeys - Multivariate monitoring and update strategies for calibration models
10:00-10:20 OL50: Sean Rozinski - What's UMAP Doing Anyway?
10:20-10:40 OL51: Ramin Nikzad-Langerodi - Does it Transfer? Assessing model generalization in domain adaptation with data fusion
10:40-11:00 OL52: Benjamin Mahieu - New developments around the VIP index
11:00-11:30 Coffee break & Poster session
11:30-12:50 Contributed Session XIV: Applications III
11:30-11:50 OL53: Dmitry Kirsanov - Chemometrics in spent nuclear fuel reprocessing
11:50-12:10 OL54: Joscha Christmann - Monitoring of fermentation processes by gas chromatography-ion mobility spectrometry (GC-IMS)
12:10-12:30 OL55: Martín Bravo - Development of an analytical platform for the identification of Fusarium circinatum in culture media, using VIS-NIR spectroscopy and chemometric methods
12:30-12:50 OL56: Tim Offermans - Retrospective Quality by Design r(QbD) using Historical Process Data and Design of Experiments
12:50-13:30 Conference Closing

POSTER LIST

P1: Fernanda Honorato - Authenticity of almond flour using handheld near infrared instruments and one class classifiers

P2: Florent Abdelghafour - Unsupervised calibration transfer between spectrometer and hyperspectral camera: challenge proposed at the congress “Chimiométrie 2022”

P3: Riccardo Aigotti - Odor concentration predictive model based on the odor activities of odorants produced by a municipal solid waste odor abatement scrubber

P4: Ricard Boqué - ATR-MIR and MCR-ALS as a tool for monitoring wine alcoholic fermentation and detecting bacterial spoilage

P5: Ricard Boqué - Prediction of beer shelf life using an HS-MS e-nose

P6: Nicola Cavallini - The NIR side of lentil

P7: Alessandro D'Alessandro - Exploiting pesto sauce by several analytical platforms: looking for most efficient information extraction and data fusion approach

P8: Tiziana Forleo - Application of chemometric approaches to answer some archeological questions for the study of the Apulian Red-Figure Pottery

P9: Gianmarco Gabrieli - Leveraging an integrated sensor array and machine learning to accelerate sensory evalution of coffee

P10: Barbara Giussani - Insights into multivariate data analysis for real-case fermentation process with miniaturized NIR spectroscopy

P11: Klaudia Glowacz - Identification of metal ions with the use of quantum dots coupled with excitation-emission matrix fluorescence spectroscopy

P12: Jule Hansen - Evaluation of preprocessing strategies for LCMS data using R

P13: Christel Kamp - Spectral identification of therapeutic allergen products

P14: Nicholas Kassouf - Comparison between colloidal and volatile profiles to create a chemometric model to classify different tomato sauce brands

P15: Nicholas Kassouf - Multivariate analysis of colloidal and volatile profiles for class-modeling of different tomato sauce brands

P16: Victor Cardoso - A comparison between artificial neural networks and partial least squares for coffee assessment by high-resolution mass spectrometry

P17: Erwin Kupczyk - Benchmarking Machine Learning approaches for hit detection in High-Content Screening

P18: Qicheng Wu - Robust quantitative analysis in Laser-Induced Breakdown Spectroscopy (LIBS) using artificial neural networks

P19: Miguel De Figueiredo - Analyzing multifactorial designed data from multiple sources with a single model using AComDim

P20: Giulia Gorla - Investigating sources of variance in miniaturized NIR spetroscopy: find clues and solve the riddle

P21: Luis A. Sarabia - Logical analysis of the sample pooling results for qualitative analytical testing: a proof-of-concept study

P22: Daniel Schorn-García - Acetic or lactic bacteria contamination? ASCA has the answer

P23: María Julia Culzoni - A fluorometric photo-induced four-way calibration method for the determination of multiclass pesticides in citrus fruits

P24: María Julia Culzoni - Chemometrically assisted high-throughput methotrexate sensing strategy based on a pH-switchable optical nanosensor

P25: Hector Goicoechea - Multiway data modeling for enhancing classification performance: fluorescence data as case of study

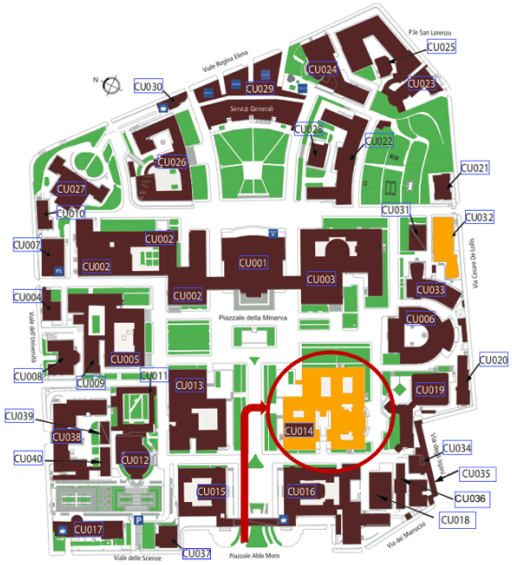
P26: Andrés Martínez Bilesio - Data fusion approach applied in chemometrics-assisted metabolomics analysis
P27: Nicola Cavallini - Mapping Chemometrics with Chemometrics
P28: Isabelle Viegas - Joint factorization of right-angle and front-face fluorescence data to improve PARAFAC pure profiles recovered from oil-in-water emulsions
P29: Marc Marín García - Multivariate Curve Resolution of incomplete and partly multilinear multi-block data sets
P30: Tobias Karakach - Low signal intensity, measurement errors and biological significance: a model for LC-MS proteomics
P31: Reza Nafari - Quantitative evaluation of red meats in kebab loghmeh samples: fourier transform infrared data and chemometric methods
P32: Justine Raeber - Fast and Convenient Authenticity Control of Natural Products using Mass Spectrometry and Chemometrics
P33: Anastasiia Surkova - Aquaphotomics study of body fluids in cancer research
P34: Soeren Wenck - Opening the Random Forest Black Box of the Asparagus Metabolome
P35: Elianna Castillo - Relationship between cadmium availability and soil properties in cacao farms at Santander – Colombia
P36: Abdelaziz Ait Sidi Mou - Application of multivariate data analysis coupled with spectroscopy to agroalimentaire investigation in Morocco: advancement and challenge
P37: Matthias Rüdt - Chemometrics – a chemometric Python package
P38: Rustam Guliev - Structuring and generalizing implementations of N-FINDR algorithm for unmixing hyperspectral data
P39: Erik Johansson - Variable removal by logical blocks in OPLS predictions
P40: John Kalivas - Rashomon effect and model interpretabiltiy: is it possible?
P41: Lyle Lawrence - Diagnostic Plots to Aid Final Model Selection
P42: Mansuk Oh - Bayesian Multivariate Receptor Modeling Software: BNFA and bayesMRM
P43: Maria Sagrario Sánchez - Compliant class-models based on PLS2 to handle several categories encoded with error correcting output codes
P44: Patrícia Valderrama - Are we there yet? efficient exploration and visualization of multivariate data with SCORXPLOR
P45: Macarena Rojas - Chemical variation of sugar beet subjected to long-term storage by Vis-NIR spectroscopy, Hyperspectral Imaging and chemometric methods
P46: Francesco Savorani - The NMR side of lentil: protein extraction and hydrolyzation, and a bit of data fusion
P47: Marek Sikorski - Explorative study of strawberry juice from various fruit varieties using absorbance-transmission and fluorescence excitation-emission matrix technique
P48: Sin Yong Teng - Chemsy: Simultaneous feature selection, pre-processing search, model selection, and hyper-parameter optimization in Python
P49: Sonia Nieto Ortega - Reliable determination of the lipidic profile of oils extracted from fish by-products through near infrared spectroscopy and chemometrics
P50: Claudete Pereira - A multivariate approach to quantify the enhancement effect on surface-enhanced spectroscopies
P51: Beatriz Quintanilla-Casas - Virgin olive oil excitation-emission matrices: exploring their usefulness to predict taste attributes
P52: Antonino Restivo - Multivariate Data Analysis and PAT in vaccines development: enabling multiple components quantification in complex formulations

P53: Elisa Robotti - Optimization of the parameters of a continuous annealing process in a steel producing company by multivariate statistics and Artificial Neural Networks
P54: Laura Rolinger - Blend uniformity design space development and verification by PAT for minibatch blending
P55: Carolina Silva - Application of class-modelling approaches for botanical and geographical origins of honey samples based on mineral content
P56: Giacomo Squeo - Application of DoE and multivariate analysis for TXRF method development and data analysis. A case-study from the agri-food sector.
P57: Mauro Tomassetti - A new survey for multicomponent analysis to solve problems linked to nano-compounds (case study)
P58: Berta Torres - Discriminant classification models applied to hazelnut unsaponifiable fingerprint for geographical and varietal authentication
P59: Patrícia Valderrama - Multivariate control chart based on PCA/Q residuals to evaluate Salmonella in meat-bone flour
P60: Helene Halberg - Fluorescence spectroscopy of wine, a complex food system
P61: Daniele Tanzilli - IMAGINE NIR to monitor Pesto sauce industrial production
P62: Lucas F. Voges - Genotyping and statistical analysis of marzipan with DMAS-PCR
P63: Andrea Junior Carnoli - Alternative approaches to untargeted LC/GC-MS data analysis
P64: Cannon Giglio - Analysis of Pinot Noir Wines Using UV-Vis Spectroscopy
P65: Milan Chhaganlal - Evaluation of the accuracy of NMR predictors for the prediction of fatty acid spectra
P66: Mohamad Ahmad - An IDEL perspective on handling spatial correlation in hyperspectral imaging
P67: Juan Araya - Identification of spectral patterns associated to different aggregation states of beta amyloid peptide in hyperspectral images through chemometric analysis
P68: Juan Araya - Supervised pattern recognition using near infrared spectrum of serum for diagnosis of gestational diabetes mellitus
P69: Issam Barra - Soil spectroscopy: use of chemometrics for fine-tuning spectra acquisition- case of scans number optimization
P70: Katharina Beier - Classification of Horsetails using Machine Learning Methods on NIR Spectra
P71: Irati Berasarte - Time-based colorimetric method for the simultaneous determination of calcium and magnesium ions with silver nanoparticles
P72: Hooriyeh Borhani - Investigation of an innovative method for classifying nanostructures based on time series analysis and fuzzy logic in microscopic images
P73: Ewa Sikorska - Multivariate models for prediction quality parametrs of berry beverages using FTIR-ATR spectroscopy
P74: Jokin Ezenarro - Olive ripening assessment methodologies using digital image analysis
P75: Davide Gattabria - An exploratory study on monitoring tomato plant growth by near infrared portable devices
P76: Hector Goicoechea - Chemometric approaches to enhance the potential of new IR spectroscopic technologies
P77: Hector Goicoechea - Feasibility of MCR-ALS to exploit the second-order advantage with first-order and non-bilinear second-order data. a systematic characterization
P78: Ivan Krylov - Approximation of Martian rock emission spectra by multiparametric optimization

P79: Saeedeh Mohammadi Tanouraghaj - An assessment of the potential of different vibrational spectroscopic techniques in classification of various types of liquid milk by using multivariate chemometric methods
P80: Arsenio Muñoz De La Peña - Discriminant analysis of three and four-way fluorescence data for classification issues
P81: Alessandra Olarini - Hyperspectral imaging data: clustering or spectral unmixing?
P82: Nicholas Pedge - Update of Transmission Raman Spectroscopy Calibration Models using Dynamic Orthogonal Projection (DOP)
P83: Jordi Riu - Classification of bitter and sweet almonds using NIR miniaturized instruments
P84: Mohamad Ahmad - A solution based on sample weighting to the leverage problem in Multivariate Curve Resolution-Alternating Least Squares
P85: Gorka Albizu - Different chemometric strategies to control PTFE in Ni-P/PTFE electroless coating baths by UV-VIS
P86: Tomass Andersons - Pure component recovery for rank-deficient problems
P87: Cristian Fuentes - Application of a segmented analysis by MCR-ALS on 1H-NMR spectroscopy for the identification of adulterations in brown sugars
P88: Adrián Gómez-Sánchez - Unmixing exponential signals by Kernelizing
P89: Jan Hellwig - Multi-Layer modeling of time series of NMR spectra
P90: Nunzia Iaccarino - Exploring the dynamic equilibria of non-canonical DNA structures by Multivariate Curve Resolution and 2D correlation spectroscopy
P91: Paulo Henrique Março - Pseudo-univariate calibration through MCR-ALS applied to electrochemical data to determine different amino acids simultaneously
P92: Nematollah Omidikia - On the Visualization of Bayesian Nonnegative Factor Analysis
P93: Nazanin Saburouhvahid - Application of PARAFAC for curve resolution of fluorescence lifetime imaging data
P94: Aina Queral Beltran - UV absorption spectrophotometry and LC-DAD-MS coupled to chemometrics analysis of the degradation of sulfamethoxazole drug by UV/chlorine advanced oxidation processes
P95: Carlos Pérez López - The potential of the ROIMCR methodology for sewage water sample characterization in environmental proteomics
P96: Eugenio Sandrucci - Monitoring the State of Health (SOH) of green batteries (GreenBat)
P97: Claudia Scappaticci – SIMCA framework for multi-block class modeling
P98: Alessandra Biancolillo - ICP-OES analysis coupled with chemometrics for the characterization and the discrimination of high added value Italian Emmer samples
P99: Juan Araya - Gestational diabetes mellitus, preterm birth and macrosomia early prediction using multivariate analysis on clinical and biochemical data
P100: Michel Rocha Baqueta – Chemometrics-assisted microNIR spectroscopy for large-scale classification and authentication of high-quality Brazilian Canephora coffees

USEFUL INFORMATION

All the conference activities (except for the social program) will take place in the Cannizzaro Building (CU014) of the Chemistry Department of Sapienza University of Rome, which is located in the main University campus (main entrance is in Piazzale Aldo Moro 5). You can see how to access the Cannizzaro building from the main entrance in the map below:



The main lecture hall is **Aula La Ginestra**, which is located on the first floor of the building (signs will guide you through the access), while the poster boards will be mounted in the space outside the main entrance of the Chemistry building, where coffee breaks and lunches will also be served.

Registration

The registration desk is located in the main entrance hall of the Chemistry building and will be open throughout the conference.

A cash point is available next to the registration desk for on-site registrations, refunds and/or purchase of optional tours and additional conference dinner tickets.

When signing in at the registration desk, you will receive your conference badge and a conference bag with the conference material. Please make sure to wear your badge at all times while attending the conference.

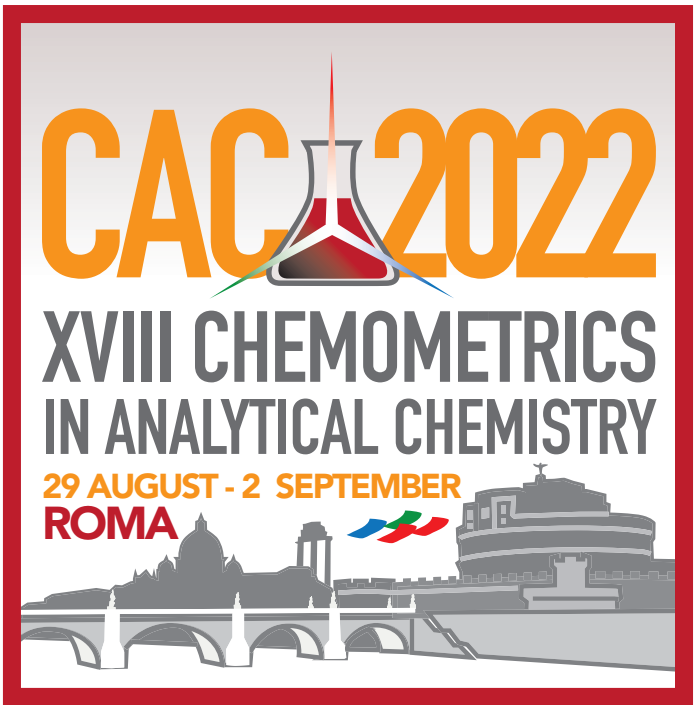
Instructions for speakers

Presenting authors are encouraged to upload their presentations (directly on the computer located in Aula La Ginestra) as early as possible and, anyway, not later than the coffee break/lunch interval before their session.

Virtual Special Issue

Analytica Chimica Acta and Chemometrics and Intelligent Laboratory Systems journals are asking for the submission of review and research papers to the virtual special issue (VSI) dedicated to the subject: *Chemometrics: Intelligent data analysis for Analytical Chemistry*.

Accepted papers will be considered for publication in either one of the two journals. Details about the submission procedure to this VSI will be sent to all CAC2022 participants just after the conference has finished.



PLENARY LECTURES

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

HUMAN-INTERPRETABLE MACHINE LEARNING WITH AN EYE FOR CAUSALITIES: Making sense of modern measurement streams inside and outside chemistry

H. Martens

Dept. Engineering Cybernetics, Norwegian U. of Science and Technology NTNU, Trondheim Norway / Idletechs AS (www.idletechs.com) / NatMat AS (motto: "Nature first, then Mathematics");
harald.martens@idletechs.com

To what extent can our ways of working in the field of chemometrics improve on today's popular, flexible and powerful but demanding, dangerous and alienating black box AI / deep learning? And what will tomorrow's measuring devices look like?

Personally, I think the chemometric culture have some good thought models and tools that are great for Explainable AI for high-dimensional technical Big Data, inside and outside chemistry. And I believe that multichannel "video-camera" systems of various sorts will become smaller and cheaper. They generate multichannel spatiotemporal streams of Quantitative Big Data, well suited for our pragmatic and relatively simple data modelling in chemometrics.

In this lecture I shall demonstrate how such relative precise, but overwhelming, non-selective Quantitative Big Data may be converted into compact streams of useful and understandable quantifications of various kinds. This extraction of interpretable information is based on multivariate hybrid modelling, using methods and tools developed in chemometrics, cybernetics and other pragmatic cultures of applied data analysis, and involves also the use of domain-specific knowledge.

It is no surprise that good technical measurements are useful – after all, the variations in such data reflect the laws of nature, whether we understand them or not. But in "machine learning" it is not cost-effective to rely ONLY on empirical training data, ignoring what is known already. For that requires ALL potentially important variation types to be spanned empirically. It is e.g. unwise to demand that the measurements should involve highly toxic constituents whose spectra are already well known, or near-to explosive states whose reactions are already well modelled. Our low-dimensional hybrid modelling gives a knowledge-based, rational but open-ended and interpretable bridge between incomplete mechanistic models in traditional academia and overly complex neural nets in Deep Learning from computer science.

In this lecture, chemometric ideas and methods will be applied to streams of Quantitative Big Data from spatiotemporal "video" cameras. This is done within a hybrid modelling framework that combines knowledge-driven and data-driven subspace modelling. The goal is to provide "Human-interpretable Machine Learning with an Eye for Causalities": Modelbased preprocessing (EMSC) involving known spectra and/or multivariate metamodeling will be used to resolve various known chemical and physical phenomena. Unknown variation patterns are discovered by PCA etc. Image analysis, e.g. IDLE-modelling, will reveal the spatial shape and motion of these known and unknown phenomena. And time series analysis will detect and describe the kinetics of how these phenomena change in level and shape relative to the camera.

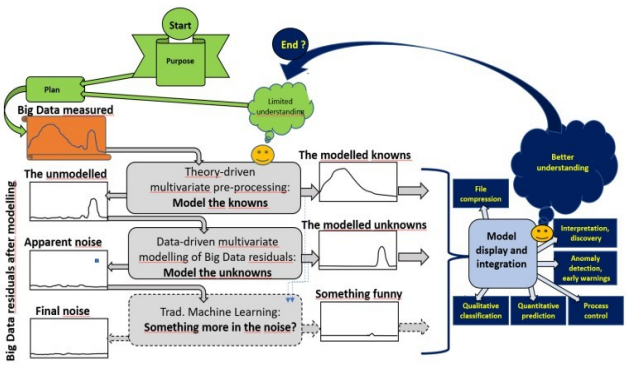


Figure 1 – A hybrid chemometric approach to Interpretable Machine Learning for multi-channel measurements

The lecture will illustrate this with two instrument types:

• Vis/NIR hyperspectral "video" to monitor the complex process of drying wood, identifying changes in water state, light scattering and physical shrinkage.

• Thermal video to monitor technical machinery, inferring that is going on inside the machinery.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

LESSONS FROM LARGE-SCALE METABOLIC PHENOTYPING

R. Goodacre
[1]*University of Liverpool, Liverpool, UK*
roy.goodacre@liverpool.ac.uk

Metabolomics is a growing discipline that allows the analysis of the thousands of structural different small molecules found within a biological system. These metabolites can be measured using a variety of different analytical approaches and we have developed gas chromatography mass spectrometry (GC-MS) and liquid chromatography mass spectrometry (LC-MS) for this purpose [https://doi.org/10.1038/nprot.2011.335].

I shall provide an overview of metabolomics and lessons learnt from of our large-scale human serum metabolome project where we profiled 1200 healthy individuals [https://doi.org/10.1007/s11306-014-0707-1]. Using these protocols we then went on to profile another ~1200 ageing individuals and identified key metabolic dysregulation which were drivers behind human frailty, which were validated in a further ~760 ageing individuals [https://doi.org/10.1038/s41467-019-12716-2].

.

INDUSTRIAL BIOPROCESSING – AN AMUSEMENT PARK FORCHEMOMETRICIANS AND ANALYTICAL CHEMISTS

I. Måge
NOFIMA, Ås, Norway
Ingrid.Mage@nofima.no

The food industry invests heavily in bioprocesses that can transform by-products from meat and fish, such as carcasses, heads, and skin, into nutritionally valuable proteins. One such process is enzymatic protein hydrolysis, where enzymes break down proteins into smaller peptides and amino acids. The resulting protein ingredients have a range of potential applications, from feed to food ingredients, nutraceuticals, and even pharmaceuticals.

There are several challenges on the way from by-products to commercially viable products, related to raw material and product characterization, process and product optimization, and consumer acceptance and marketing. My colleagues and I have worked closely with this industry for more than ten years, and it has turned out to be an “amusement park” for us with a wide selection of scientific rides and attractions. In this talk I will present a selection of these, showing how chemometrics and analytical chemistry is the key to solving real industrial challenges.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

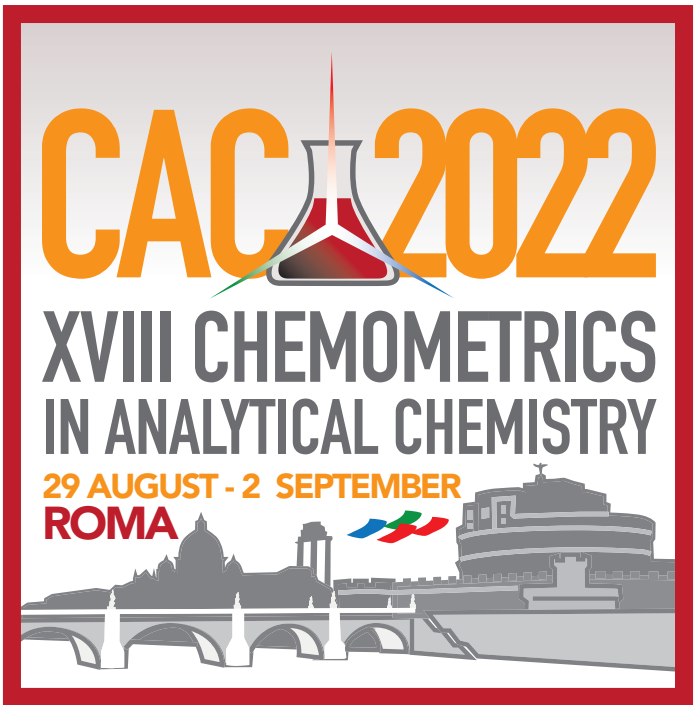
Bilinear model factor decomposition: a general mixture analysis tool

N. Omidikia[1], M. Ghaffari[1], J. Jansen[1], L. Buydens[1] and R.Tauler[1,2]
1 Department of Analytical Chemistry, Institute for Molecules and Materials (IMM), Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
2 IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, Spain
Roma.Tauler@idaea.csic.es

Mixture analysis is a very general and ubiquitous problem encountered in many research areas and applied fields, and its solution has been addressed within different frameworks. The goal of mixture analysis is to translate the measured raw data into physico-chemically meaningful profiles from their unknown sources combined in an undetermined way. In many circumstances a possible way to solve this very general problem is by means of a bilinear model factor decomposition of a data table (or data matrix) having the experimentally measured multivariate data under a set of conditions and constraints. This situation is frequently encountered in Analytical Chemistry, where the measured mixture signals should be detangled in their chemical constituents to know their identity, composition and apportion, and often also with the goal of their chemical interpretation. Different bilinear factor decomposition strategies have been proposed such as Multivariate Curve Resolution Alternating Least Squares (MCR-ALS [1]), Non-Negative Factor Decomposition (NMF, [2]), Positive Factor Decomposition (PMF [3]) and Bayesian Non-Negative Factor Analysis (BNFA [4]) These methods use different type of optimization algorithms to generate the solutions of the mixture analysis parameters of the bilinear model and differ also in the constraints implementation. This is an essential aspect to reduce the uncertainties associated with the bilinear model decomposition. Although these decomposition algorithms have been frequently addressed in the literature in different contexts, there is little knowledge about their comparison. This comparison should consider different aspects, such as the speed of the calculations and the convergence of the algorithms, the initialization effect, the flexibility in the constraints implementation, and the easiness of results interpretation, including the possibility to estimate the uncertainties associated to them. In this work, these different algorithms will be compared and tested using different data examples related with environmental source apportionment studies and with chromatographic analysis of chemical mixtures.

References

[1] Tauler R. Multivariate curve resolution applied to second order data, , Chemomet. and Intell. Lab. Syst., 1995, 30, 133-146; 18.
[2] Lee, D.D. and H.S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature, 1999. 401, 6755, 788-791
[3] Paatero, P., Tapper, U., Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 1994, 5, 111–126
[4] Sug Park E., Kyung Lee E., Suk Oh M., Bayesian multivariate receptor modeling software: BNFA and bayesMRM, Chemometrics and Intelligent Laboratory Systems 2021, 11, 104280



KEYNOTE LECTURES

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

ON THE AMBIGUITY UNDERLYING THE SPECTRAL RECOVERY PROBLEM AND ITS ANALYSIS BY THE AREA OF FEASIBLE SOLUTIONS

M. Sawall[1]
[1]*University of Rostock, Rostock, Germany*
mathias.sawall@uni-rostock.de

Spectral recovery is a blind source separation problem. Observing a chemical reaction system or a dynamic multi component system with modern spectrometers usually results in a large number of high-resolution spectra containing overlapping signals of all absorbing species. Multivariate curve resolution (MCR) techniques can help to uncover the pure component information from the series of multicomponent spectral data. According to the Lambert-Beer law, the data matrix can be factorized into matrices of pure component spectra and their associated concentration profiles. Typically there is not only one factorization, but a continuum. Potentially pure component profiles can be represented by the area of feasible solutions (AFS) in a low-dimensional manner in U- resp. V-space.

The talk focuses on the analysis of the ambiguity of factorizations as well as the concepts of the low-dimensional representations of profiles and the AFS. Inner and outer polytopes and their limiting character are discussed. We introduce duality relations between points in U-space and hyperplanes in V-space and vice versa. Special attention is on the handling of experimental data. For noisy data at least one restriction of the nonnegative factorization problem must be weakened to allow meaningful results. This means that either the data is no longer reconstructable as error-free as possible or the profiles contain negative entries.

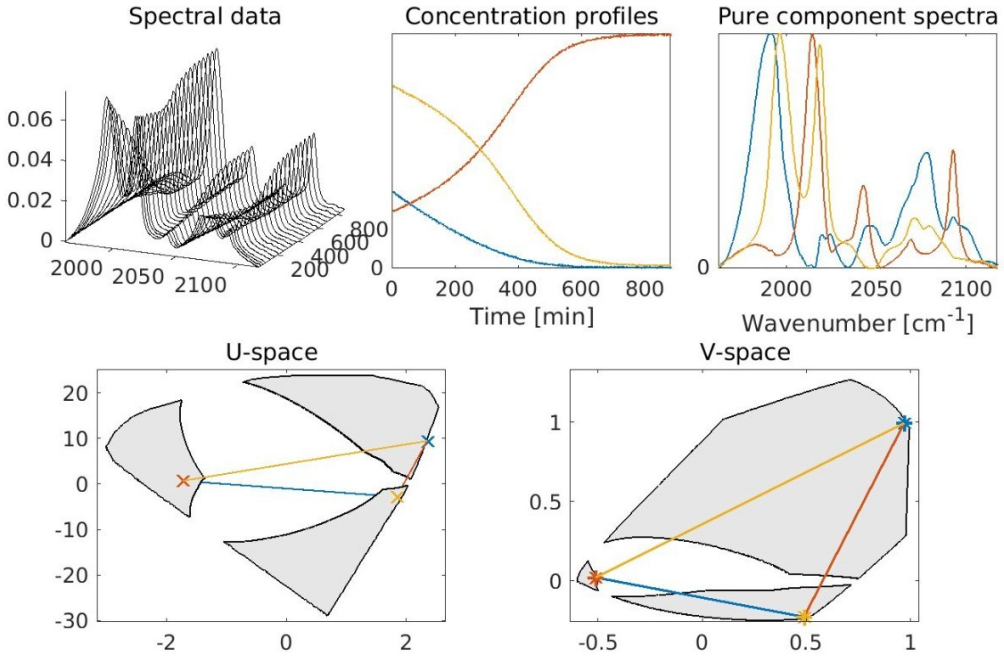


Figure – Top: Mixed spectral data & pure component profiles, Bottom: areas of feasible solutions (gray) and low-dimensional representations of the pure profiles (vertices of the triangles). Each edge/vertex of the triangle in U-space is dual to its associated vertex/edge of the triangle in V-space.

References

[1] H. Abdollahi, R. Tauler, *Chemom. Intell. Lab. Syst.* **108**(2) 2011, 100-111.
[2] M. Sawall, H. Schröder, D. Meinhardt, K. Neymeyr, *In Comprehensive Chemometrics: Chemcial and Biochemical Data Analysis*, Eds. S. Brown, R. Tauler, B. Walczak, Elsevier 2020, 199-231.
[3] M. Sawall, C. Kubis, K. Neymeyr et al., *J. Chemom.* **34**(2) 2020, e3159.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

INTEGRATION OF HANDHELD SPECTROMETERS AND CHEMOMETRICS FOR FOOD AUTHENTICATION

H. Parastar[1]
[1] *Department of Chemistry, Sharif University of Technology, Tehran, Iran*
h.parastar@sharif.edu

Nowadays, miniaturized spectrometers have been emerged as new technologies with many different applications [1]. Owing to the important role of miniaturized near-infrared (NIR) spectrometers, this technology and its applicability is explored for food authentication [2]. Unlike a rather uniform design of a mature benchtop FT-NIR spectrometer, miniaturized instruments employ diverse technological solutions, which have an impact on their operational characteristics. Continuous progress leads to new instruments appearing on the market. The current focus in analytical NIR spectroscopy is on the evaluation of the devices and associated methods, and to systematic characterization of their performance profiles. The technologies are not specifically aimed at certain commodities or product features, and no single technology can be applied for authentication of all commodities. Nevertheless, many useful applications have been developed for many food commodities [3].
The trade-off for using these devices is that the spectral region and resolution are limited compared to benchtop technologies. Additionally, scattering effects and instrumental and ambient noise make robust chemometric and machine learning methods crucial to extract the relevant information from the spectra.
The focus of the present contribution is summarizing miniaturised technologies, commercially available devices, chemometric data analysis methods and device applications for food authentication or measurement of features that could potentially be used for authentication [3]. We focus on the handheld technologies and their generic characteristics: (1) technology types available, (2) their design and mode of operation, and (3) chemometric data handling. Subsequently, two examples of recent applications are reviewed with details [4, 5]. It is important to note that the use of these applications in practice is still in its infancy. This is largely because for each single application, new spectral databases need to be built and maintained. Therefore, apart from developing applications, a focus on sharing and re-use of data and calibration transfers is pivotal to remove this bottleneck and to increase the implementation of these technologies.

References

[1] J. Muller-Maatsch, S. M. van Ruth, *Foods*, **2021**, 10, 2901.
[2] K.B. Bec, J. Grabska, C.W. Huck, *Chem. Eur. J.,* **2021**, 27, 1514.
[3] S. Lohumi, S. Lee, H. Lee, B.-K. Cho, *Trends Food Sci. Technol*., **2015**, 46, 85.
[4] H. Parastar, G. van Kollenburg, Y. Weesepoel, L. Buydens, J. Jansen, *Food Control*, **2020**, 112, 107149.
[5] S. Ehsani, E.M. Dastgerdi, H. Yazdanpanah, H. Parastar, *J. Chemom*., **2022**, xx, xxx.

ANALYTICAL QUALITY BY DESIGN USING A COMPUTATIONAL APPROACH FOR THE INVERSION OF A PLS MODEL

MC. Ortiz[1],
[1]*Dpto. de Química. Facultad de Ciencias. Universidad de Burgos, Burgos, Spain*
mcortiz@ubu.es

This work presents a computational approach for the inversion of a PLS model [1] in the framework of Quality by Design (QbD). In the context of Process Analytical Technology (PAT), one has process variables measured over 'n' samples and over the samples themselves the variables which define the quality of the product. The aim is to build predictive models to estimate the expected quality of the product (Critical Quality Attributes, CQA) as a function of the characteristics of the process variables (Control Methods Parameters, CMP). PAT can be applied to the development of analytical methods in chemical laboratories, in which case QbD is the so-called Analytical Quality by Design (AQbD) [2].

Given the frequently large number of process variables and quality characteristics and their correlation, in both spaces, these prediction models are usually based on latent variables as Partial Least Squares, which will be the model used in this work. However, it is also important to maintain a given quality, for example, in the framework of the Analytical Quality by Design. For this task, the interest is focused on the inversion of the PLS to discover at which values of the CMP the given CQA are fulfilled.

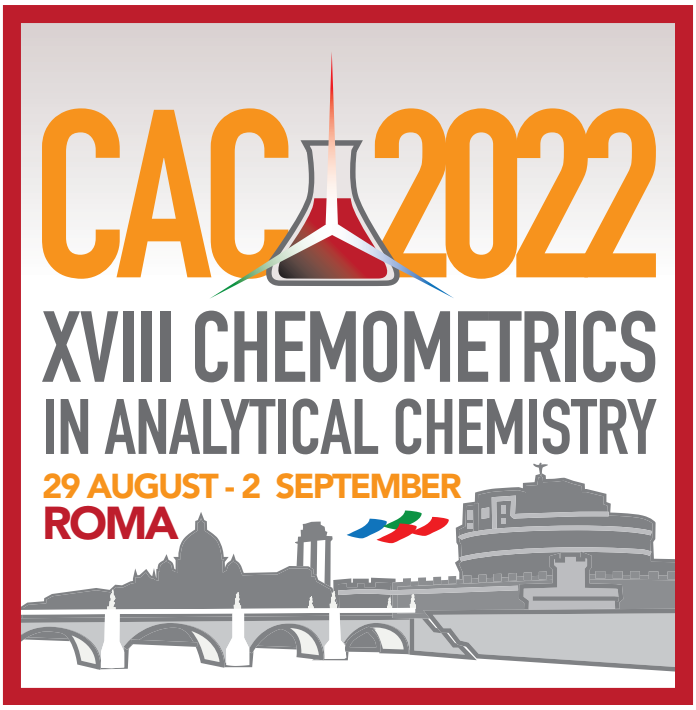
This work shows the procedure for selecting the optimal CMPs to obtain a preset 'analytical target profile' (CQAs) when a liquid chromatographic technique is going to be carried out for the simultaneous determination of five bisphenols, some of them regulated by the European Union for their toxicity [3]. Furthermore, in a second application it will show how to obtain the Method Operable Design Region (MODR) that is the core of the AQbD. This last approach has been carried out to determine 10 PAHs measured by liquid chromatography with fluorescence detection [4].

The global chemometric procedure has four steps: i) to build a D-optimal experimental design to reduce the number of experiments to carry out with the CMP ii) to fit a PLS2 model to predict the analytical responses (QCA), namely the resolution between each pair of contiguous peaks and final chromatographic time, as a function of the CMP. iii) to invert the PLS2 model, by means of a computational approach powered by an evolutionary algorithm, to obtain the conditions needed for attaining a preset QCA and iv) to obtain the Method Operable Design Region, as the convex envelope of the Control Method Parameters that provided compliant chromatograms.

Acknowledgments: This work is part of the project with reference BU052P20 financed by Junta de Castilla y León with the aid of European Regional Development Funds.

References

[1] S. Ruiz., M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, *Chemometrics Intell. Lab Systems*, 182 (2018) 70-78.
[2] T. Tome, N. Žigart, Z. Časar, A. Obreza, *Org. Process Res. Dev*. 23:9 (2019) 1784-1802.
[3] M.M. Arce, S. Ruiz, S. Sanllorente, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, *Anal. Chim. Acta* 1149 (2021) 338217.
[4] M.M. Arce, S. Sanllorente, S. Ruiz, M.S. Sánchez, L.A. Sarabia, M.C. Ortiz, *J. of Chromatography A*, 1657 (2021) 462577.



ORAL LECTURES

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

ENHANCING LC-MS/MS SPECTRAL SEARCHING WITH MULTI-TASK NEURAL NETWORKS AND MOLECULAR FINGERPRINTS

D. Ballabio[1], V. Consonni[1], F. Gosetti[1], V. Termopoli[1], R. Todeschini[1]
[1]*Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, Milan, Italy*
davide.ballabio@unimib.it

Liquid chromatography with tandem mass spectrometry (LC-MS/MS) is one of the most effective analytical techniques to characterize biological samples [1]. However, the identification of molecules usually requires searching for a match in a spectral library, that is based on creating a library of annotated spectra against which individual spectrum can be searched for. Building a spectral library is time-consuming and dependent on the LC-MS/MS instrumentation, while freely available libraries typically cover a limited number of molecules [2]. Moreover, it is likely that a molecule detected in a metabolomics trial may not be present in the reference library.
Recent studies have shown that deep-learning-based approaches can be used for structure elucidation of unknown compounds with their MS spectra [3]. These tools can be applied to directly predict molecular structures from large databases of spectra measured under different experimental conditions. Molecular structures can be numerically represented through in-silico fingerprints, which are binary vectors that encode features of molecules. Prediction of molecular fingerprints starting from the LC-MS/MS spectra would consequently assist the match of target compounds, which would benefit from the increased dimension of fingerprint databases.

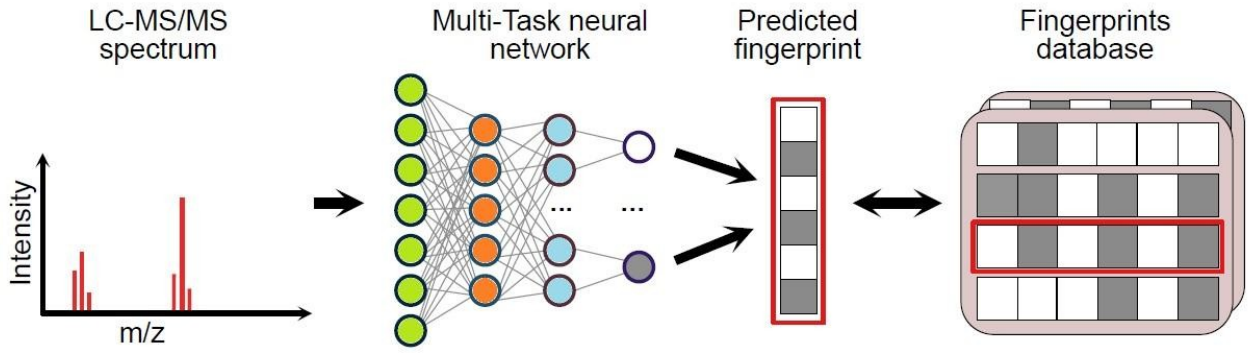


Figure 1 - Workflow for the fingerprint prediction from LC-MS/MS through a multi-task neural network and fingerprint matching.

In this study, we trained multi-task neural networks to predict molecular fingerprints starting from the LC-MS/MS spectra (Figure 1). MS spectra from available sources (MassBank of North America [4]) were initially collected and cured, leading to a dataset including around 40'000 spectra. Fingerprints were calculated (MACCS167 keys) and classification tasks were carried out by training multi-task feedforward neural networks to predict the binary bits of molecular fingerprints. Due to the extremely sparse nature of MS spectra, data reduction methods based on sparse PCA were also tested, using the PCA scores as input of the neural network training. Models were validated through specific protocols (training, test and validation sets) and demonstrated to have suitable performances in terms of accuracy (85% of bits correctly predicted).
Fingerprints predicted for the validation chemicals were used as targets for similarity searching in a very large fingerprint database (around 14 million structures), leading to promising preliminary results, with 26% of target chemicals found among the 10 nearest similar structures in the 14 million structures database.

References

[1] SK. Grebe, RJ. Singh. Clin Biochem Rev. 2011, 32, 5–31.
[2] H. Lam. Mol Cell Proteomics. 2011, 10(12), R111.008565.
[3] H. Ji, H. Deng, H. Lu, Z. Zhang, Anal. Chem. 2020, 92, 8649-8653
[4] S. Mehta, https://mona.fiehnlab.ucdavis.edu/

On the use of Bayesian statistics for (big) data analysis: automation for both qualitative and quantitative chemometrics

Gabriel Vivó-Truyols[1]

[1]*Tecnometrix, finca Mallaui, Camí de St. Joan de Missa, s/n, 07760 Ciutadella de Menorca – Balearics - Spain*

g.vivotruyols@tecnometrix.com

Data analysis methods applied to chemical data (either chromatographic or spectroscopic), are a routine part of most modern analytical workflows. With the emergence of n-th order instruments, the large data sets pose a new challenge for the data analysis. Basically, we are witnessing a boom of the amount of data to be processed, up to the point we can talk about Big Data in Analytical chemistry. Analysing these enormous and complex quantities of data becomes a tremendous challenge, especially because of the need of automation.

Automation is always a challenging task. In most of cases, the scientist has to "rely" on the algorithm taking (automated) decisions on both qualitative (e.g. peak identification) or quantitative (e.g. calibration) processes. However, Bayesian statistics offers an actual paradigm shift on the automation process. Contrary to classical methods mentioned above, it is not the algorithm but the scientist the one who takes the decisions, and the role of the algorithm is to calculate the probabilities of the variables of interest. This way of thinking opens a new world of possibilities. In this way, the scientist has no longer to "trust" the results of the algorithm, but (s)he has to decide on the different configurations that explain the data, based on the probabilities of each one.

This way of thinking has been applied to a broad range of situations. One example concerns toxicological screening, in which the probabilities of a list of compounds being present in the sample, analysed with LC-MS. Using a Bayesian approach, it is easy to build up evidence about the presence/absence of a compound by taking into account adduct formation, isotope ratios, retention times and mass values, resulting in more accurate values of probability. Another example is a Bayesian version of MCR-ALS, in which the classical multivariate curve resolution is applied probabilistically, so the question of parsimony of the model (i.e. how many compounds are present) is solved using Bayesian model averaging. Another application to be discussed concerns peak assignment, which is tackled from a combinatorial optimization perspective.

All in all, the use of Bayesian statistics to deal with massive data treatment constitutes a shift in the way we think about data analysis. Basically, we are proposing to work with probabilities of hypotheses (and update them as long as more information/data is taken into account), opposed to deliver the final answer to the user.

Chemometrics with Amazon Web Services (AWS)

Michael Sorochan Armstrong[1], Wenwen Li[1], A. Paulina de la Mata[1], James Harynuk[1]
[1]*University of Alberta, Edmonton, Canada*
mdarmstr@ualberta.ca

Cloud computing enables users to “rent” cores and memory on large-scale infrastructure to accelerate development and deployment of production-ready models. This is especially popular for deep learning routines that are computationally expensive to train, and may require large volumes of storage for high-dimensional data. Conventional chemometrics largely focuses on interpretable, linear models that are computationally inexpensive to train and implement, so cloud computing has not been prioritized in the same way as it has been for adjacent fields.

However tensor decomposition methods such as PARAFAC/PARAFAC2 and wrapper methods for feature selection can return simple and interpretable multi-linear or linear models that may yet require a great deal of computational time. For PARAFAC-type models, reliance on the standard Alternating Least Squares (ALS) algorithm is a major bottleneck, although recent developments have demonstrated that speed and convergence to a global optimum can be improved upon[1,2]. Wrapper methods for feature selection are popular for chromatographic data that is often plagued with missing values; and while this can help avoid the problem of building a model based off poorly integrated information, there is an inherent degree of redundancy that is costly to overcome with conventional hardware.

In this presentation, basic workflows for integrating a popular tool for cloud computing (Amazon Webs Services - AWS) into a workflow that is relevant to chemometrics will be discussed. A part of the presentation will include an example of a wrapper algorithm as a web-app powered by AWS (Feature Selection by Cluster Resolution)[3]. Although a popular platform for cloud computing, AWS offers a number of products that are not always intuitive to navigate and build into a pipeline for the uninitiated. This presentation will seek to provide guidelines for researchers interested in using cloud-based infrastructure, and demonstrate its potential for better communicating algorithms developed by chemometricians, with researchers in applied fields such as metabolomics.

References

[1] Yu, Huiwen, Dillen Augustijn, and Rasmus Bro. "Accelerating PARAFAC2 algorithms for non-negative complex tensor decomposition."*Chemometrics and Intelligent Laboratory Systems* **214** (2021): 104312.
[2] Yu, Huiwen, and Rasmus Bro. "PARAFAC2 and local minima." *Chemometrics and Intelligent Laboratory Systems* **219** (2021): 104446.
[3] Armstrong, Michael Sorochan, A. Paulina de la Mata, and James J. Harynuk. "An efficient and accurate numerical determination of the cluster resolution metric in two dimensions." *Journal of Chemometrics* 35.7-8 (2021): e3346.

Process economy, efficiency and sustainability go hand in hand, how chemometrics can build a greener industry

Jeroen J. Jansen[1]
[1]*Analytical Chemistry & Chemometrics, Radboud University, Nijmegen, the Netherlands*
jj.jansen@science.ru.nl

Chemometrics, or Process Analysis, has played an indispensable role in industry for several decades already. The value of PAT for monitoring and prediction of Quality is still continuously increasing. Chemometrics, through the omics revolution, has also taught us how a systems perspective is essential for accurate and early disease diagnosis and understanding of the system. Industrial processes are engineered and therefore **systems knowledge** is also widely available, from **process diagrams** to empirical **operational knowledge**. Therefore, there is a new challenge of translating such diagrams into modeling structures, where well-known paradigms like path modeling may provide new avenues to improve diagnosis and predictions. The systems perspective may also lead to much higher operational value, as improved process operations and control may not only benefit end-product quality, but through the entire process may also lead into more cost-efficient operations and several ways to increase **Industrial Sustainability**. This requires integration of process control with continuous improvement, where semi-quantitative tools, like the SQDCME process may be used to translate process observations, control operations and evidence-based process improvements into **KPIs for Sustainability, Quality and Economics**.
I will show in several of our recent industrial case studies [1, 2, 3] how developing new chemometrics methods and paradigms to include more and complementary process knowledge from engineering and process operations, lead to good processes that are cheaper and greener. This both increases the value of industry to our society and greatly increases the value of chemometrics to process industry.

References

[1] van Kollenburg, G. H., van Es, J., Gerretzen, J., Lanters, H., Bouman, R., Koelewijn, W., ... & Jansen, J. J. (2020). Understanding chemical production processes by using PLS path model parameters as soft sensors. *Computers & Chemical Engineering*, *139*, 106841.
[2] Galvis, L., Offermans, T., Bertinetto, C. G., Carnoli, A., Karamujić, E., Li, W., ... & Jansen, J. J. (2022). Retrospective quality by design r (QbD) for lactose production using historical process data and design of experiments. *Computers in Industry*, *141*, 103696.
[3] Teng, S.Y. *et al, Machine-Learned Digital Phase Switch for Sustainable Chemical Production. Submitted to Journal of Cleaner Production, May 22, 2022*

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

QSRR for small pharmaceutical compounds in RPLC: A Machine learning approach

Priyanka Kumari[1], Thomas Van Laethem[1,2], Philippe Hubert[1], Marianne Fillet[2], Pierre-Yves Sacré[1], Cédric Hubert[1]

1. University of Liège (ULiege), CIRM, Laboratory of Pharmaceutical Analytical Chemistry, Liège, Belgium

2. University of Liège (ULiege), CIRM, Laboratory for the Analysis of Medicines, Liège, Belgium

Priyanka@student.uliege.be

Reversed-Phase Liquid Chromatography (RPLC) is a common liquid chromatographic mode used to characterize drug substances by separating e.g.the principal active ingredient from its impurities or matrix excipients. Nevertheless, determining the optimal chromatographic conditions that enable this separation is time-consuming and requires a lot of lab work. Quantitative Structure retention Relationship models (QSRR) are helpful for doing this job with minimal time and cost since they allow for predicting retention times of known samples without performing experiments.
In the current work, we developed QSRR models and compared the strength of various machine learning algorithms which were based on a combination of linear and non-linear algorithms such as Multiple Linear regression (MLR), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosted Regression (GBR). to predict the retention times. The dataset comprised small molecules covering a wide range in terms of physicochemical properties related to the selected retention mechanism. Models were built for data acquired at two pH conditions, i.e., at pH 2 and 5 at a gradient time of 20 minutes (0 to 95% methanol) using a C18 T3 column. In the end, the model predictions were combined using stacking, and the performances of each model were compared. Here, QSRR models of the retention prediction have been built using structured derived physicochemical molecular descriptors, under consideration of the OECD principles in regulation for their acceptability and validation check during model construction and assessment. The KNN-based application domain filter was established to assess the reliability of the prediction for further compound prioritization. The best model was selected based on comparative values of R^2 and RMSE values on 10-fold cross-validation. Then the model was assessed on a holdout test data.
Out of all models, the stacked model at pH5 outperformed with RMSECV= 1.93, RMSEP = 2.02 minutes and R^{2pred} = 0.94.
The strategy used in this work is proposed as a generic workflow for the QSRR modelling of any RP-LC dataset when only a few number of samples is available. The QSRR models developed may then be used in an “in vitro” screening step to reduce the lab work and the lead time for chromatographic method development.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

COMPARISON OF MID-LEVEL FUSION STRATEGIES FOR THE MULTI-OMIC ANALYSIS OF TOXICOLOGICAL DATA

A. Menendez-Pedriza[1], L. Navarro-Martín[1] and J. Jaumot[1]
[1]1Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Barcelona, Spain.
ampgam@idaea.csic.es

An emerging interest related to the ecotoxicological assessment of substances of concern, such as Endocrine Disrupting Chemicals (EDCs), is that they are able to modify cellular regulatory mechanisms long after exposure, a long-term effect linked to epigenetics [1]. However, the mechanisms of action remain unclear when applying single-omic workflows. For that reason, multi-omic approaches are required to understand toxicity mechanisms [2]. The development of these multi-omics approaches has resulted in various tools, methods, and platforms provisioning data analysis, visualization, and interpretation.
However, multi-omic data approaches have to face some pending issues to achieve the full potential of combining high-throughput data obtained from different molecular layers.These challenges include the heterogeneity across omics technologies, the treatment of missing values, the difficulty of interpreting multilayered systems models, and the problems related to data annotation, storage and computational resources [3].
In the present study, we compare the capability of different multi-omic approaches taking as a study case the toxicity assessment of tributyltin (TBT) in zebrafish embryos, widely considered an excellent alternative animal model [4], integrating three-omic levels: epigenomics, transcriptomics and metabolomics. On the one hand, standard approaches such as PaintOmics or Multiple Factor Analysis (MOFA) were used to perform the data analysis and determine benchmark results. This multi-omic integration was performed in two steps following a mid-level fusion strategy (Figure 1). Firstly, data sets were normalized and analyzed individually to identify differentially altered features. After that, selected features were merged and integrated multi-omic analyses were conducted. On the other hand, the same strategy has been followed to analyze the data using chemometric methods able to perform this mid-level data fusion, such as Multivariate Curve Resolution (MCR). In this case, different data arrangements and block scaling strategies have been tested in order to improve the interpretability of the obtained results. Finally, the results obtained by the various tested methods have been compared, and the advantages and drawbacks of each approach are discussed.

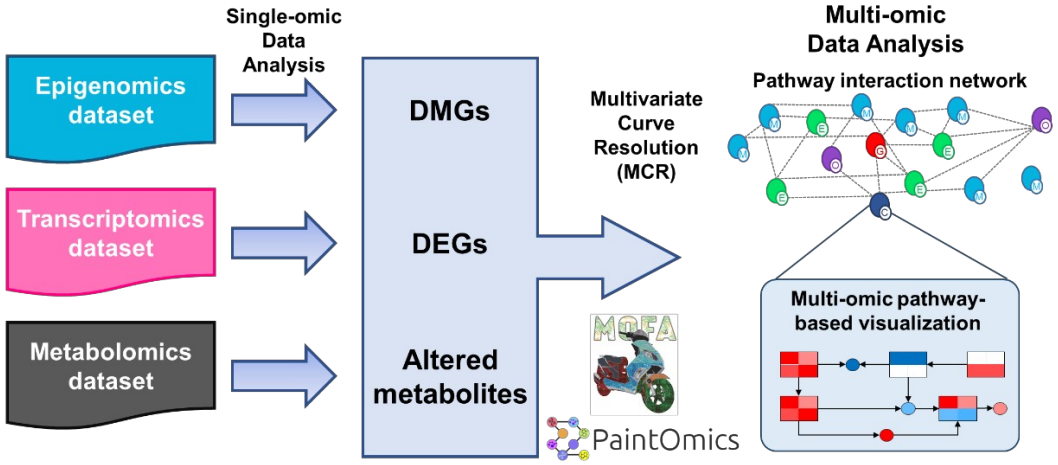


Figure 1 – General scheme of the mid-level fusion strategy conducted taking into consideration 3 different omics data (epigenomics, transcriptomics and metabolomics).

References

[1] I. Da Silva, *et al.*, *Frontiers in Endocrinology* **8** (2018) 366.
[2] H. Lee, *et al.*, *Environment International,* **157** (2021) 106802.
[3] S. Tarazona, *et al., Nat. Comput. Sci.* **1** (2021) 395-402
[4] A.J. Hill, *et al.*, *Toxicol. Sci*. **86** (2005) 6–19

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Metabolomics-guided insights on Bariatric Surgery: a longitudinal chemometrics approach over ^{1}H NMR spectra from serum samples

Martínez Bilesio AR[1,2]; Argüello MA[3]; Matellicani G[3]; Nasurdi A[3]; Barrera MM[3]; Paz J[3]; Rocca L[3]; Sciara M[4]; Fay F[4]; Jaumot Soler J[2]; Rasia R[1]; García-Reiriz AG[5]; Burdisso P[1]

[1] *Argentine Platform for Structural Biology and Metabolomics (PLABEM), Institute of Molecular and Cell Biology of Rosario (IBR), National Scientific and Technical Research Council (CONICET), Rosario, Santa Fe, Argentina*

[2] *Institute of Environmental Assessment and Water Research (IDAEA), Spanish National Research Council (CSIC), Barcelona, Catalunya, Spain*

[3] *Interhospital Morbid Obesity Unit (UIOM), Provincial Hospital of Rosario, Rosario, Santa Fe, Argentina*

[4] *Cibic Laboratories S.A., Rosario, Santa Fe, Argentina*

[5] *Institute of Chemistry of Rosario (IQUIR), National Scientific and Technical Research Council (CONICET), Rosario, Santa Fe, Argentina*

martinezbilesio@ibr-conicet.gov.ar

Bariatric surgery is considered the most efficient treatment for diseases related to morbid obesity [1]. This surgical procedure has proven successful for weight loss, but also for the progression control of type 2 diabetes considering its metabolic impact [2]. However, the effect of bariatric surgery on metabolism is still not well defined. In this sense, metabolomics analysis through high-throughput nuclear magnetic resonance (NMR) coupled with chemometric processing has emerged in biomedical research as a field able to shed some light on obesity-related metabolic diseases [3].

In the present study, we aimed to discriminate metabolic signatures linked to bariatric surgery and determine potential adaptations of different patients. A chemometrics-assisted ^{1}H NMR metabolomics approach was used in order to analyze serum samples of subjects with morbid obesity (n = 15), before (2-3 weeks) and after (48 hours, 5 days, 1, 6 and 12 months) bariatric surgery.

In the first step, different multivariate analyses were applied over the ^{1}H NMR spectra. Chemometric methods, including principal component analysis (PCA), ANOVA simultaneous component analysis (ASCA) and partial least squares discriminant analysis (PLS-DA), allowed identifying the main metabolic responses associated with the bariatric surgery. We defined two metabolic phenotypes of response (metabotypes) independently of gender, age or body mass index (BMI). In addition, it was possible to elucidate general metabolic profiles over time, distinguishing three primary temporal trends throughout the bariatric surgery evolution.

In the second step, a multivariate curve resolution (MCR)-based strategy was applied to obtain and evaluate the peak integrals profiles of the previously identified metabolites (biomarkers) [4]. By this means, a significant reduction in the dataset dimensionality was achieved, without losing the potential for biological interpretation. Then, the chemometric evaluation of this reduced features matrix using the same tools as in the first step allowed the identification of those metabolites associated with the metabolic changes induced by the bariatric surgery.

Although further studies are needed, our results open new hypotheses in the study of obesity-linked co-morbidities and provide a comprehensive view of the metabolic changes after the surgery.

References

[1] Rubino F, et al. Metabolic surgery in the treatment algorithm for type 2 diabetes: a joint statement by international diabetes organizations. *Diabetes Care*. 2016;39: 861-877.

[2] Mingrone G, et al. Bariatric-metabolic surgery versus conventional medical treatment in obese patients with type 2 diabetes: 5 year follow-up of an open-label, single-centre, randomised controlled trial. *Lancet*. Elsevier Ltd; 2015;386: 964-973.

[3] Palau-Rodriguez M, et al. Metabotypes of response to bariatric surgery independent of the magnitude of weight loss. *PLoS One*. 2018;13(6): e0198214

[4] Puig-Castellví F, et al. Untargeted assignment and automatic integration of ^{1}H NMR metabolomics datasets using a multivariate curve resolution approach. *Analytica Chimica Acta*. 2017;964: 55-66.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Evaluation of mid-infrared spectra of serum and synovial fluid in predicting early post-traumatic osteoarthritis in an equine model

S. Malek[1], F. Marini[2], T. Trumble[3], S.C. Buono [2]
[1]*Dept. of Veterinary Clinical Sciences, Purdue University College of Veterinary Medicine, W. Lafayette, USA*
[2]*Dept. of Chemistry, University of Rome La Sapienza, Rome, Italy*
[3]*Dept. of Veterinary Population Medicine, University of Minnesota, St. Paul, USA*
maleks@purdue.edu

Trauma to the joint results in acute inflammation and depending on the severity can significantly increase the risk of developing osteoarthritis (OA). Detecting acute changes leading to early OA can be challenging due to the insidious onset and lack of validated and reliable serum or synovial fluid biomarkers. The aim of this study was to evaluate Fourier-transform infrared (FTIR) spectroscopy of serum and synovial fluid as a candidate screening tool in diagnosing early post-traumatic OA (PTOA) in a non-terminal equine model. Twenty-two horses were included in this prospective study. Unilateral PTOA in the metacarpophalangeal (MCP) joint of 11 horses in the PTOA group was induced by arthroscopically creating an osteochondral (OC) fragment (i.e., OC joint subgroup). The contralateral MCP joint in these horses was arthroscopically evaluated but no fragment was created (i.e., sham joint subgroup). Eleven horses were used as Controls (i.e., two control MCP joints). Synovial fluid sample from both MCP joints and serum samples were obtained from all control and preoperatively from the PTOA horses at 0 and overtime at 2, 4, 6, 12 and 16 weeks. Samples were processed and stored in -80°C until batch analysis. Fourier-transform infrared spectroscopy of serum and synovial samples was performed by drying triplicates of each thawed-out sample on 96-welled silicone microplates. Absorption spectra at 400–4000 cm^{-1} wavenumbers was recorded at 4cm^{-1} resolution, 512 scans)[1]. After preprocessing, partial least squares discriminant analysis (PLS-DA) and multilevel-simultaneous component analysis (MSCA) were used to compare serum spectral results from samples from PTOA and control groups for serum and synovial fluid. For synovial fluid samples, OC joint samples were compared to the controls. In detail, performance of predictive (PLS-DA) models built for discriminating OA from control samples based on spectra from each time point were compared; multi-level simultaneous component analysis (MSCA) was used to evaluate the significance of spectral variations at each time point. All OC joints developed early PTOA at 16 weeks based on arthroscopic and radiographic examination. The performance of predictive models based on serum spectra at each time-point between PTOA and control groups were poor. The synovial fluid spectra at 0 time point had minimal variability between OC joint and control groups. The synovial fluid spectra showed significant discriminatory capability from 2 weeks after the injury that was reduced at 16 weeks **(Table 1)**. In conclusion, FTIR spectra of synovial fluid from joints with PTOA can be distinguished from controls. This discriminatory capabilities is reduced by 16 weeks which may be due to natural reduction in the initial inflammatory response to the trauma. The results demonstrate the potential for FTIR spectroscopy of synovial fluid from joints with acute trauma as a screening tool for early PTOA.

Sampling intervals	Accuracy	Classification error	Sensitivity	Specificity
0	53.6±8.3	46.3±8.3	51.8±9.9	55.6±12.6
2	93.2±3.2	6.8±3.2	98.4±4.4	88.0±4.7
4	80.4±4.5	19.6±4.5	88.2±7.4	72.6±6.0
6	69.6±4.2	30.4±4.2	68.9±6.1	70.4±7.1
12	72.9±5.6	27.1±5.6	75.8±7.5	70.0±7.6
16	56.7±6.8	43.3±6.8	60.7±9.1	52.7±10.1

Table 1 – Predictive model performance for discriminating synovial spectra of control and PTOA samples. Sampling intervals are based on weeks. The model performance is presented in % and ± standard deviation.

1. Malek S, Marini F, Rochat MC, Béraud R, Wright GM, Riley CB. Infrared spectroscopy of synovial fluid as a potential screening approach for the diagnosis of naturally occurring canine osteoarthritis associated with cranial cruciate ligament rupture. Osteoarthritis and Cartilage Open. 2020;2(4):100120. doi: https://doi.org/10.1016/j.ocarto.2020.100120.

CLASS-MODELLING – OPTIMIZATION, VALIDATION AND APPLICATION

Z. Małyjurek, B. Walczak
Institute of Chemistry, University of Silesia, Katowice, Poland
zuzanna.mitrega@op.pl

Class-modelling, also known as one-class classification, is used for individual class-model construction based on the similarities among samples of the class studied, the target class. The obtained model is used to predict whether a new sample of unknown origin belongs to the class of interest. Class-modelling is widely applied for, e.g., authentication, quality control, and novelty detection. The class-model construction is a multistep process that includes the construction of the training and test sets, selection of the class-modelling method, optimization of the model, and validation. Each step has an impact on the final classification outcomes thus they should be handled carefully [1].

The aim of this study was to indicate the optimal strategy for class-model optimization, authentication of similar classes, limitations and scopes of applicability of selected class-modelling methods.

Concerning the optimization of the class-model, i.e., the selection of the model complexity and classification rules, two scenarios can be distinguished: rigorous and compliant. The rigorous strategy is considered when only target class samples are used for class-model optimization, whereas in the compliant strategy, target and nontarget class samples are taken into account. There is an ongoing discussion on whether nontarget class samples can be used for class-model optimization since they cannot be treated as representative. For this reason, in the thorough comparative study upon the example of SIMCA model optimization, the influence on final classification results of rigorous and compliant optimization strategies were tested. It was demonstrated that class-models optimized in the compliant scenario combined with the optimized decision threshold lead to the most effective models. However, it was also shown that these models can be biased, thus the rigorous scenario is a safer option for class-model optimization [2].

Once the optimal optimization strategy was selected, the scope of applicability of different class-modelling methods was tested based on the several datasets of various structures. The SVDD (Support Vector Description Domain) models led to the highest classification results most often. However, when data of complex structure are analysed, the density-based class-modelling methods, such as PFM (Potential Functions Method), are more suitable than SVDD.

In the case of authentication of several similar classes that overplap in the feature space, the individual class-models can lead to unsatisfactory results. In such situations, the discriminant model allows obtaining better classification results than class-models, but classical discriminant approach cannot be applied for authentication. Therefore, a two-step approach that combines class-modelling and discriminant approaches has been proposed, and its performance was compared with so-called soft discriminant methods [3] and SIMCA optimized using the ROC curve [4].

The conclusions obtained from the aforementioned aspects studied allowed to propose a successful strategy for authentication of rooibos, honeybush, and three Cyclopia species used for honeybush production.

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] P. Oliveri, *Anal. Chim. Acta* **982** (2017) 9-19.
[2] Z. Małyjurek, R. Vitale, B. Walczak, *Talanta* **215** (2020) 120912.
[3] A. L. Pomerantsev, O.Y. Rodionova, *J. Chemom,* **32** (2018) e3030.
[4] R. Vitale, F. Marini, C. Ruckebusch, *Anal. Chem.* **90** (2018) 10738-10747.

LIKELIHOOD RATIO IN FORENSIC DISCRIMINATION/CLASSIFICATION TASKS

A. Martyna*[1], A. Nordgaard[2], E. Alladio[3,4]
[1]*Institute of Chemistry, University of Silesia in Katowice, Poland*
[2]*Department of Computer and Information Science, Linköping University, Sweden*
[3]*Chemistry Department, University of Turin, Italy*
[4]*Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Turin, Italy*
agnieszka.martyna@us.edu.pl

Classification or discrimination tasks, which aim at establishing the class membership of the analysed evidence materials based on the predefined decision rules, are very frequent in the forensic expert practice. According to the recommendations of the European Network of Forensic Science Institutes (ENFSI), the conclusions drawn in the forensic investigations must be formulated in the probabilistic manner. One of the widely studied and accepted means is the *likelihood ratio*, LR, defined as $LR=Pr(E|H_1)/Pr(E|H_2)$. It describes how many times the evidence (E) supports the H_1 in contrast to H_2 [1]. In forensic classification the hypotheses state that the evidence comes from either one (H_1) or the other (H_2) class. However, originally, the pioneering LR models for interpreting the physicochemical data for forensic purposes were developed by Aitken and Lucy [1] in the aim of establishing whether the glass samples may come from the same source or not. This poses a slightly different problem than typical classification, since 'same source' or 'different source' cannot be ragarded a particular class. Nevertheless, the undisputed advantage of these models is the determination of how high the degree of similarity between compared samples is in the light of the frequency of occurrence of the analysed features in the entire population. Their main idea is that the similarity found between very rare features is much more valuable than the same similarity observed between frequent features.
This potential of LR models may also be applied in the typical classification tasks after undergoing appropriate changes. If rare features occur in one class only (despite their rarity), the questioned sample with such rare features strongly supports the hypothesis that it comes from this class. This corresponds to high LR values for such comparisons. The necessary changes mainly involve variance estimates definitions. Each questioned sample is decided to come or not from the same class as each of the labelled samples. The LR values received in such comparisons are then translated to probabilities using, e.g., logistic regression. These probabilities can be seen as the reliability that the LR value indicates the same class of the compared samples. Finally, the class membership of the analysed sample is established by defining the appropriate and balanced decision rules which take into account that the highest LR values (and probabilities) refer to the very strong indications of the common class membership of the samples. The proposed workflow will be presented using a few examples from the forensic field.

The research was conducted within the project No. 2019/35/D/ST4/00933 financed by the National Science Center in Poland.

References

[1] C. Aitken, D. Lucy, Appl. Statist **53** (2004) 109-122.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

SUPERVISED AND UNSUPERVISED CHEMOMETRIC METHODS TO DEAL WITH SAFFRON AGING AND ITS QUALITY CONTROL

M. Foschi[1], A. Biancolillo[1], A.A. D'Archivio[1]
[1]*Dept. of Physical and Chemical Sciences, University of L'Aquila, Coppito, L'Aquila, Italy*
martina.foschi@univaq.it

The saffron production process, authentication, and quality control remain topical due to the interesting biological activities and multiple applications of its phytocompounds. Saffron production consists of several time-consuming phases; whether the cultivation conditions and harvesting stages can affect the final yield of the product and, together with the high costs, the propensity of producers to commit fraud, the drying and the storage conditions phases actually affect the commercial quality of the spice. Indeed storage induces oxidative or hydrolytic decomposition, whereas spice stability depends on relative humidity, temperature, and light exposure. In this context, the quality control is regulated by the ISO 3632-1 and 2, which allows saffron to be assigned to quality classes according to its moisture, quantity of extraneous material, and concentration of its secondary metabolites (picrocrocin, crocins, and safranal). Thus, the purpose of the work was to assess if, in the context of UV-Vis quality control of the spice, it is possible, by means of chemometrics, to differentiate between FRESH products (i.e., saffron produced and marketed within a year) COMPLIANT (product not yet expired) and legally EXPIRED products (produced over two years). Therefore, spectra were collected on 104 FRESH Umbrian samples between the years 2016 and 2020; exactly the same batches were preserved by simulating common home storage and reanalyzed in 2021 from the interval of 5 years (the 2016 samples) to the interval of 8 months (the 2020 samples).

Dataset	Production Year	Aging	N samples
S2 0 1 6 (F)	1	1	18
S2 0 1 7 (F)	2	1	27
S2 0 1 8 (F)	3	1	28
S2 0 1 9 (F)	4	1	14
S2 0 2 0 (F)	5	1	15
S2016(AE)	1	2	18
S2017(AE)	2	3	27
S2018(AE)	3	4	28
S2019(AC)	4	5	14
S2020(AC)	5	6	15

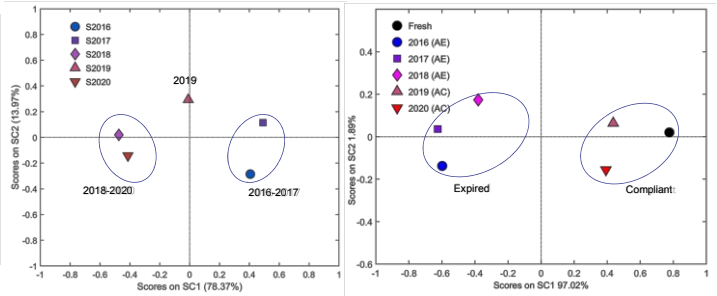


Figure 1 – Dataset of Fresh (F), Aged but Compliant (AC) Aged and Expired (AE) samples, factor columns, and number of samples; ASCA centroids for the Production Year (left) and Aging (right) factors

Anova-Simultaneous Component Analysis confirmed a significant effect of both the Production Year and Aging. The loading inspection confirmed the loss of the spice color power due to the crocins degradation [1]. In a preliminary stage, to release the treatment from the Production Year variability, Partial Least Squares-Discriminant Analysis and Soft Independent Modeling of Class Analogy were performed on signals obtained by subtracting, from each sample, its FRESH spectrum. Although excellent results have been obtained with these methods, they have the limitation of requiring, to be applied, a reference spectrum of the fresh samples. Alternatively, SIMCA constructed on FRESH samples showed a sensitivity of 81% (6 samples erroneously refused over 32 in external validation), a specificity of 91% for the EXPIRED class and of 89% for the COMPLIANT class. The reported results are excellent, considering the model was built on a class consisting of samples produced over a 5-year period (variability recognized as significant at the exploratory stage). In conclusion, following the ISO regulation and with an appropriate database, it is possible to evaluate the compliance and the freshness of the spice during its quality control. Although the understanding of degradation phenomena is challenging with this kind of analytical technique, Multivariate Curve Resolution-Alternating Least Squares could provide further valuable information about the temporal evolution of the main components of saffron aqueous extracts.

References

[1] T.S. Cid-Perez, G.V. Nevárez-Moorillón, C.E. Ochoa-Velasco, A.R. Navarro-Cruz, P. Hernández-Carranza, R. Avila-Sosa, *Molecules* **26** (2021) 6954.

Real time prediction of ABS properties through multiblock and local regression methods.

L. Strani[1], R. Vitale[2], D. Tanzilli[1], F. Bonacini[3], A. Perolo[3], A. Ferrando[3], E. Mantovani[3] and M. Cocchi[1]

[1]*Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Modena, Italy*
[2]*Centre National de la Recherche Scientifique (CNRS), Laboratoire de Spectroscopie pour les Interactions, la Réactivitè et l'Environnement (LASIRE), University Lille, Lille, France*
[3]*Versalis S. p. A. - ENI group, Mantova, Italy*
lostrani@unimore.it

Acrylonitrile-Butadiene-Styrene (ABS) is a styrenic thermoplastic polymer characterized by high toughness and impact resistance. It is used for the manufacture of a large number of different kind of products, which is why is one of the most popular product on the market. ABS quality is evaluated every day through numerous and time-consuming off-line analyzes of several parameters, in order to ensure the production of a high quality and, therefore, a more desirable product. These parameters can be predicted in real time through the huge amount of plant data. In fact, along the ABS production plant are installed a large number of sensors, which continuously acquires information about the status of both the process (temperature, pressure, flows) and the forming product (NIR spectra). In this context, multiblock and local approaches can help to improve both process understanding and the performances of the prediction models, as the data collected by the sensors can be easily partitioned in different data blocks according to specific plant areas. Furthermore, through these approaches it is possible to obtain predictions of quality parameters without taking into account measurements related to the final area of the plant, i.e. before the product is complete.
The present study aims to apply different multiblock and local regression methods, fusing process sensors and spectroscopic (NIR) data, to calculate real time monitoring models for the prediction of ABS quality parameters. With the aim to assess which were the most relevant sensors and plant areas for the prediction of the ABS properties, and to be sure that the plant setting changes did not negatively affect the prediction quality, data collected by four NIR probes and more than 50 process sensors was analysed through MultiBlock-PLS, Response-Oriented sequential alternation (ROSA) [1] and ParSketch [2]. Several prediction models were computed involving data acquired by sensors present in different process stages. Low prediction errors were obtained through the explored approaches that allowed to identifying the most crucial data blocks for the ABS quality prediction. Additionally, prediction obtained by models computed without considering data blocks belonging to the final areas of the process were comparable to those obtained involving all the available data blocks. Therefore, a good estimation of ABS quality can be obtained in real time before the final product is completed, radically reducing the off-line laboratory analysis.

References

[1] K.H. Liland, T. Næs, U.G. Indahl. J. Chemom., **30** (2016) 651–662.
[2] M. Metz, M. Lesnoff, F. Abdelghafour, R. Akbarinia, F. Masseglia, J.M. Roger, Chemom. Intell. Lab. Syst., **203** (2020) 104076.

SPECTRAL IMAGING –PRE-HARVEST MALTING BARLEY ERMINATION CLASSIFICATION WITH SEQUENTIAL ORTHOGONALISED MULTIBLOCK DATA FUSION METHODOLOGIES

S.H. Orth[1], F. Marini[1,2], G.P. Fox [1,3], S. Hayward[1], M. Manley[1]

[1] Department of Food Science, Stellenbosch University, Private Bag X1, Matieland, Stellenbosch, 7602, South Africa,

[2] Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, 1-00185 Rome, Italy,

[3] Food Science and Technology, University of California Davis, 1Shields Ave, Davis, CA 95616, USA,

stefanh@sun.ac.za

Visible and near-infrared (NIR) imaging with multiblock methodologies, i.e., sequential and orthogonalised partial least squares-linear discriminant analysis (SO-PLS-LDA) and sequential and orthogonalised covariance selection-linear discriminant analysis (SO-CovSel-LDA) allow for modelling with a high degree of precision and sensitivity [1,2]. These sequential and orthogonalised methods do not introduce redundant spectral information commonly found in conventional low-and mid-level approaches. Malting barley (*Hordeum vulgare* L.) undergoes a series of steeping phases followed by controlled germination to produce malt for beer brewing. Pre-harvest germination, triggered by untimely adverse wet and humid environmental conditions, contributes to non-uniform malting and huge economic losses. An objective, rapid, user-friendly method, capable of early-stage germination detection may circumvent such losses [3]. Hyperspectral images, in the visible near-infrared (VNIR; 186 wavebands) and shortwave infrared (SWIR; 288 wavebands) wavelength regions were combined with SO-PLS-LDA and SO-CovSel-LDA classifiers to detect early-stage germinated single barley kernels. Classification accuracies (based on cross-validation) with SO-PLS-LDA and SO-CovSel-LDA, were 100% (full spectrum) and 99.19% (17 variables), respectively, with similar test set performance. An improvement was noted for the SO-PLS-LDA method and a slight decrease in accuracy for SO-CovSel-LDA when compared to PLS-DA (99.42 % SWIR, 99.48 % VNIR). The selected 15 SWIR and VNIR wavelengths were evenly spread across the entirety of spectral ranges. These included the extreme wavelengths which compensate for additive/multiplicative effects [2]. Using the multiblock variable selection procedure, SO-CovSel-LDA, the selected wavelengths could be further reduced to 13 (SWIR and VNIR), which corresponds to 2.7% of the variables of the original data blocks. Despite this great reduction in the number of variables, the classification accuracy remains comparable (97%) to that obtained with SO-PLS-LDA using the full-spectrum data set. This alludes to a multispectral approach with reduced sensor cost and added benefit of in-line/on-line implementation and real-time monitoring of malting barley pre-harvest germination – a sizable industry problem. VNIR and SWIR spectral information and multiblock data fusion methods provide viable research and industry solutions for the malting and brewing sector.

References

[1] A. Biancolillo, I. Måge, T. Næs, T., *Chemometr. Intell. Lab. Syst.* **141** (2015) 58–67.
[2] A. Biancolillo, F. Marini, J.M. Roger, *J. Chemometr.* **34** (2020) e3120.
[3] D.J. Mares, K. Mrva, K., *Planta* **240** (2014) 1167–1178.

PROCRUSTES CROSS-VALIDATION OF MULTIVARIATE REGRESSION MODELS

S. Kucheryavskiy[1], O. Rodionova[2], A. Pomerantsev[2]
[1]*Department of Chemistry and Bioscience, Aalborg University, Esbjerg, Denmark*
[2]*Semenov Federal Research Center for Chemical Physics RAS, Moscow, Russia*
svk@bio.aau.dk

When multivariate data is being fitted by a model, it is important to account for (and minimize) both fitting and sampling error. Estimating sampling error is the most difficult task in this case and usually this is done by applying the model to a new set of data, that belongs to the same population as the calibration set. This procedure is well known as the test set validation, which has been proven to be the most reliable and efficient.

However, if an intermediate step for optimization of the model and related parameters (e.g. preprocessing or variable selection) is necessary, two independent test sets should be used for a proper validation. In case when measurements are expensive, time consuming, or there are other obstacles, the test set validation can be substituted with a cross-validation – resampling-based approach, which merges the results obtained from several local models created for subsamples of the calibration set.

Recently we proposed an alternative approach, named Procrustes Cross-Validation (PCV), which takes the best parts of both validation methods [1]. PCV uses cross-validation to estimate the sampling error and then introduces the sampling error directly to the calibration set, thus creating a new set of measurements — a pseudo-validation set. This set can then be used to validate the global model in the same way as it is done using an independent test set.

The PCV method has been initially developed for PCA and the PCA based applications. The first example is SIMCA, in which a direct cross-validation is questionable. In this contribution we firstly present a new improvement, which makes PCV calculation several times faster. Additionally, this enhancement makes it possible to generalize the approach to a broader set of modelling methods, including PCR- and PLS-regressions. This generalization also extends the usability of PCV, which can be applied both for validation and optimization of the models and also for data augmentation. Results based on several simulated and real-life examples will be demonstrated.

References

[1] S. Kucheryavskiy, S. Zhilin, O. Rodionova, A. Pomerantsev, *Anal. Chem.* **92** (2020) 11842-11850.
[2] A. Pomerantsev, O. Rodionova, *Talanta* **226** (2021), 122104

OPENING THE RANDOM FOREST BLACK BOX WITH SURROGATE MINIMAL DEPTH

S. Seifert[1], S. Gundlach[2], Silke Szymczak[3]
[1]*Hamburg School of Food Science, University of Hamburg, Hamburg, Germany*
[2]*AG Software Engineering, Kiel University, Kiel, Germany*
[3]*Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany*
stephan.seifert@chemie.uni-hamburg.de

In order to comprehensively exploit the complex data generated by spectroscopic and spectrometric techniques, unsupervised and supervised multivariate chemometric methods are utilized. The latter, machine learning approaches, are often applied as black boxes meaning that only the class assignment is reported, while the background that led to this decision remains unknown.
Random forest (RF) is a non-parametric machine learning approach that consist of a large number of individual binary decision and has many advantages, such as flexibility in terms of input and output variables and the possibility of internal validation. Another advantage is the ability to generate variable importance measures that are used to select relevant features. However, the relationships between the predictor variables are usually not examined. We developed a novel RF based variable selection approach called Surrogate Minimal Depth (SMD) that incorporates relations into the selection process of important variables [1]. This is achieved by the exploitation of surrogate variables that have originally been introduced to deal with missing predictor variables. In addition to improving variable selection, surrogate variables and their relationship to the primary split variables can also be utilized as proxy for the relations between the different variables (see Fig. 1). This relation analysis goes beyond the investigation of ordinary correlation coefficients because it is based on the mutual impact on the outcome.
I will present the basic concept of surrogate variables, SMD and mean adjusted agreement, as well as their application to simulated data as proof of concept. In addition, I will present successful applications in different fields ranging from metabolomics analysis of food [2] to the exploitation of data from surface-enhanced Raman scattering experiments [3], e.g. to illuminate the interaction of drugs with proteins and lipids in living cells [4].

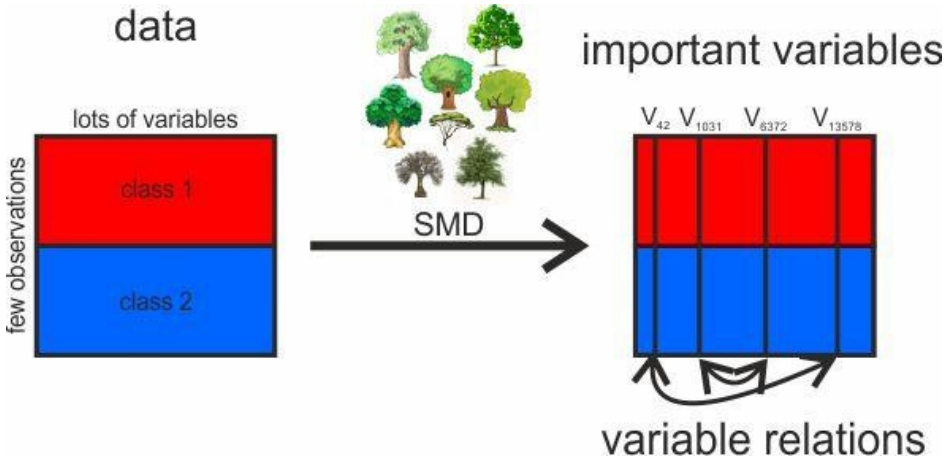


Figure 1 Illustration of the application of SMD to high-dimensional data.

R Package
https://github.com/StephanSeifert/SurrogateMinimalDepth.

References
[1] S. Seifert, S. Gundlach, S. Szymczak, Bioinformatics **35** (2019) 3663-3671.
[2] S. Wenck, M. Creydt, J. Hansen, F. Gärber, M. Fischer, S. Seifert, Metabolites **12** (2022) 5.
[3] S. Seifert, Sci Rep. **10** (2020) 5436.
[4] V. Živanović, S. Seifert, D. Drescher, P. Schrade, S. Werner, P. Guttmann, G. P. Szekeres, S. Bachmann, G. Schneider, C. Arenz, J. Kneipp, ACS Nano, **13** (2019) 9363-9375.

EXPANSION of the DD-SIMCA CONCEPT

O. Rodionova[a], N. Kurysheva[b], G. Sharova[c], A. Pomerantsev[a]
[a]*Federal research center for Chemical Physics RAS, Moscow, Russia*
[b] *Ophthalmological Center FMBA, Moscow, Russia*
[c] *Belikova MD Eye Clinic, Moscow, Russia*
oxana.rodionovaname@gmail.com

Data-driven SIMCA (DD-SIMCA) has proven to be highly effective in solving one-class classification problems even in fairly complex cases. At the same time, the DD- SIMCA concept can be used much more widely. Full distance (FD) that is a statistics calculated as a weighted sum of the score distance and orthogonal distance was reported as an efficacious tool for the outlier detection in projection-based methods [1], for the estimation of the limits of detection in qualitative analysis [2], and for selection of similar samples in nonlinear local weighted modeling [3].

In this contribution, we go further and present several novel and unusual applications of the DD-SIMCA concept, using a very interesting medical data [4] as an illustrative example.

The following problems are considered.

(1) Comparison of several multivariate datasets for similarity, which is a basic problem in statistics and machine learning and a very popular task in practical applications. We propose using β-value [5] and the extreme plots [6] as a reasonable alternative to the p-value approach.

(2) Evaluation of the efficiency of treatment. This a common problem in medicine and other fields, when it comes to assessing which method is preferable for curing of a particular group of patients. Our approach is calculation of FD and β values to assess proximity to the target class of the control (healthy) subjects.

(3) Selection of a personalized treatment. In this case, the FD value is used as a response in the PLS model, which predicts the distance of the subject (object) to the target class after a certain treatment, and, therefore, the probability of recovery. This approach can be used not only in medicine, but also in restoration work to preserve and restore cultural heritage sites.

Acknowledgements: We acknowledge partly funding from the IAEA in the frame of projects D5240 and G42007.

References

[1] O.Ye. Rodionova, A.L. Pomerantsev, *Anal. Chem*. **92** (2020) 2656−2664
[2] A.L. Pomerantsev, O.Ye Rodionova, *Trends. Anal. Chem.* **143** (2021) 116372.
[3] R. Zhao, X. Liu, .., S.Yang, *Foods*, **11** (2022) 846
[4] N. Kurysheva, A. Pomerantsev, O. Rodionova, G. Sharova *J.Glaucoma*, submitted (2022)
[5] A.L. Pomerantsev, O.Ye. Rodionova, *J. Chemometrics* **28** (2014) 518-522.
[6] O. Rodionova, S. Kucheryavskiy, A. Pomerantsev, *Chemom.Intell.Lab.Syst* **213** (2021) 104304.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Speeding up PARAFAC2 dramatically

P. Schneide[1], R. Bro[2]
[1]*BASF SE, Ludwigshafen, Germany*
[2]*University Copenhagen, Copenhagen, Denmark*
paul.schneide@basf.com

PARAFAC2 has shown great capabilities for extracting chemical information from chromatographic data like GC-MS. [1,2] A completely automized workflow has been published recently. [3] If a PARAFAC2 model converges to a global optimum, a unique solution can be obtained. However, especially for complicated data, the non-convex optimization might get stuck in a local minima or won't converge due to a degenerate solution. [4] Often, non-unique and degenerate solutions deviate drastically from the true underlying factors and thus are meaningless. [4] Recently published results indicate, that the problem of local minima can be reduced by applying constraints on the different modes (e.g. non-negativity), using sophisticated initialization schemes and refitting the model multiple times. [5] However, the study on the local minima reducing effect of constraints was so far limited to constraints on the non-shifted modes. Further, there is currently no strategy for how to efficiently handle the problem of degenerate solutions in PARAFAC2 models.
This work presents, a simple and highly efficient strategy for handling degenerate solutions in PARAFAC2 models based on Tucker's congruence coefficient. The procedure was tested on real GC-MS and HPLC-DAD data sets and could almost completely remove the occurrence of degenerate solutions and drastically reduces the computation time required to achieve optimal solutions.

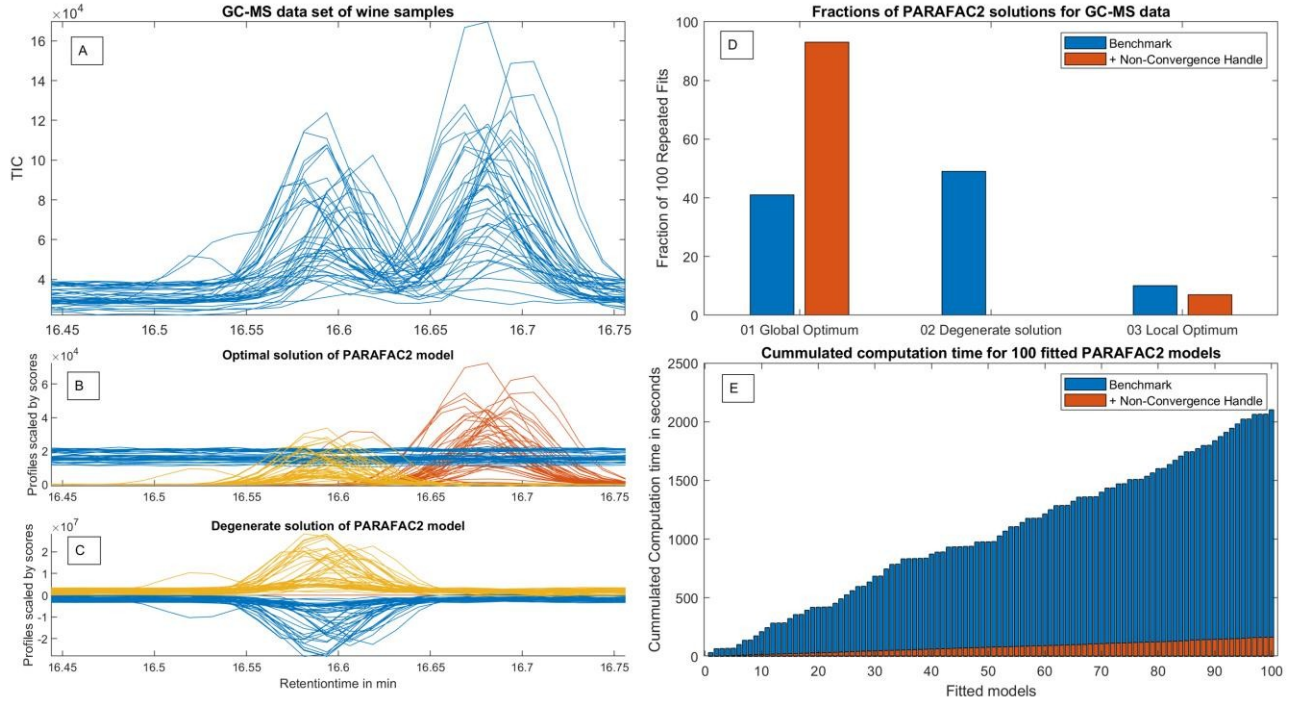


Figure 1 – **A**: Interval of GC-MS data set of wine samples [7], **B**: Elutionprofiles of an optimal 3 component PARAFAC2 model, **C**: Elutionprofiles of a degenerate solution of a 3 component PARAFAC2 model, **D**: Non-Convergence Handle increases the fraction of optimal solutions from 41 % to 93 %, **E**: Non-Convergence Handle reduces the computation time for 100 fitted 3 component PARAFAC2 models by 91 % for the GC-MS data set shown in **A**

References

[1] R. A. Harshman, UCLA Working Papers Phonet. **22**. (1972)
[2] J. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, TrAC. **27**. (2008) 714-725
[3] G. Baccolo, B. Quintanilla Casas, S. Vichi, D. Augustijn, R. Bro, TrAC. **145.** (2021)
[4] R. Bro, Chemometrics and Intelligent Laboratory Systems. **38** (1997) 149-171
[5] H. Yu, R. Bro, Chemometrics and Intelligent Laboratory Systems, **219** (2021)
[6] J.E. Cohen, R. Bro, Latent Variable Analysis and Signal Separation, **10891** (2018) 89–98
[7] T. Skov, D. Balabio, R. Bro, *Analytica Chimica Acta*, **615** (2008) 18-29

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

COUPLED FACTORIZATION OF FLUORESCENCE DATA OF PROTEINS AND QUANTUM DOTS TO ASSESS THEIR CONJUGATION PROCESS

I. M. A. Viegas*[1,2], R. Bro[3], J. M. Amigo[3,4,5], G. A. L. Pereira[1], C. F. Pereira[1]
[1] *Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, Brazil*
[2] *Danish Offshore Technology Center, Technical University of Denmark, Kgs. Lyngby, Denmark*
[3] *Department of Food Science, University of Copenhagen, Frederiksberg, Denmark*
[4] *IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*
[5] *Department of Analytical Chemistry, University of the Basque Country UPV/EHU, Bilbao, Spain*
iviegas@dtu.dk

Quantum dots (QDs) are semiconductor nanocrystals with an active surface that enables their conjugation with a variety of biomolecules, leading to inorganic-biological hybrid nanoparticles with exceptional optical characteristics from the QDs and the biochemical functions from the conjugated molecules [1]. The bioconjugates are usually characterized by separation-based, scattering, microscopy, spectroscopy, mass spectroscopy, and thermal techniques. Depending on the application type, nanomaterial, and biomolecule, some characterizations may be considered more important than other ones, hence there is not a broad agreement on what the essential characterizations are [2]. Many characterization techniques are laborious, sample-destructive, and/or costly, besides requiring sample preparation procedures that may interfere with the desired properties of the bioconjugates and lead to misinterpretation of the results [2]. In this context, we proposed the coupled factorization of fluorescence excitation-emission matrices (EEMs) measured in two spectral regions, over nine weeks, to extract underlying characteristics of the QDs-protein conjugation and monitor it over time. That originated two tensors: **A** for the spectral range of proteins (herein named "lower spectral range") and **B** for the region of QDs (named "upper spectral range"), which were decomposed separately by PARAFAC into three and two components, respectively for **A** and **B**. Subsequently, those two tensors were jointly decomposed into three components by Advanced Coupled Matrix and Tensor Factorization (ACMTF), which allowed improving the recovered profiles in the shared modes: relative concentrations and time. Another important observation is that the 3rd component recovered by ACMTF was absent in the upper range's excitation and emission profiles, which agrees with the spectral features of the proteins assigned to that component. We concluded that fluorescence spectroscopy associated with multi-way analysis has great potential as a non-destructive, quick technique to extract implicit information about the conjugation of QDs to molecules of biological interest.

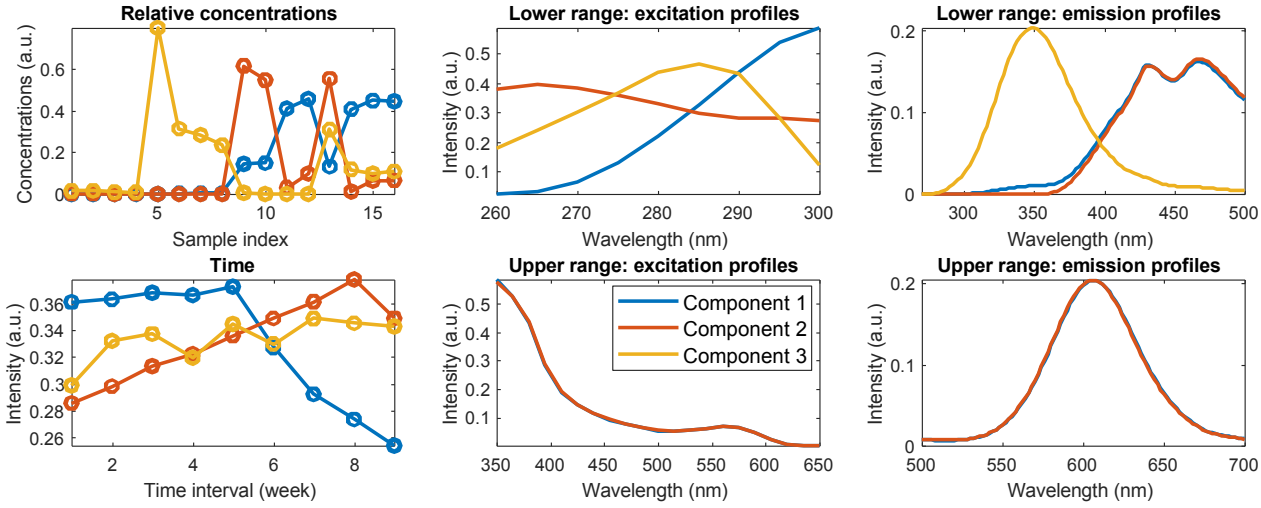


Figure 1 – Profiles recovered by a PARAFAC model fit with joint tensors.

This work was funded by CNPq, CAPES, INCTAA and Instituto Serrapilheira.

References

[1] L. Jing, S. V. Kershaw, Y. Li, X. Huang, Y. Li, A. L. Rogach, M. Gao, *Chem. Rev.* **116** (2016) 10623-10730.
[2] K. E. Sapsford, K. M. Tyner, B. J. Dair, J. R. Deschamps, I. L. Medintz, *Anal. Chem.* **83** (2011) 4453-4488.

New tools for designing food ingredients structures

O. Mykhalevych[1,2], R. Bro[2], A. Benie[1], J. Larsen[1]
[1] *CP Kelco, 4623 Lille Skensved, Denmark*
[2]*Department of Food Science, University of Copenhagen, 1958 Frederiksberg, Denmark*
oksana@food.ku.dk

Carrageenan is a hydrocolloid used as a gelling and stabilizing agent in food and cosmetics. Carrageenan is a very complex polysaccharide, and it has a backbone of 3,6-anhydrogalactose and potassium, sodium, magnesium, and calcium sulphate esters of galactose copolymers that are alternately linked by α-1,3 and β-1,4. The different types differ in molecular structure, e.g., in the 3,6-anhydrogalactose and ester sulphate content. The carrageenan molecule is a very large molecule and contains about 1000 residues. Therefore, there are many structural variations [1]. Structural variations influence functional properties of the carrageenan, i.e., gel strength, texture, solubility, syneresis, synergy, melting and setting temperature [2].
The aim of this project is to provide novel knowledge of the structural characteristics of the carrageenan molecule with regard to its functionality and application, through the development of a data fusion (DF) based model. The model will be applied to predict carrageenans functionality in food and cosmetics.

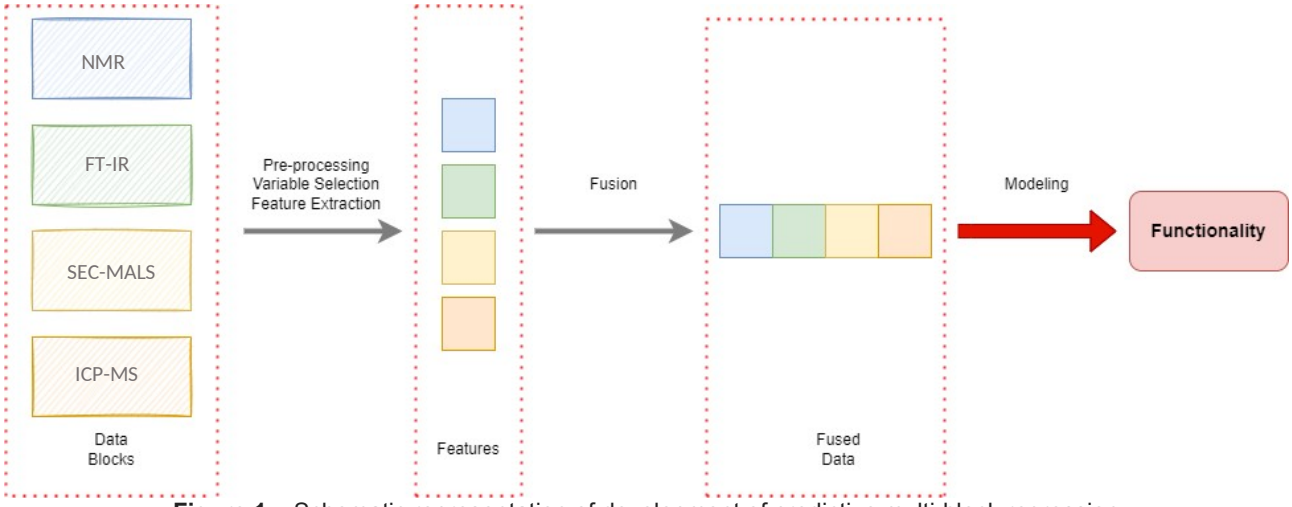


Figure 1 – Schematic representation of development of predictive multi-block regression.

The structure of carrageenan can be characterized by the following parameters: the ion composition, molecular weight distribution, intrinsic viscosity, carrageenan concentration and type, and degree of alkali modification. The mentioned parameters can be determined by four analytical techniques: inductively coupled plasma (ICP), nuclear magnetic resonance (NMR) spectroscopy, Fourier-transform infrared spectroscopy (FT-IR) coupled to Partial Least-Squares (PLS) regression, and size-exclusion chromatography with multi-angle light scattering detection (SEC-MALS). The main hypothesis of this project is that combining all structural data by data fusion methodologies [3] would enable determination of an empirical relationship between chemical structure and functionality of carrageenan. Figure 1 shows a flow diagram of the planned data treatment and model development work.

References

[1] CP Kelco, GENU ® Carrageenan Book, Kelco CP. (2001) 4–23.
[2] A.P. Imeson, Carrageenan and furcellaran, Woodhead Publishing Limited, 2009. https://doi.org/10.1533/9781845695873.164
[3] A. Biancolillo, R. Boqué, M. Cocchi, F. Marini, Data Fusion Strategies in Food Analysis, 2019. https://doi.org/10.1016/B978-0-444-63984-4.00010-7

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

MONITORING POLLUTION PATHWAYS IN RIVER WATER BY PREDICTIVE PATH MODELLING USING UNTARGETED GC-MS MEASUREMENTS

Maria Cairoli[1], André van den Doel[1], Berber Postma[1], Tim Offermans[1], Henk Zemmelink[2], Gerard Stroomberg[1,3], Lutgarde Buydens[1], Geert van Kollenburg[1,4], Jeroen Jansen[1]

[1] *Radboud University, Nijmegen, The Netherlands*
[2] *Rijkswaterstaat, Lelystad, The Netherlands*
[3] *RIWA Rijn, Nieuwegein, The Netherlands*
[4] *Eindhoven University of Technology, Eindhoven, The Netherlands*
maria.cairoli@ru.nl

Heading toward a holistic approach to protect river water quality is among the most compelling needs advocated by the European Water Framework Directive (WFD), one of the most comprehensive European water policies [1]. Non-target screening is an essential building block toward this approach: the analysis of unknowns provides a complete chemical fingerprint of the aquatic ecosystem and favors the identification of chemicals of emerging concern [2]. Tracking chemicals' spatial pathways is equally critical: path modelling is an advanced statistical tool which allows incorporating the spatial dimension into predictive modelling, diagnosing chemicals' origin and dispersion patterns. We propose an integrated approach which couples non-target screening with spatiotemporal path modelling to reveal intrinsic patterns of unknown pollution in the river. We studied a path model for 9 different sites along the Rhine, after extracting characteristic chemical features from their GC-MS measurements with PARAFAC2 [3] (**Figure 1**). We show how path modelling can be employed on untargeted data as a quantitative tool to investigate chemicals' pathways, favouring the prioritization of unidentified chemicals for further investigation. For this study we utilized Process PLS [4], a path modelling method which accounts for multicollinearity and multidimensionality, thus highly suited to analyze the water system complexity and the heterogeneity in the untargeted data. Our approach offers the unique opportunity to implement early-warning strategies in watershed management, complying with the holistic need outlined by the WFD.

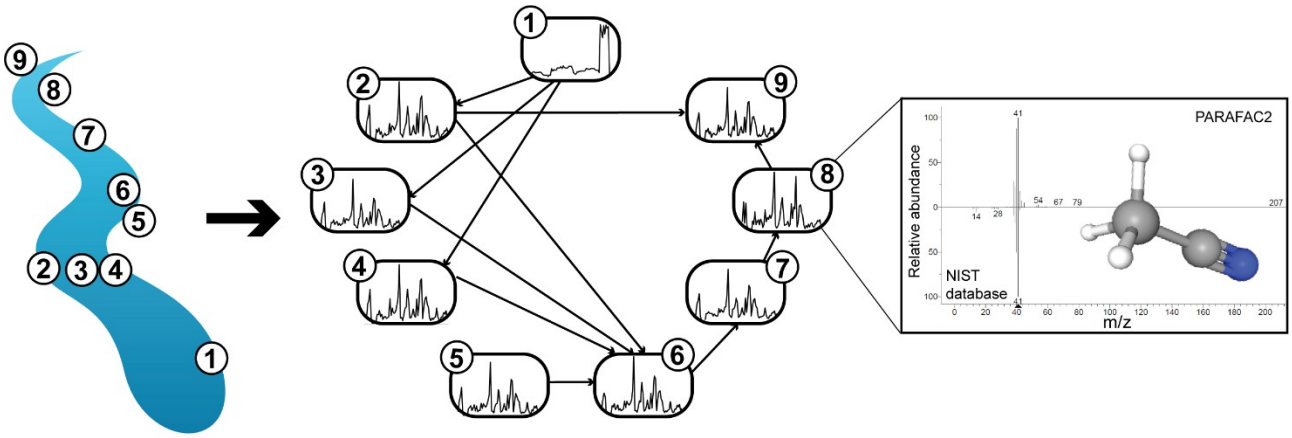


Figure 1 – Schematic representation of path modelling for chemical prioritization and tentative identification.

References
[1] Voulvoulis N., Arpon K.D., Giakoumis T., *Sci. Total Environ* **575** (2017) 358–366.
[2] Reichenbach S.E., Tian X., Cordero C., Tao Q, *J. Chromatogr. A* **1226**, (2012) 140-148
[3] Johnsen L.G., Skou P.B., Khakimov B., Bro R., *J. Chromatogr. A* **1503**, (2017) 57–64.
[4] van Kollenburg G., Bouman R., Offermans T., Gerretzen J., Buydens L., van Manen H.J., Jansen J., *Comput. Chem. Eng.* **154**, (2021) 1–29.

Acknowledgements
This project is co-funded by TKI-E&I with the supplementary grant 'TKI-Toeslag' for Topconsortia for Knowledge and Innovation (TKI's) of the Ministry of Economic Affairs and Climate Policy. The authors thank all partners within the project 'Measurement for Management (M4M)' (https://ispt.eu/projects/m4m/), managed by the Institute for Sustainable Process Technology (ISPT) in Amersfoort, the Netherlands.

FLUORESCENCE AND SCATTERING MODEL ESTIMATION

I. Krylov[1], T. Labutin[1], Å. Rinnan[2], R. Bro[2]
[1]*Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia*
[2]*Department of Food Science, University of Copenhagen, Frederiksberg, Denmark*
ikrylov@laser.chem.msu.ru

Canonical tensor decomposition (also known as PARAFAC) is well matched to the underlying physical model of fluorescence signal and is frequently used to model fluorescence excitation-emission matrices (EEMs). However, EEMs typically contain scattering signal, which has to be handled separately as it doesn't fit the assumptions of the PARAFAC model.
Use of missing data in place of the scattering areas and their interpolation [1] are two widely used methods. In particular, interpolation makes it possible to avoid the local minima and convergence problems resulting from use of the PARAFAC model with a lot of missing data. On the other hand, interpolation may discard potentially useful information or result in artefacts, e.g. in case when second diffraction order scattering band overlaps with a fluorescence peak.
In this work, a combination of PARAFAC and a bilinear model is suggested in order to model fluorescence and scattering signals together, similar to multivariate curve resolution (MCR) with trilinear constraints [2]. Currently, the model is fitted by alternating between the PARAFAC step and a constrained matrix factorisation step. Implementations in MATLAB and R programming languages are available.

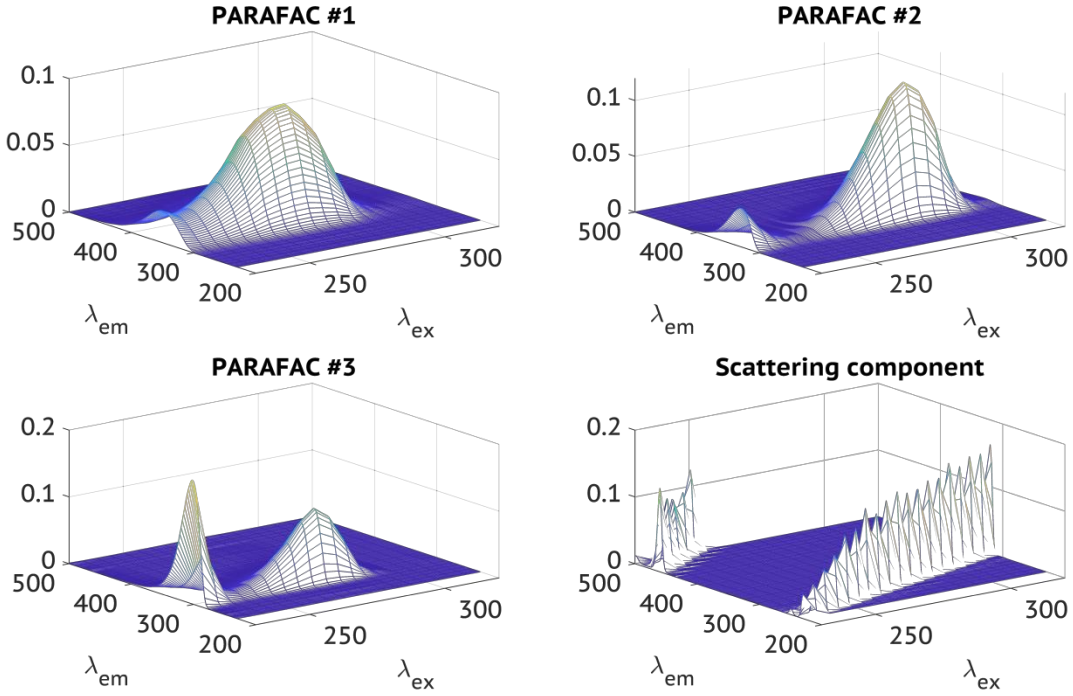


Figure 1 – Results of the decomposition of a subset of Fluorescence data measured by Åsmund Rinnan and Jordi Riu.

The reported study was funded by the Russian Foundation for Basic Research according to the research project No. 20-33-90280.

References

[1]. M. Bahram, R. Bro, C. Stedmon, A. Afkhami. *Journal of Chemometrics*. **20** (2006), 99–105.
[2]. R. Tauler, I. Marqués, E. Casassas. *Journal of Chemometrics*. **12** (1998), 55–75.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

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

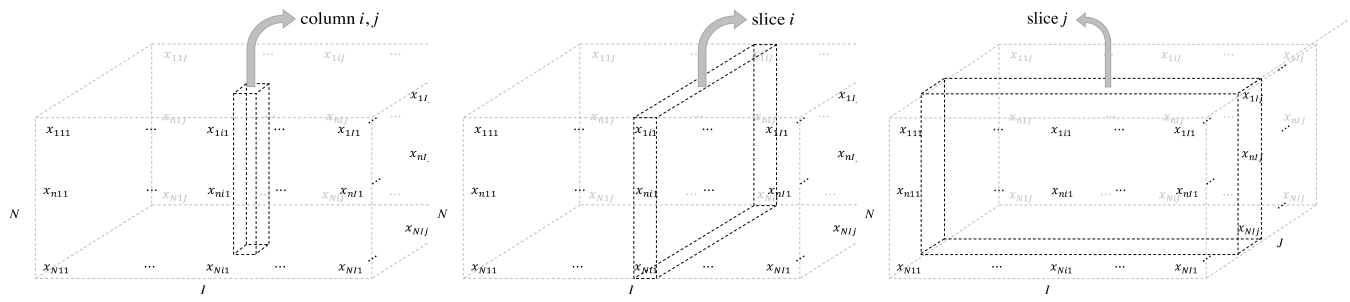
A. BIANCOLILLO [1], J.M. ROGER[2], F. MARINI[3]
[1]Department of Physical and Chemical Sciences, University of L'Aquila, L'Aquila, Italy
[2]ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France
[3]Department of Chemistry, University of Rome "La Sapienza", Rome, Italy
jean-michel.roger@inrae.fr

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.



[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.

Using Multi-Block Non-Negative Matrix Factorization for Multi-layer Plastic Sorting

M. Ghaffari[1], G. H. Tinnevelt[1], M. van Eijk[2], S. Podchezertsev[2], J.J. Jansen[1]
[1]*Radboud University, Institute for Molecules and Materials, Analytical Chemistry, P.O. Box 9010, 6500, GL, Nijmegen, the Netherlands*
[2]*National Circular Plastics Test Centre (NTCP), Germanydreef 7, 8447SE Heerenveen, the Nettherlands*
Mahdiyeh.ghaffari@ru.nl

Multilayer plastics are widely employed to improve the functional properties of packaging i.e. thickness of packaging, mechanical strength, and heat tolerance. On one hand, 26% of the flexible packaging market is multilayer plastic packaging. On the other hand, one of the main challenges in plastic sorting is the detection, identification, and separation of multilayer packaging. Although some companies succeeded in the post-industrial multilayer packaging sorting, the available technologies are limited to some particular polymer types [1].

In recent years, automated sorting of plastic packaging significantly increased thanks to technological improvements, especially ones based on Near Infrared-Hyperspectral Imaging (NIR-HSI) [1]. HSIs are non-destructive and fast with minimum sample preparation steps which help for the identification of single/multilayer plastic streams integrated with pattern recognition and/or curve resolution techniques.

In this contribution, a Multi-Block Non-negative Matrix Factorization model (MB-NMF) [2] is conducted for the identification of single/multilayer plastics. In the proposed strategy, the recorded HSI of single/multilayer plastics is jointly analyzed using Multi-Block-NMF under predefined constraints to tackle the possible collinearity of concentration contribution maps of polymers in the multilayer block. For this, augmented HIS images are analyzed by MB-NMF and the results are present in figure 1. The first two sub-matrices in figure 1 are hyperspectral images of two different polymers (Polypropylene and Polyethylene). However, the last image contains multilayer packaging and is made of both Polypropylene and Polyethylene. Joint analysis of HSIs with zero-region constraints together with selecting appropriate spectral domain resulted in the accurate unraveling of hyperspectral images and correct identification.

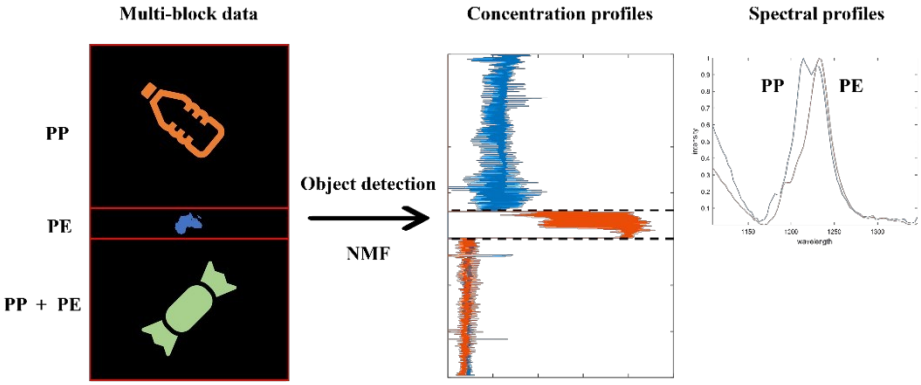


Figure 1 – Using MB-NMF for the analysis of HSI data sets for multilayer plastics sorting. The left panel shows recorded HSIs (PP; PE; PP/PE). The middle panel shows the unfolded concentration contribution map of PP and PE. The right panel indicates the spectral profile of identified polymers.

References

[1] Chen, X. et. al., Determination of the Composition of Multilayer Plastic Packaging with NIR Spectroscopy, Detritus, 2020.
[2] Lee, D. D., et. al., Learning the parts of objects by non-negative matrix factorization, letters to nature, 1999.

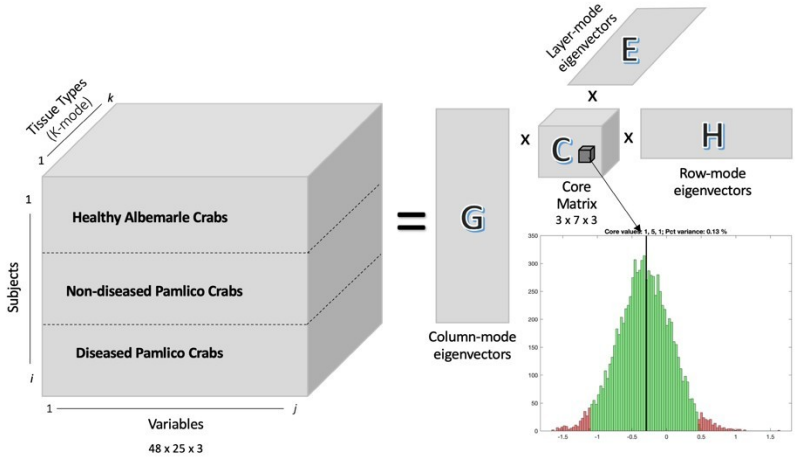
CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Combining ASCA and Tucker3 models to explain high-dimensional data

P. Gemperline[1], S. Hugelier[2], F. Koleini[1], H. Abdollahi[3] M. Akbari Lakeh[3] F. Maddahi[3]
[1]*East Carolina University, Greenville, NC, USA*
[2]*University of Pennsylvania, Philadelphia, PA, USA*
[3]*Institute for Advanced Studies in Basic Science, Zanjan, Iran*
gemperlinep@ecu.edu

We report the use of ANOVA simultaneous component analysis (ASCA) [1] and Tucker3 modeling [2] to analyze multivariate data with an underlying experimental design. By comparing the spaces spanned by different model components we show how the two methods can be used for confirmatory analysis and provide complementary information. ASCA is used to determine the statistical significance of experimental factors and their interactions in a blue crab data set [3]. We demonstrate the novel use of ASCA to analyze the residuals of Tucker3 models and determined that the original 4x5x2 model [3] was insufficient to fully describe the experimental factors in the data set. Increasing the model complexity to 3x7x3 factors removed the last remaining ASCA detectable structure in the residuals. Bootstrap analysis of the core matrix values of the 3x6x3 model compared to the 3x7x3 model showed that one additional triad of eigenvectors was needed to describe the remaining structure in the residuals.

We developed a new simple, novel strategy for aligning Tucker3 bootstrap models with the Tucker3 model of the original data so that eigenvectors of the three modes, the order of the values in the core matrix, and their algebraic signs match the original Tucker3 model without the need for complicated bookkeeping strategies or performing rotational transformations [4]. Concerned that the 3×7×3 Tucker3 model with 63 core values was overparameterized, we used the bootstrap method to determine 95% confidence intervals of the loadings and core values. Important variables for clustering were identified by inspection of loading confidence intervals. We found that 21 of the 63 core values were statistically equivalent to zero at the 95% confidence level and collectively only contributed 0.22% of the variance explained by the model, suggesting that the 3x7x3 model is parsimonious. While triads of eigenvectors represented by these 21 core values do not contribute significantly to the model variance, it was not possible to eliminate them because they share needed combinations of eigenvectors in one or two of the other modes that explain a significant amount of variance in the data. Because the core matrix in our Tucker3 models is three-way orthogonal, constraining even the smallest of these core values to zero produces a completely different model structure, making comparative visualizations difficult.



Considering that tensor decompositions are becoming increasingly important strategies for variable reduction in multiway biomarker and bioinformatic studies, the above methods offer a new reliable strategy for selection of model complexity in tensor decompositions.

Figure 1: Tucker3 model with a sample histogram of a core element. The data set is elemental analysis of blue crab muscle, hepatopancreas, and gill tissue by ICP-AES, 48 crabs x 25 elements x 3 tissue types [3].

1. Liland, K.; Smilde, A.; Marini, F.; Næs, T; Confidence ellipsoids for ASCA models based on multivariate regression theory. *J. Chemom.* **2018**; *32*, 113-125.
2. Kroonenberg, P. *Applied Multiway Data Analysis*, Wiley-Interscience, 2008.
3. P.J. Gemperline, P.; Miller, K; Li, S.; Bray, J.; West, T; Principal component analysis, trace metals, chemometrics and blue crab shell disease. *Anal. Chem.* **1992**, 64, 523A-531A.
4. Kiers, H.; Bootstrap confidence intervals for three-way methods. *J. Chemom.* **2004**, 18, 22-36.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

DATA FUSION BASED ON SELF-ORGANIZING MAP ALGORITHM FOR THE INTEGRATION OF DIFFERENT SOURCE/FREQUENCY INSTRUMENTAL DATA AND SPOT SAMPLING CONTEXTUALIZATION FOR ENVIRONMENTAL MONITORING

S. Licen[1], E. Greco[1], S. Cozzutto[2], V. Bandelj[3], P. Barbieri[1]
[1] *Dept. of Chemical and Pharmaceutical Sciences, University of Trieste, Trieste , Italy*
[2] *ARCO SolutionS s.r.l., spin-off company of the University of Trieste, Trieste , Italy*
[3] *National Institute of Oceanography and Experimental Geophysics-OGS, Trieste, Italy*
slicen@units.it

The evaluation of the spatial and temporal distribution and dynamic variation of pollutants is an important issue to assess the anthropogenic burden on the environment. Modern analytical techniques with a high level of automation allow to process several samples for multi-pollutant analysis purposes in a short time interval. Moreover real-time or quasi real-time instruments/sensors allow to collect high frequency data[1,2]. The integration of the collected information for interpretation is usually challenging considering the overall amount of data and their different collection frequency, especially for long term environmental monitoring (several months).

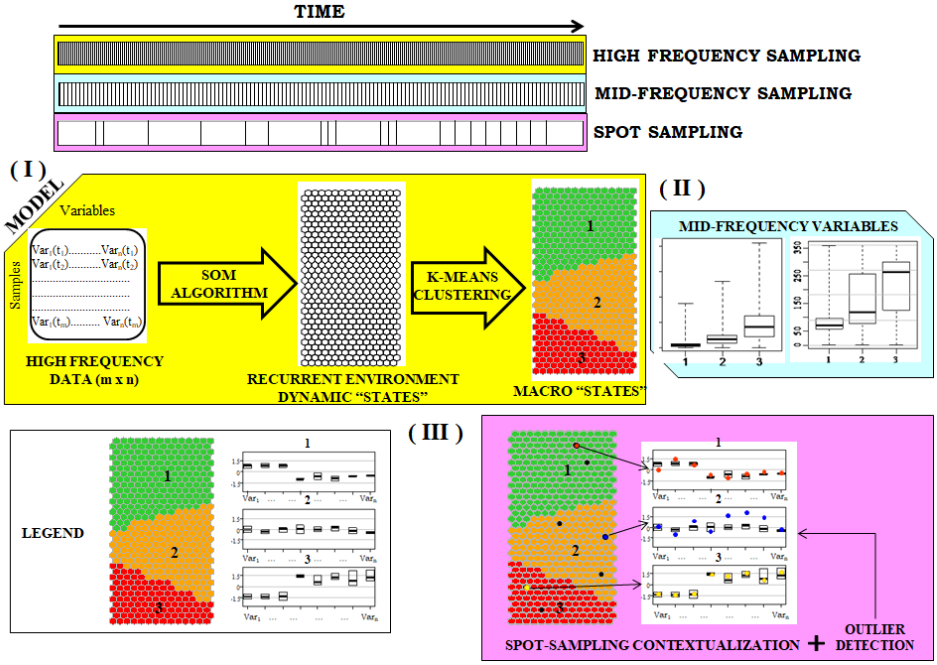


Figure 1 – Data fusion steps (I – II – III)

We propose a data fusion approach based on Self-Organizing Map (SOM) neural network algorithm [3] and 2 nd level abstraction by k-means clustering with the following steps: (I) the high frequency data are mined by the algorithms for obtaining recurrent environment "states" (by SOM) and "macro-states" (by k-means) described by modeled variable profiles; (II) mid-frequency data not used to build the model are related to the "macro-states" to better characterize them; (III) the resulting model is used as a legend to contextualize spot sampling and to detect possible outliers (Figure 1). Results about air impact assessment near an industrial site [4,5] and preliminary results concerning sea water monitoring in the Adriatic Sea (Italy) will be presented. The data were mined by SOMEnv, a package with a Graphical User Interface that works in R software environment and has several built-in visualization features for high frequency data [6].

References

[1] J. Chapman, V.K. Truong, A. Elbourne, S. Gangadoo, et al.,. Chem. Rev. **120** (2020) 6048
[2] M.F. Dupont, A. Elbourne, D. Cozzolino, J. Chapman, et al. *Anal. Methods* **12** (2020) 4597
[3] T. Kohonen, Self-Organizing Maps, *Springer series in Information Sciences*, (2001).
[4] S. Licen, G. Barbieri, A. Fabbris, S.C. Briguglio, et al. *Sensors Actuators, B Chem.* **263** (2018) 476
[5] S. Licen, S. Cozzutto, G. Barbieri, M. Crosera, et al. *Chemom. Intell. Lab. Syst*. **190** (2019) 48
[6] S. Licen, M. Franzon, T. Rodani, P. Barbieri, *Microchem. J.* **165** (2021), 106181

APPLICATION OF DIFFERENT CHEMOMETRIC APPROACHES FOR MALDI-MSI DATA SET OF HETEROGENEOUS TISSUES. CASE STUDY: PAROTID TUMOUR

V. Caponigro[1], E.Salviati[1], F. Marini[2], M. Grimaldi[1], A. M. D'Ursi[1], E. Sommella[1], P. Campiglia[1]

[1] *Department of Pharmacy, University of Salerno, I-84084 Fisciano, (SA), Italy*
[2] *Dipartimento di Chimica, Sapienza Università di Roma, I-00185 Rome, Italy.*

vcaponigro@unisa.it

MALDI Mass Spectrometry Imaging (MALDI-MSI) has increasingly emerged as a valid tool for early diagnosis. The potential of MALDI-MSI approaches resides in providing the spatial distribution of different biomolecules, which can be exploited for both diagnostic and prognostic purposes [1–3]. However, the application of MALDI-MSI imaging on the heterogeneous cellular composition of tissues such as parotid tissues can bring out difficulties during the data analysis. In fact, two levels of complications may occur. Firstly, due to the diverse histological appearances of tumour lesions, the recognition of cancerous mass, as well as the distinction between tumour types can be difficult in the case of parotid tumour diagnosis. So, it is complicated to identify and label each pixel correctly, avoiding both an error during the model training step and evaluating the results. Secondly, reducing the number of variables and selecting a correct number of pixels representing the tissue is a must for the massive amount of data that this technique generates. Finally, a suitable data pre-processing is required to reduce the influence of individual patient signature and generalise the changes related to cancer cell signalling, disease onset mechanisms and tumour progression.

This study compares different approaches in order to identify the optimal workflow during the analysis of heterogeneous tissues using a case study regarding cancer on parotid tissues. The raw data set was imported and analysed using MATLAB R2021a. Firstly, an exploratory analysis was conducted using Principal Component Analysis (PCA). Then, Partial Least Squares Discriminant Analysis (PLS-DA) was carried out to identify specific alterations between the pathological and healthy groups and to compare the different approaches. Each approach has been investigated at pixel level as well as object-wise for the sake of interpretation. As far as the pixels are concerned, random selection and the use of a mask based on the histological information (after image registration) have been compared. In order to maximise the information related to cancer cell signalling, disease onset mechanisms and tumour progression, different pre-processing and normalisation have been compared including full range, PQN, MSROI [4] and co-localization approaches. The optimal approach presented a 95.00% accuracy at pixels level in cross-validation.

ACKNOWLEDGEMENT: This work was funded by project PIR01_00032 BIO OPEN LAB BOL "CUP" J37E19000050007 to P. Campiglia.

References

[1] N. Ogrinc, C. Attencourt, E. Colin, A. Boudahi, R. Tebbakha, M. Salzet, S. Testelin, S. Dakpé, I. Fournier, Frontiers in Oral Health **3** (2022) 1–11.
[2] S. Mas, A. Torro, L. Fernández, N. Bec, C. Gongora, C. Larroque, P. Martineau, A. de Juan, S. Marco, Talanta **208** (2020) 120455.
[3] S. Meding, U. Nitsche, B. Balluff, M. Elsner, S. Rauser, C. Schöne, M. Nipp, M. Maak, M. Feith, M.P. Ebert, H. Friess, R. Langer, H. Höfler, H. Zitzelsberger, R. Rosenberg, A. Walch, Journal of Proteome Research **11** (2012) 1996–2003.
[4] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, Chemometrics and Intelligent Laboratory Systems **215** (2021) 104333.

COMPREHENSIVE CHEMOMETRIC STRATEGY FOR THE HIGH-THROUGHPUT SCREENING OF IN-LINE SPECTROSCOPIC SENSORS FOR MILK COMPOSITION TRAITS

Ewa Szymanska, Frank Gielens, Emanuela Cavatorta, Santosh Lohumi, Fons Jacobs, Bram Dekkers
FrieslandCampina, Amersfoort, the Netherlands
ewa.szymanska@frieslandcampina.com

Development of a new in-line vibrational spectroscopy application often starts with a small feasibility study in the lab where a potential sensor is tested on a limited set of real samples and when possible calibrated against a reference. What if you would like to perform a bigger study and screen multiple sensors for different composition traits at the same time? How to efficiently perform such high-throughput screening and how to critically evaluate results afterwards?
In our study, a comprehensive strategy combining different chemometric criteria, methods and approaches was developed and used in high-throughput screening of in-line sensors. Screening included four different spectroscopic sensors and more than 20 composition traits in milk. More than 200 milk samples were tested including different milk types and sources: whole milk, skimmed milk, standardized milk from multiple production locations and raw milk from a farm.
Developed chemometric strategy included three main steps: data quality check (both reference and spectral data), data preparation (including spectral pre-processing) and development of calibration models. During data preparation spectral regions were selected based on the state-of-the-art knowledge and data-driven approaches. Performance of calibration models was evaluated not only by standard figures of merit (RMSEP, R2, RPD) but also on consistency of predictions across different cross-validation rounds and different sample types. Finally, practical utility of each sensor for each composition trait was assessed to be used in the evaluation of potential benefits of the specific in-line application.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

COMBINING HYPERSPECTRAL IMAGING DATA WITH CLIMATE DATA TO PREDICT PHYSIOLOGICAL VARIABLES OF GRAPEVINE PLANTS

M. Ryckewaert[1], D. Héran[1], T. Simonneau[2], R. Boulord[2], N. Saurin[3], F. Abdelghafour[1], D. Moura[1], S. Mas-Garcia[1], R. Bendoula[1]

[1]*ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France*
[2]*LEPSE, Univ Montpellier, INRAE, Institut Agro Montpellier, Montpellier, France*
[3]*Pech-Rouge, Univ Montpellier, INRAE, Institut Agro, Montpellier, France*
maxime.ryckewaert@inrae.fr

Digital agriculture driven by new intelligent sensors is one of the main ways to improve farm management. Accessing physiological variables such as transpiration (E) and stomatal conductance (gs) in real time with optical instruments is challenging. These are the privileged variables to detect water stress. In this study, the objective is to evaluate visible-near-infrared spectral imaging data combined with climate data to predict transpiration (E) and stomatal conductance (gs) of grapevine (Vitis vinifiera L.) plants by using Sequentially-Orthogonalized Partial-Least-Square Regression (SO-PLS).

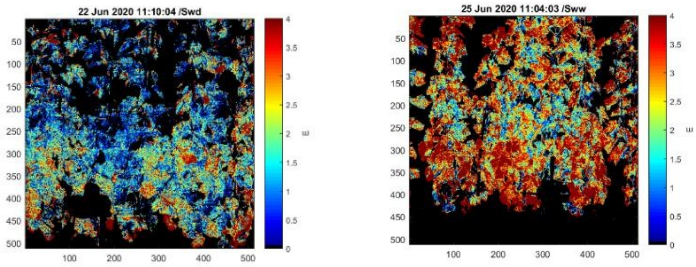


Fig. 1 Prediction of E on hyperspectral images of vines under (left) water deficit condition and (right) well watered condition.

A water stress gradient was obtained using pots of three grape varieties (Syrah, Merlot, Riesling) tested under two water conditions where precise monitoring of physiological variables was performed. Hyperspectral images were acquired and a weather station provided radiation (Rg), relative humidity (RH), temperature (Ta) and wind speed (Ws). For gs, best model is obtained by using only spectral data (R^2= 0.656, bias=8.76, RMSE=64.7 mmol.m².s-1). For E, the best model is obtained by using both blocks (R^2= 0.699, bias=0.055, RMSE=0.614 mmol.m².s-1). While E prediction model has a lower performance using only spectral data (R^2= 0.625, bias=-0.02, RMSE=0.67 mmol.m².s-1). These encouraging results offer prospects to combine data from several sources, including spectral imaging, to detect water stress in grapevines.

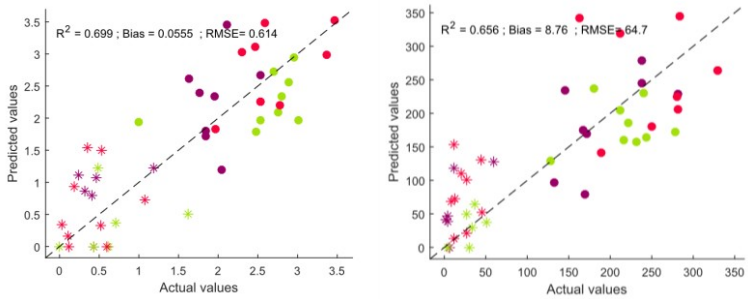


Fig. 2 SO-PLS-model evaluation on the test set of (a) stomatal conductance and (b) transpiration

Reference

M. Ryckewaert, D. Héran, T. Simonneau, F. Abdelghafour, R. Boulord, N. Saurin, D. Moura, S. Mas-Garcia, R. Bendoula, « Physiological variable predictions using VIS–NIR spectroscopy for water stress detection on grapevine: Interest in combining climate data using multiblock method », *Computers and Electronics in Agriculture,* **vol**. (197), p. 106973, juin 2022, doi: 10.1016/j.compag.2022.106973.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Using ATR FT-IR and MCR as a method to understand the crystal state of chocolates tempered under different conditions

E. Ioannidi[1,2], E. Aarøe[2], J. Risbo[1], A. de Juan[3], F.W.J. van den Berg[1]
[1]*University of Copenhagen (dep. of Food Science), Copenhagen, Denmark*
[2]*Aasted Aps, Copenhagen, Denmark*
[3]*University of Barcelona (dep. of Chemical Engineering and Analytical Chemistry), Barcelona, Spain*
eleni.ioannidi@food.ku.dk

The proper and controlled crystallization of cocoa butter is critical during the production of chocolate. Tempering and cooling/ are the two main process steps where cocoa butter fat nucleation and propagation of crystallization takes place [1]. It is important in these steps to achieve the formation of the right crystal form (β V) of sufficient small size to obtain a tight crystal network that will provide proper organoleptic characteristics and long shelf-life to the chocolate. Nowadays, the methods used to evaluate the tempering quality of industrial products (i.e., DSC, temper meter) are not sensitive enough to capture small differences on the crystalline state of chocolate [2]. Fourier Transform Infrared spectroscopy (FT-IR) has been used to study the conformation and lateral packing of the acyl chains of pure TAGs (i.e., POP, SOS) [3–6]. These studies have shown that different crystal states (i.e., γ, α, β', β) of the TAGs present unique spectral profiles that express a disordered or ordered acyl chain conformation and subcell packing.

In our study we explored the use of FT-IR by Attenuated Total Reflection (ATR) for the evaluation of the crystal state of dark chocolates tempered under different conditions and after different periods of storage (i.e., 1 day and 2 months). Since cocoa butter is a complex multiphase system of mixed TAGs, the acquired spectra were further processed with Multivariate Curve Resolution (MCR), a method that can explain complex chemical systems and find the dominant spectral profiles (S) and their related contributions (C) [7]. In our study, a variant of MCR, including a dedicated hard-modeling constraint, has been applied [8]. The hard model implemented in the algorithm is the Avrami equation, designed specifically to describe parametrically the crystallization process undertaken by chocolate during cooling. Such an implementation has allowed a hybrid hard- and soft-modeling of a multiset, where observations acquired during the cooling time range and modelled according to Avrami's equation are coupled with blocks of spectra obtained after a certain period of storage.

For our MCR analyses we analyzed the spectral region 1800-1700cm^{-1} that represents the C=O vibrational modes of the TAG chain. In general, all chocolates at all stages (under cooling and after storage) showed three absorption bands at 1744, 1735 and 1730cm^{-1}, whose relative intensity changed according to the physical state (from melt/amorphous to solid/crystalline). The MCR model generated two main profiles, one "metastable" that represents a disordered state of the TAGs and one crystalline profile of the ordered state with sharper absorption bands. Both profiles were present in the chocolate after a few minutes of cooling, with the contribution of the crystalline profile increasing after prolonged storage of up to two months. Depending on the tempering procedure, initial spectral differences were observed within the first days, which indicates that some tempering procedures provide ordered crystal network than others. However, the chocolates reached very similar stage after prolonged storage (at 18˚C). This indicates that the TAG acyl chains are rearranging over time in order to reach the most favorable and stable thermodynamic state.

References

1. E. J. Windhab, in *Beckett's Ind. Choc. Manuf. Use* (John Wiley & Sons, Ltd, 2017), pp. 314–355.
2. E. Ioannidi, J. Risbo, E. Aarøe, and F. W. J. Van Den Berg, Food Anal. Methods **14**, 2556 (2021).
3. F. Kaneko, K. Tashiro, and M. Kobayashi, J. Cryst. Growth **198–199**, 1352 (1999).
4. J. Yano and K. Sato, Food Res. Int. **32**, 249 (1999).
5. K. Sato, T. Arlshlma, Z. H. Wanga, K. Ojimab, N. Sagib, and H. Morib, JAOCS, J. Am. Oil Chem. Soc. **66**, 664 (1989).
6. F. Kaneko, K. Oonishi, H. Uehara, and H. Hondoh, Cryst. Growth Des. (2021).
7. A. de Juan and R. Tauler, Anal. Chim.Acta **1145**, 59 (2021).
8. A. de Juan, M. Maeder, M. Martínez, R. Tauler, Chemom. Intell. Lab. Sys, 54 (2000) 123.

HSI-NIR AND CHEMOMETRICS FOR THE QUANTIFICATION OF COLLAGEN IN BONES: HOW CHEMICAL MAPPING CAN HELP IN PRESERVING ARCHEOLOGICAL FINDS.

C. Malegori[1], Giorgia Sciutto[2], Paolo Oliveri[1], Silvia Prati[2], Stefano Benazzi[3], Emilio Catelli[2], Silvia Cercatillo[4], Dragana Paleček[4], Rocco Mazzeo[2], Sahra Talamo[4]
[1]University of Genova, Department of Pharmacy, Viale Cembrano 4, 16148, Genova, Italy
[2]University of Bologna, Department of Chemistry "G. Ciamician", Ravenna Campus, Via Guaccimanni, 42, 48121, Ravenna, Italy
[3] Department of Cultural Heritage, University of Bologna, Via degli Ariani 1, 48121 Ravenna, Italy
[4] University of Bologna, Department of Chemistry "G. Ciamician", Via Selmi 2, Bologna, 40126, Italy
malegori@difar.unige.it

Many rarest bones and human remains in Prehistory are much too precious and considered a cultural and historical patrimony, and so the application of destructive methods -such as 14C- must be as limited as possible. The identification and quantification of collagen in archaeological bones play a crucial role in the 14C dating protocol because a significant amount of proteins (1% yield of collagen) [1] is requested for performing a reliable radiocarbon analysis. The advantages of the HSI-NIR technology can be of high impact in this context, in particular regarding the non-destructiveness and the possibility to map, from a spatial extent, the area of the bone in which the proteins are more present. A first milestone in mapping collagen in bones by means of HSI-NIR was achieved by the authors [2], applying the NDI approach to few archaeological samples. Acting upon this, a quantification model is now proposed thanks to the analysis of 60 archaeological bones, ranging back from the modern age to more than 50,000 years ago. For the development of the PLS regression model, NIR images of both bone powders and fragments were acquired in the spectra range 1000-2500 nm (Specim SWIR 3 camera) and then the samples were submitted to collagen extraction at the BRAVHO 14C lab, following the procedures of Talamo et al. 20211. With a RMSECV of 2.2% wt, the proposed strategy was applied to unknown samples demonstrating how HSI-NIR can be a sustainable pre-screening method to quantify the presence of collagen on bone samples in a non-destructive way.

References

[1] Talamo, Sahra, et al. "Here we go again": the inspection of collagen extraction protocols for 14C dating and palaeodietary analysis. STAR: Science & Technology of Archaeological Research 7.1 (2021): 62-77.
[2] Lugli, F., et al. Near-infrared hyperspectral imaging (NIR-HSI) and normalized difference image (NDI) data processing: An advanced method to map collagen in archaeological bones. Talanta 226 (2021): 122126.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

SPECTROSCOPY AND CHEMOMETRICS FOR SORTING WASTE WOOD MATERIAL ACCORDING TO THE BEST-SUITED APPLICATION

M. Mancini[1,2], Å. Rinnan[1] , V.-M. Taavitsainen[3]

[1]*Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

[2]*Department of Agricultural, Food and Environmental Sciences, Università Politecnica delle Marche, via Brecce Bianche, 60131 Ancona, Italy*

[3]*Department of Computational Engineering, Lappeenranta University of Technology, Skinnarilankatu 34, 53850 Lappeenranta, Finland*

manuela@food.ku.dk

Wood is a highly exploited resource in several sectors (e.g. pulp, construction, energy), but it is also limited. For this reason, waste wood is becoming an appealing alternative material since nowadays a large amount remains unused [1]. By 2030, waste wood is estimated to contribute with 59-67 million m^3 to annual European Union wood demand [2].

There are a bunch of reasons why recovery is important and why waste materials should be reused/recycled if possible. Most waste wood can be reused as a building material or for producing new composite wood material, recycled into pulp for paper production, or used profitably as a biofuel increasing the share of renewable energy production. Only wood with hazardous substances needs to be sent to the disposal with no benefits. Even if wood can be recycled into a variety of useful products and materials, its recycling potential is still low because of the presence of contaminants [3]. In fact, wood is commonly subjected to heat, chemical or mechanical treatments that involve preservatives containing organic and inorganic contaminants. Some of those are hazardous and move the bar in the direction of landfilling instead of the more sustainable and cost-efficient recovery of the material. It is also important to mention that waste wood recycle or reuse could be maximized if the material is properly sorted and handled based on its quality and characteristics. As for other sorting processes, near-infrared (NIR) spectroscopy could represent a valid solution for the rapid screening of the waste wood material and the assessment of its best reuse.

To this aim, more than 100 waste wood samples have been collected in different recycling centers and a panel board company located in Italy and Denmark. The waste wood material has been collected as large pieces of wood in their original form (e.g. items of furniture, fiber board or pallet) and each sample has been coded with the most appropriate waste wood category, according to the suitable end-use. In detail, three waste wood categories have been considered: i) virgin wood that can be used both for panel board production and bioenergy applications, ii) treated wood that can be used for panel board production and iii) impregnated and painted wood (disposal wood) that should be sent to the disposal with no reuse. All the samples have been analysed both in their original particle size and after reduction to about 5 cm of particle size to simulate real sorting process. Spectra have been acquired both with bench-top and a handheld NIR device. Chemometrics has been used for the development of different classification models and evaluating the possibility to separate the waste wood material according to the most suitable reuse, i.e. energy production, panel board production or landfill (no reuse). The results demonstrated as spectroscopy and chemometrics could be a perfect tool for the rapid screening of the waste wood material improving its reuse as a valuable resource for the generation of secondary materials.

References

[1] G. Deroubaix, Optimisation of material recycling and energy recovery from waste and demolition wood in different value chains (FCBA), in 5th WWNet Symposium (2013).

[2] U. Mantau et al., EUwood–real potential for changes in growth and use of EU forests. Final report. Project: Call for tenders No. TREN/D2/491-2008 (2010) 160 p.

[3] G. Faraca, D. Tonini, T.F. Astrup, Dynamic accounting of greenhouse gas emissions from cascading utilisation of wood waste, *Science of The Total Environment*, **651**, (2019) pp. 2689–2700.

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.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

AN ACTIVE CONSTRAINT APPROACH TO IDENTIFY ESSENTIAL SPECTRAL INFORMATION IN NOISY DATA

M. Beese[1,2], M. Sawall[1], C. Ruckebusch[3], R. Francke[1,2], A. Prudlik[1,2], K. Neymeyr[1,2]
[1]*Universität Rostock, Rostock, Germany*
[2]*Leibniz-Institut für Katalyse, Rostock, Germany*
[3]*Université de Lille, Lille, France*
martina.beese@uni-rostock.de

To extract pure component profiles from mixed spectral data multivariate curve resolution (MCR) methods can be used. Often certain rows and columns of the spectral data matrix are essential for the outcome, whereas other rows and columns are of minor importance.
The rows and columns, i.e. the spectra and frequency channels, are represented by data points in the U- or V-space from the singular vectors of the spectral data matrix. The data points are to be classified as essential or non-essential. They represent essential spectra and frequency channels. We address the question how to detect these essential data points. For model data it is easy to use only the vertices of the inner polyhedron from the low-dimensional representation of the data matrix, as known from MCR analysis, but this approach cannot be used for noisy data. Thus, this noisy data is analysed with the goal of minimal computational costs for large data sets.
An algorithm is presented for this purpose. Active nonnegativity constraints in combination with duality arguments are responsible for determining the essential spectral information.
The essential spectral information also reduces the dimension, ideally without the loss of chemical information, which is especially desirable for high-dimensional data. This can speed up MCR analyses. The suggested algorithm is tested for noisy experimental data. An example is given in Figure 1.

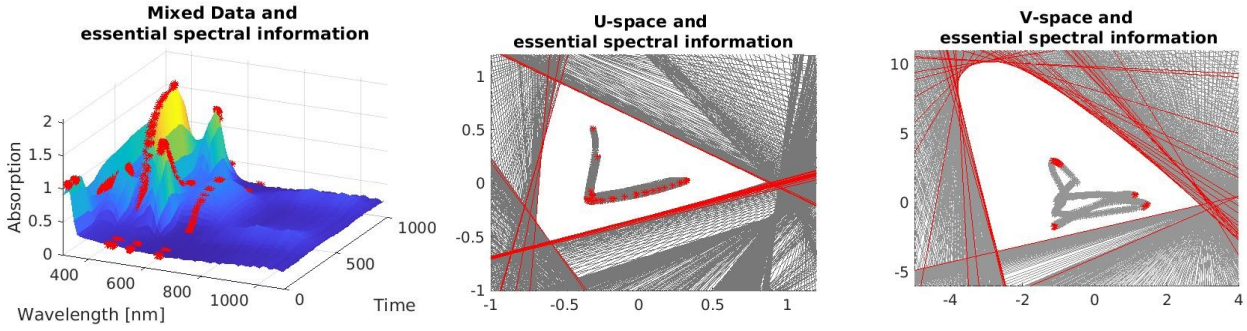


Figure 1 – Spectroelectrochemical data of an anthraquinone system and their representation in the U- and V-space with marked essential spectral information

References

[1] M. Sawall, C. Ruckebusch, M. Beese, R. Francke. A. Prudlik, K. Neymeyr, Submitted (2022)

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

A PHASOR VIEW OF MULTIVARIATE CURVE RESOLUTION

L. Coic, R. Vitale, C. Ruckebusch
[1]*Université de Lille, Dynachem, LASIRE CNRS, Lille, France*
laureen.coic@univ-lille.fr

This work shows that the identification of the purest information encoded in a spectral mixture dataset can be performed more easily and reliably by computing first a phasor transform of the original measurements. A phasor is a two-dimensional polar plot representing the values of the sine-cosine transform of a signal provided by the fast Fourier transformation. Originally, the phasor approach was proposed in electrical engineering as a simple way of reducing the complexities resulting from handling single frequency signals [1] and was later adapted to other fields [2,3].

The phasor transform enables visualizing the information associated to the rows (spectra) of a linear mixture dataset by converting each one of them into a point in a bidimensional plot. This transformation has interesting features to be exploited for robust and automatic multivariate curve resolution. It is fast, it can be performed independently on every spectrum (it does not need to process the whole spectral dataset to evaluate the "score" of a given spectrum) and, most importantly, it is a linear transformation. This means that the phasor of a mixed spectrum is a linear combination of the phasors of spectra of individual species. Thus, under proper normalization, pure(st) spectra can be found at the vertices of a convex geometry [4,5] in a two-dimensional phasor plot, irrespective of the data dimensionality and the number of such individual species.

We illustrate here the results obtained from the processing of a seven-component Raman hyperspectral dataset [6]. The following figure permits to show results of the convex hull calculation in a phasor representation.

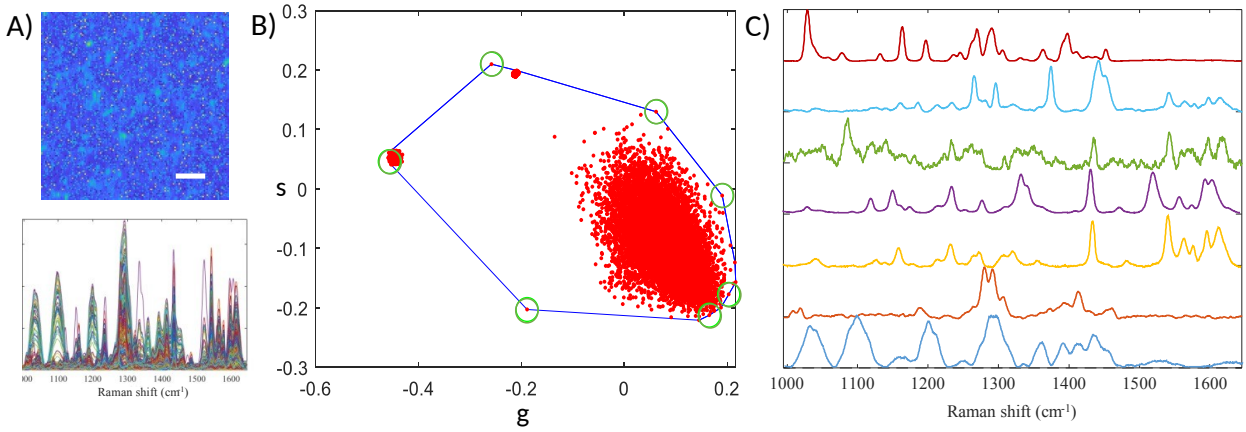


Figure 1 – *A) Raman hyperspectral data (134x134x1600) of a seven-constituent pharmaceutical formulation. B) Phasor representation of all the spectral data (red dots) with their convex hull highlighted in blue and the purest spectral pixels identified by applying SIMPLISMA* circled in green. *C) Profiles of the purest spectral pixels identified*

References

[1] D.I. Crecraft, S. Gergely, in *Analog Electronics: Circuits, Systems and Signal Processing*, 2002
[2] A Clayton, A. Andre, S. Quentin, S. Hanley, P.J. Verveer, *Journal of Microscopy*, 213, 2005.
[3] F. Fereidouni, A. Bader, C. Gerritsen, *Optics Express*, 20, 2012.
[4] M. Ghaffari, N. Omidikia, C. Ruckebusch, *Analytical Chemistry*, 17, 2019
[5] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, N. Omidikia, *Trends in Analytical Chemistry*, 132, 2020.
[6] L. Coic, P.Y. Sacre, A. Dispas, C. De Bleye, C. Ruckebusch, P. Hubert, E. Ziemons, *Analytica Chimica Acta*, 1198, 2022.

Trilinearity in Multivariate Curve Resolution: hybrid modeling and missing data.

Anna de Juan[1] and Adrián Gómez-Sánchez[1,2].
[1]*Chemometrics group, Universitat de Barcelona, Barcelona, Spain*
[2]*LASIRE, Université de Lille, Lille, France*
anna.dejuan@ub.edu

Multivariate Curve Resolution (MCR) is, by nature, a bilinear decomposition method devoted to solve the mixture analysis problem. However, the versatility of the MCR framework allows incorporating trilinear, multilinear or factor interaction models in a multiset analysis context [1].

In difference with multi-way trilinear decomposition methods, the main asset of MCR is the possibility to work in a hybrid model context, where part of the multiset information is modeled in a trilinear way and the rest following the natural MCR bilinear decomposition. Hybrid modeling is possible because the implementation of the constraint enables the *per component* and *per block* selection of the information forced to obey the trilinear model. The *per component* implementation of trilinearity appeared first and was applied in environmental data and analytical measurements [2]. It is more recent the *per block* use of the constraint, which has allowed combining the information of matrices and trilinear tensors in a single multiset. In contrast with approaches such as Combined Tensor and Matrix Factorization (CTMF), where merging this information requires separate factorizations, the *per block* application of trilinearity in MCR allows solving the problem with a single multiset factorization via a hybrid bilinear/trilinear model, where the blocks of the matricized tensor obey the trilinear condition and the matrix blocks follow a bilinear behavior. This methodology has been successfully applied in 3D/4D image fusion problems [3] and in time-resolved fluorescence methods, where exponential signals, easily treated by trilinear slicing methodologies, are affected by non-exponential contributions at short time ranges [4].

A recent advantage of the use of trilinearity in the MCR context is related to the analysis of trilinear data with systematic patterns of missing values, such as those encountered in excitation-emission measurements (EEM) when excitation and emission ranges overlap. In this case, the EEM slices of the data cube have a large proportion of triangular-patterned missing values. In trilinear decomposition methods, the necessary data imputation to solve this problem presents serious limitations. An important fact is that these ragged tensors, when matricized, become a multiset without missing values, with emission spectra of different lengths. A new implementation of trilinearity, applied sequentially to small full sections of the emission spectra information, allows obtaining at the end complete trilinear profiles skipping the limitations of the data imputation step [5].

References

[1] A. de Juan, R. Tauler. (2021). Anal. Chim. Acta, **1145** (2021) 59-78.
[2] M. Alier, M. Felipe, I. Hernández, R. Tauler. Anal. Bioanal. Chem. 399 (2011) 2015-2029..
[3] A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, A. de Juan. Anal. Chem. **92** (2020) 9591-9602.
[4] O. Devos, M. Ghaffari, R. Vitale, A. de Juan, M. Sliwa, C. Ruckebusch. (2021). Anal. Chem., **93** (2021) 12504-12513.
[5] A. Gómez-Sánchez, P. Loza, C. Ruckebusch, A. de Juan (submitted).

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Combining spectral and spatial features extracted from hyperspectral images: Application on the detection of scab disease

Belal Gaci [1,2,3], Florent Abdelghafour [2,3], Maxime Ryckewaert[2,3], Silvia Mas-Garcia [2,3], Marine Louargant 1, Yohana Laloum [1], Ryad Bendoula [2,3], Jean-Michel Roger [2,3].

Belal.gaci@ctifl.fr

[1] CTIFL, France
[2] ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France
[3] ChemHouse Research Group, Montpellier, France

The ability to detect and map early infectious outbreaks in orchards with non-destructive methods would constitute a substantial advantage in crop protection [1]. In this context, hyperspectral imaging has proven to be a promising tool. In order to better exploit the richness of the data, a new method is proposed. It consists in exploiting jointly spatial and spectral properties of the hyperspectral images with a data fusion method [2]. To apply this method, 16 hyperspectral (111*169*256) SWIR images [1000 nm-2500nm] of healthy and scab infected apple leaves were used. For the spatial features, the first step is to extract a monochromatic image. It is achieved by projecting the hypercubes on the loadings of a PCA performed on all the calibration images. On this image, the spatial features are estimated on sub-images of 3*3 px on infected and healthy parts of the limbus. The features are the Haralick's indexes computed from grey level cooccurence matrixes. The second step for spectral features, consists in a set of Principal components resulting from a non centered SVD performed in each sub-image. The third and last step is to model jointly both type of features with a multiblocks method (ROSA-FDA algorithm). The proposed method results in accurate discrimination of healthy and infected plants with 8% of errors for the infected spot and 14% of error for the healthy part of the leaves. These results outperform the purely spectral approach (14% of error infected spot, 13% healthy spot).

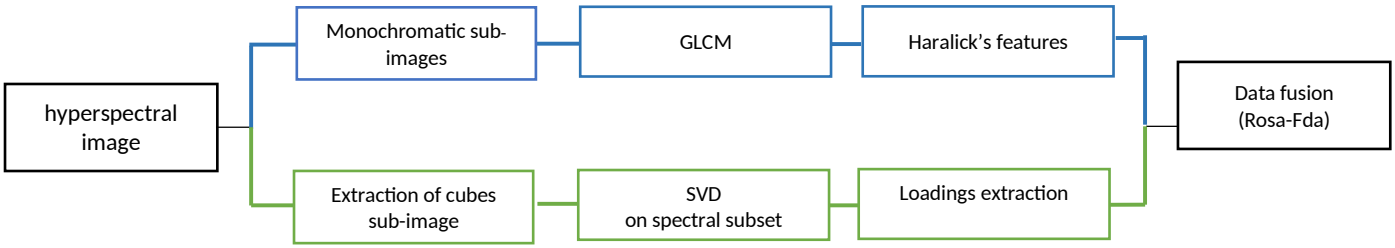


Figure 1 : Workflow of the methodology

References:

[1] Meunkaewjinda, A., Kumsawat, P., Attakitmongcol, K. & Srikaew, A. Grape leaf disease detection from color imagery using hybrid intelligent system. 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology vol. 1 513–516 (2008).
[2] Liland, K. H., Næs, T. & Indahl, U. G. ROSA—a fast extension of partial least squares regression for multiblock data analysis. Journal of Chemometrics 30, 651–662 (2016).
[3] Nouri, M. et al. Near infrared hyperspectral dataset of healthy and infected apple tree leaves images for the early detection of apple scab disease. Data in Brief 16, 967–971 (2018).

2-D WAVELET IMAGE DECOMPOSITION AND MULTIVARIATE STATISTICAL PROCESS CONTROL FOR BLENDING END-POINT DETECTION

R. Rocha de Oliveira[1], M. Cocchi[2], A. de Juan[1]
[1]*Chemometrics group, Universitat de Barcelona, Spain*
[2]*Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Modena, Italy*
rodrigo.rocha@ub.edu

A critical aspect during the industrial production of solid mixtures is the assessment of heterogeneity either during the blending processes or in the final product. The use of process analytical technology (PAT) tools based on chemical imaging systems, such as near-infrared hyperspectral images (NIR-HSI), are useful to assess heterogeneity information during mixing processes, because they provide chemical and spatial information about samples, i.e., they tell about sample spatial composition [1]. Still, when handling chemical images, common chemometric approaches usually overlook this spatial information.

This work presents a new PAT chemometric tool to take advantage of the spatial information from chemical images for blending process endpoint detection. The strategy is based on a first step linked to HSI unmixing analysis [2], which provides distribution maps that offer a good qualitative visual representation of the evenness in the spatial distribution of every mixture ingredient in the image collected during the blending process. From these distribution maps, an extraction of spatial features is carried out with 2D wavelet decomposition [3]. The last step consists of the use of the wavelet spatial features as input to a multivariate statistical process control (MSPC) model to obtain statistical indicators, such as the *Q*-residuals or Hotelling's T^2, and test whether the blending process has reached the endpoint, i.e. the statistical indicator are below the control charts limits. The MSPC model and statistical control limits are previously built based on features extracted from distribution maps at the blending endpoint of acceptable batches.

The methodology has been tested for the blending process endpoint detection of a pharmaceutical formulation monitored atline with a pushbroom NIR-HSI system. The results obtained have proven that this new PAT tool is a powerful methodology allowing the detection of blending endpoint based on the spatial information extracted from chemical images. This strategy can be easily adapted to the control of blending processes continuously monitored with any kind of machine vision system or instrumental technique that can provide spatially resolved responses.

References

[1] J.M. Amigo, ed., Hyperspectral imaging, in: Data Handl. Sci. Technol., Elsevier, 2020.
[2] A. de Juan, Multivariate curve resolution for hyperspectral image analysis, in: Data Handling in Science and Technology. Vol. 32. Elsevier, 2020. 115-150.
[3] M. Li Vigni, J.M. Prats-Montalban, A. Ferrer, M. Cocchi, Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA), J. Chemom. 32 (2018) 10–28.

Modelling and preprocessing of sparse infrared spectra

V. Tafintseva, M. Aledda, A. Kohler, N. Patel, B. Zimmermann, V. Shapaval
Faculty of Science and Technology
Norwegian University of Life Sciences, 1430 Ås, Norway
valeria.tafintseva@nmbu.no

The development of Fourier Transform Infrared (FT-IR) spectroscopes in 1950s gave rise to the development of pre-processing and modelling approaches to analyze spectroscopic data. Data from conventional FT-IR instruments are broadband spectra comprising thousands of highly collinear variables. Nowadays, there exists a number of methods for the preprocessing and modelling of the FT-IR broadband spectra. Different preprocessing approaches range from basic baseline corrections, peak normalizations to model-based preprocessing methods such as Multiplicative Signal Correction (MSC) [1]. Among modelling approaches, standard chemometrics method such as Partial Least Squares Regression (PLSR) is known to perform well on the broadband FT-IR spectral data [2]. The strength of the method is that it reduces dimensionality of the variable space and transforms it into much smaller dimensional space of latent variables. Methods like Random Forest (RF) is also known to work well on the broadband FT-IR spectra [3], however the collinearity of the variables makes the method difficult to use when variable selection or interpretation is of a high importance for the analysis.

Modern IR devices such as light-emitting diodes (LEDs) and quantum cascade lasers (QCLs) produce data with limited number of wavenumber channels which we will refer to as sparse IR spectral data. The increasing interest in the photonics solutions in food industries and other fields creates a need for preprocessing and modelling approaches for the sparse data produced by the devices. This is a challenge since only a few spectral variables are available for the analysis.

In this study we focus on selecting the best method for the analysis of the sparse data: both preprocessing and modelling. We compare PLSR, RF and standard Multiple Linear Regression (MLR) methods to find the optimal modelling algorithm. We employ model-based pre-processing approaches like MSC, baseline corrections and peak normalization as well as raw spectra where appropriate. To improve modelling results, we suggest an approach to increase the number of the spectral wavelengths which has direct application for photonics solutions' architecture. The approach improves dramatically the performance of the models even for the very limited number of spectral wavelengths - three to five.

Both regression and classification problems are considered in the study. Different datasets are used to show the results. To mimic photonics data, the sparse data were obtained by reducing broadband FTIR spectra comprising several thousand spectral variables into datasets comprising only few (three to nine) spectral variables. The results of the modelling are compared for the sparse spectral data, for selected wavelength regions and the full spectral range available in the infrared.

References

[1] H. Martens, E. Stark, *J. Pharm. Biomed. Anal*. **9** (1991), 625.
[2] S. Wold, H. Martens, H. Wold, *in: Matrix Pencils: Lecture Notes in Mathematics,* (Eds: B. Kagström and A. Ruhe), Springer, Berlin, Heidelberg (1983) p. 286.
[3] V. Tafintseva, E. Vigneau, V. Shapaval, V. Cariou, E.M. Qannari, A. Kohler, *J. Biophotonics*, **11**, (2018).

TRACING THE IDENTITY OF MOUNTAIN PRODUCT PARMIGIANO REGGIANO PDO CHEESE USING ^{1}H-NMR SPECTROSCOPY AND MULTIVARIATE DATA ANALYSIS

N. Cavallini[1], P.P. Becchi[2], C. Durante[2], L. Strani, V. Pizzamiglio[3], S. Michelini[3], F. Savorani[1], M. Cocchi[2]

[1]*Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italia*

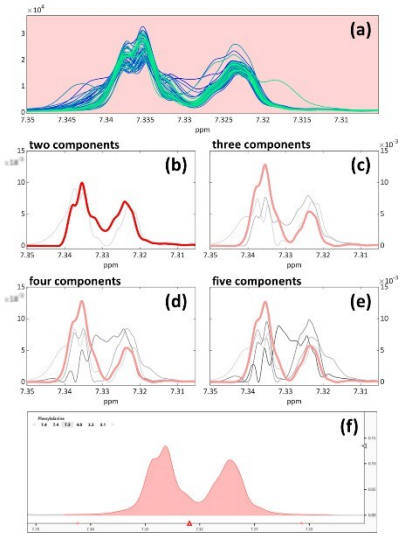
[2]*Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi 103, 41125 Modena, Italia*

[3]*Consorzio Formaggio Parmigiano Reggiano, via Kennedy 50, 42124, Reggio Emilia, Italia*

nicola.cavallini@polito.it

Parmigiano Reggiano cheese is one of the most appreciated and famous Italian cheeses and even if its name is usually associated with a single product idea, different varieties can be found. In particular, the product "Prodotto di Montagna - Progetto Qualità Consorzio ("Mountain Product - Consortium Quality Project") represents a quality denomination for Parmigiano Reggiano PDO (Protected Designation of Origin) cheese that must comply with rather strict rules about its aging, geographical origin and the cow feed and breeding.

In this scenario, there has been an increasing request from both dairy farmers and consortia to protect the authenticity of Mountain Product Parmigiano Reggiano PDO from analogues, and to promote it as a higher quality product. To this aim, comprehensive analytical techniques can provide objective quality and identity assessments and proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy can be used as a tool for metabolic fingerprinting of milk and its derivatives [1-2]. NMR spectroscopy can provide huge amounts of information directly related to many metabolites with a single analytical run, and it can be therefore used for the identification of sugars, small organic acids, vitamins, nucleotides, and aromatic compounds. Due to the NMR signals complexity, multivariate data analysis is needed to interpret and extract information from this type of data, generally leading to optimal results in food characterization and authenticity assessments [3–4].



In this study the metabolic profile of "Mountain Product - Consortium Quality Project" and conventional Parmigiano Reggiano PDO samples were analyzed by means of ^{1}H-NMR spectroscopy, with the aim of finding information useful to distinguish the two denominations. To this aim, two different data analysis approaches were employed: the full spectra dataset (i.e., without any compression) was compared with a "features dataset" obtained by applying Multivariate Curve Resolution (MCR, [5]) to carefully defined small intervals. The extracted features were compared and matched with literature and reference libraries to obtain a putative identification of the resolved compounds/metabolites.

Figure 1 – Chemical identification of MCR-resolved components (b–e) by comparison with a reference library (f).

References

[1] P. Scano, E. Cusano, P. Caboni, R. Consonni, *Int. Dairy J.*, **90** (2019) 56–67

[2] Celso F. Balthazar, Jonas T. Guimarães, Ramon S. Rocha, Tatiana C. Pimentel, Roberto P.C. Neto, Maria Inês B. Tavares, Juliana S. Graça, Elenilson G. Alves Filho, Mônica Q. Freitas, Erick A. Esmerino, Daniel Granato, Sueli Rodrigues, Renata S.L. Raices, Marcia C. Silva, Anderson S. Sant'Ana, Adriano G. Cruz, *Trends Food Sci. Technol.*, **108** (2021) 84-91

[3] N. Cavallini, F. Savorani, R. Bro, M. Cocchi, *Molecules*, **26**(5) (2021) 1472

[4] L. Laghi, G. Picone, F. Capozzi, *TrAC*, **59** (2014) 93–102

[5] A. De Juan, J. Jaumot, R. Tauler, *Anal. Methods*, **6** (2014) 4964–4976

A COMBINED CHEMOMETRIC STRATEGY FOR A NON-DESTRUCTIVE AGE ESTIMATION OF BIOLOGICAL FLUID STAINS

P. Oliveri[1], C. Malegori[1], C. Manis[2], E. Alladio[2,3], M. Vincenti[2,3], P. Garofano[3], F. Barni[4], A. Berti[4]
[1]*Dipartimento di Farmacia (DIFAR), Università degli Studi di Genova, Genova, Italy*
[2]*Dipartimento di Chimica, Università degli Studi di Torino, Torino, Italy*
[3]*Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Orbassano, Italy*
[4]*Reparto Carabinieri Investigazioni Scientifiche, Sezione di Biologia, Roma, Italy*
paolo.oliveri@unige.it

From a criminalistic point of view, the accurate dating of biological traces found at the crime scene, together with its compatibility with the estimated crime perpetration timeframe, enables to limit the number of suspects by assessing their alibis and clarifying the sequence of events. The present study delineates the possibility of dating biological fluids such as semen and urine, as well as blood traces, by using a non-destructive analytical strategy based on hyperspectral imaging in the near infrared region (HSI-NIR), coupled an integrated strategy that combines object-based image processing and multivariate regression.

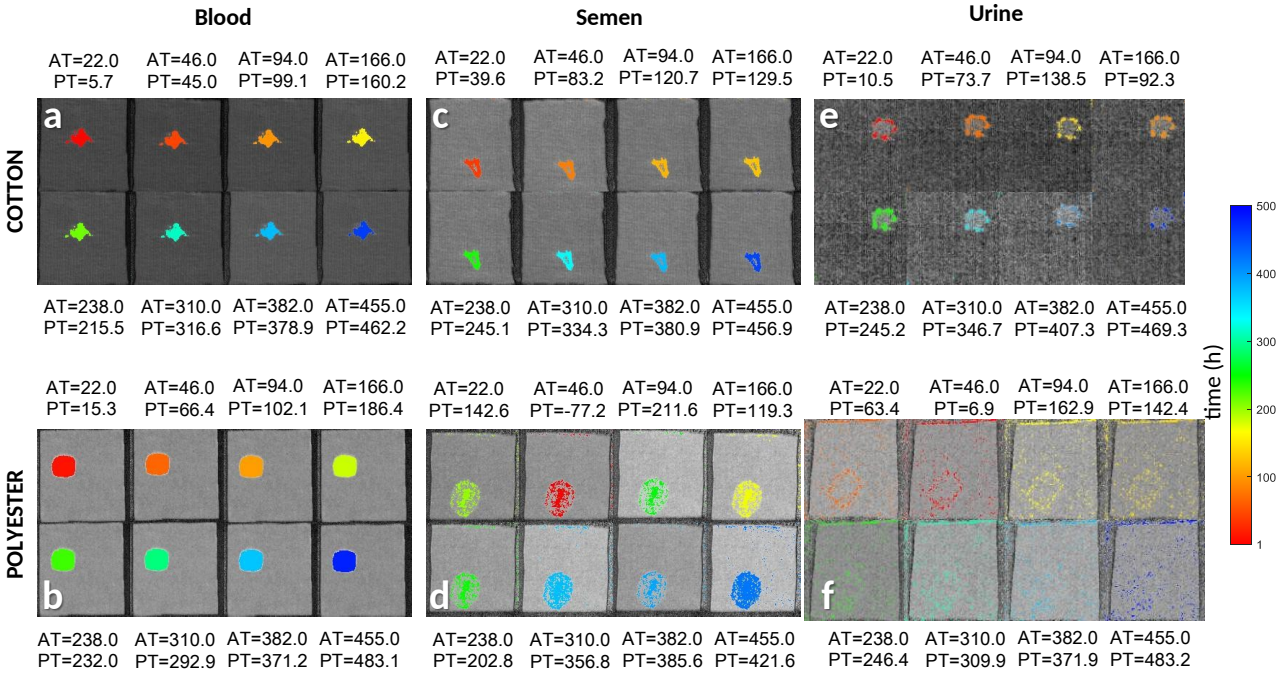


Figure 1 – Color maps related to the different fluids on the two supports. AT = actual time (hours); PT = predicted time (hours) obtained by application of PLS regression.

Investigated aspects of the present study include not only the progressive degradation of the biological trace itself, but also the effects of its interactions with the support on which it is absorbed, in particular the hydrophilic vs. hydrophobic character of fabric tissues (Fig. 1). Results are critically discussed, highlighting potential and limitations of the proposed approach for a practical implementation [1].

References

[1] C. Manis, C. Malegori, E. Alladio, M. Vincenti, P. Garofano, F. Barni, A. Berti, P. Oliveri, *Talanta* **245** (2022) 123472.

Rebalanced ASCA (RASCA) to handle unbalanced multifactorial designs

M. de Figueiredo[1,2,3], S. Giannoukos[1], S. Rudaz[2,3], R. Zenobi[1], J. Boccard[2,3]
[1]*Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland*
[2]*School of Pharmaceutical Sciences, University of Geneva, Switzerland*
[3] *Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Switzerland*
miguel.defigueiredo@unige.ch

A novel chemometric approach is proposed to analyze high-dimensional data collected from unbalanced designs of experiments. It combines a rebalancing strategy based on unique combinations averages with the ASCA method under the name Rebalanced ASCA (RASCA) [1]. The ability of RASCA to handle unbalanced designs was compared with classical ASCA, as well as some of its latest improvements, such as ASCA+ [2] and WE-ASCA [3]. Figure 1 briefly describes the RASCA algorithm.

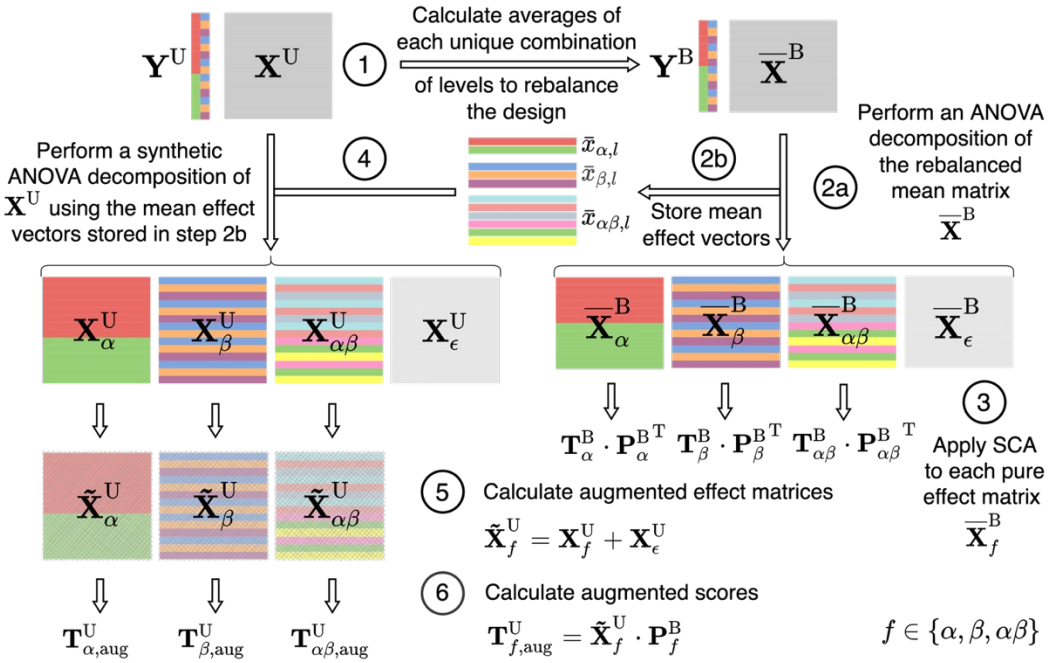


Figure 1 – Description of the main steps (1 to 6) of the RASCA algorithm.

A framework was designed to provide a systematic comparison of the various approaches. For that purpose, a real dataset obtained from an initially balanced design was gradually unbalanced in a controlled fashion by removing observations belonging to specific combinations of factor levels. This allowed an objective evaluation of the ability of the different methodologies to handle increasingly unequal group sizes. ASCA+ and WE-ASCA both aim at solving biased parameter estimators in unbalanced designs using different effect coding approaches under the general linear models methodology. However, the effect matrices obtained with both methods are not all mutually orthogonal. Even though WE-ASCA interaction effects are orthogonal to the main effects, the latter are not mutually orthogonal. RASCA offers the great advantage to solve this issue, which may be of utmost importance for the interpretation of models when facing unbalanced designs, while also providing unbiased parameter estimators. Overall, the comparison of RASCA with state-of-the-art methods demonstrated its adequacy to handle unbalanced designs.

References

[1] M. de Figueiredo, S. Giannoukos, S. Rudaz, R. Zenobi, J. Boccard, J Chemomtr. (2022) e3401.

[2] M. Thiel, B. Féraud, B. Govaerts, *J Chemomtr.* **31** (2017) e2895.

[3] N. Ali, J. Jansen, A. van den Doel, G. H. Tinnevelt, T. Bocklitz, *Molecules*, **26** (2020) 1–16.

LMWiRe: an R package for Linear Modeling of Wide Responses based on ASCA family of methods

Michel Thiel[1], Nadia Benaiche[1], Manon Martin[1], Sébastien Franceschini[2], Robin Van Oirbeek[1], Bernadette Govaerts[1]

[1]*Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA),Louvain-La-Neuve, Belgium*

[2] *Gembloux Agro-Bio Tech Département GxABT Modélisation et développement TERRA Research Centre Modélisation et développement, Gembloux, Belgium*

Michel.thiel@uclouvain.com

Many modern analytical methods are used to analyze samples issued from an experimental design; for example in medical, biological, chemical or agronomic fields. Those methods generate most of the time highly multivariate data like spectra or images, where the number of variables (descriptors) tends to be much larger than the number of experimental units. Therefore, multivariate statistical tools are needed to highlight variables that are consistently modified by experimental factors.

Two methods developed around ten years ago, ASCA [1] and APCA [2], combine ANOVA decomposition and PCA to visualize those data in a reduced space and to take into account the information from the experimental design. However, these methods are only correct in the case of a balanced design, which represents a major limitation in real case studies.

In this regard, Thiel et al. [3] exposed two new approaches, ASCA+ and APCA+, in which the General Linear Model (GLM) is used instead of ANOVA. These methods give the same results for the balanced cases and correct the bias for the unbalanced cases by using Ordinary Least Squares (OLS) estimators rather than simple differences of means to estimate the parameters. The GLM approach is applicable to all ANOVA models with fixed categorical effects and can also be extended to models including quantitative factors.

This work presents the implementation of the R package LMWiRe, standing for Linear Models for Wide Responses, which allows the statistical analysis of wide response matrices using the ASCA+ and APCA+ methods [4]. The package offers a comprehensive workflow first to explore the data set of interest, then to fit a chosen model to the data and finally to analyze the model results through a wide variety of statistics, tables and graphics.

LMWiRe presents different advantages compared to other software in this field: (1) it is able to treat any balanced or unbalanced experimental design for fixed categorical factors, (2) it offers optimized methods to calculate effect importance and test their significance, (3) it allows the user to represent data and ASCA/APCA results with various and rich ggplot2-based graphical outputs that are highly customizable and (4) the package is open to future extensions to more sophisticated statistical models. The package is now available on GitHub with two detailed vignettes and a user's manual. Its use will be illustrated on a metabolomic data set obtained by 1H-NMR spectroscopy.

References

[1] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, M. E. Timmerman, *Bioinformatics* **21 (13)** (2005) 3043–3048.

[2] P. Harrington, N. E. Vieira, J. Espinoza, J. K. Nien, R. Romero, A. L. Yergey, *Analytica chimica acta* **544 (1)** (2005) 118–127.

[3] M. Thiel, B. Féraud, B. Govaerts, *Journal of Chemometrics* **31 (6)** (Jun. 2017).

[4] N. Benaiche, Stabilisation of the r package lmwire - linear models for wide responses, Master's thesis, UCLouvain, Belgium, Louvain-la-Neuve, Belgium (2021).

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

An Experimental Design Perspective on Cross-Validation

C. Beleites[1,2]
[1]*Julius Kühn Institute, Federal Research Centre for Cultivated Plants, Institute for Ecological Chemistry, Plant Analysis and Stored Product Protection, Berlin, Germany*
[2]*Chemometrix GmbH, Wölfersheim, Germany*
claudia.beleites@chemometrix.gmbh

Experimental designs (DoE) are widely employed to ensure optimal or good experimention and data analysis. We propose to utilize a DoE perspective also for the study of chemometric models. Cross-Validation in various flavors is widely used as part of internal validation as well as to guide model selection, e.g. in hyperparameter optimization.

From an experimental design perspective, repeated cross validation yields data with test cases partially crossed with the surrogate models. We discuss two models of this structure:

(a) $(\hat{y} - y)_{case,\ model} = b + \sigma_{case} + \sigma_{model} + \sigma_{case \times model}$
(b) $(\hat{y} - y)_{case,\ model} = b_0 + b_1 \cdot y_{centered} + \sigma_{case} + \sigma_{model} + \sigma_{case \times model}$

to decompose cross-validation residuals $(\hat{y} - y)_{case,\ model}$ into bias and variance terms: bias b or b_0 for systematic over- or underprediction and b_1 covering bias towards the mean such as due to heavy regularization, and variance terms for case-to-case variance σ^2_{case}, overall performance variation between the surrogate models σ^2_{model}, and a residual (interaction) term $\sigma^2_{case \times model}$ for instability in the predictions that is individual for cases and surrogate models.
Partitioning **(a)** is a bias-variance decomposition of MSE_{CV} up to differences between degrees of freedom according to structure **(a)** and the total number of predictions in MSE_{CV}.

The models may be further adjusted to the situation at hand, e.g. with additional terms for replicate cases, by dropping non-significant terms, or by grouping σ^2_{model} and $\sigma^2_{case \times model}$ into a single $\sigma^2_{instability}$.

As an example, figures **1** and **2** illustrate models **(a)** and **(b)** for 10 × 10-fold cross-validation of PLS models with up to 10 latent variables of the well-known "gasoline" data set [1].

In future, we expect such insight into the structure of prediction error to be useful for model tuning.

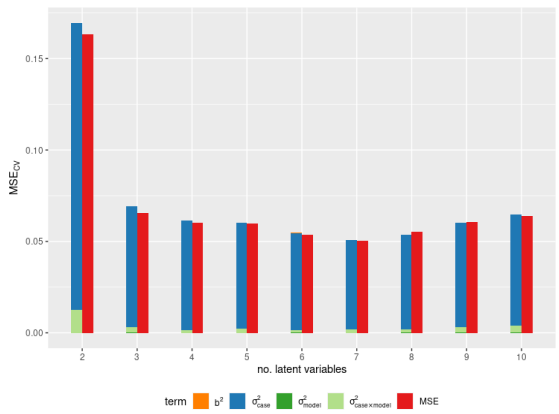


Figure 1 – MSE-like decomposition according to model **(a)**. CV error is dominated by case-to-case variance, there is no noticeable additive bias b_0. *(1 l.v. model omitted due to dominating the scale)*

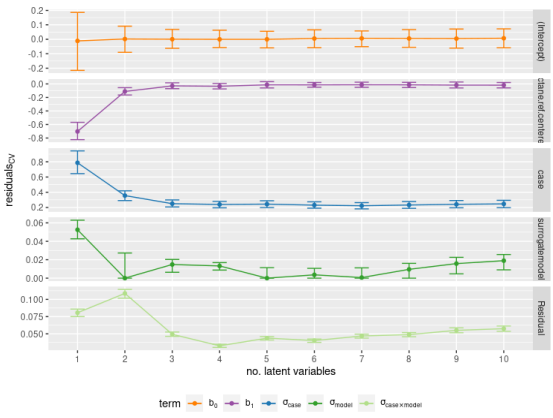


Figure 2 – Model terms/effect sizes for model **(b)**, in y-units, with bootstrapped confidence intervals (2.5 to 97.5 %). Model instability ($\sigma^2_{model} + \sigma^2_{case \times model}$) is minimal at 4 l.v., while σ^2_{case} decreases until 7 l.v. Bias towards the mean b_1 is insignificant for > 2 l.v.

References

[1] J. Kalivas, Two Data Sets of Near Infrared Spectra, *CILS*, 37 (1997) 255–259.

Quantifying the Tau protein aggregation degradation process by classification of super-resolution fluorescence microscopy localizations

S. Hugelier[1], H. Kim[1], M. Gyparaki[1], M. Lakadamyali[1]
[1]*Lakadamyali lab, University of Pennsylvania, Philadelphia, USA*
siewert@upenn.edu

The neuronal protein Tau plays an important role in facilitating the assembly and stabilization of neuronal microtubules, but, in pathological conditions it detaches from the microtubules and forms aggregates in the cytosol [1]. Whereas the accumulation of these Tau aggregates is important in many neurodegenerative diseases such as Alzheimer's or Parkinson's disease, more recent studies in animal models have shown neurodegeneration can also happen without these accumulated aggregates [2]. Little is known about these smaller aggregates because their visualization requires advanced imaging techniques such as super-resolution fluorescence microscopy.

In this study, we used super-resolution microscopy to visualize and quantify this Tau aggregate degradation process in an engineered cell model (QBI-293 Clone 4.1 cells [3]) expressing 4R-P301L-GFP tau. The Tau aggregates are maintained in these cells using the drug doxycycline, but once it is removed, they will degrade over the course of several days (Fig. 1).

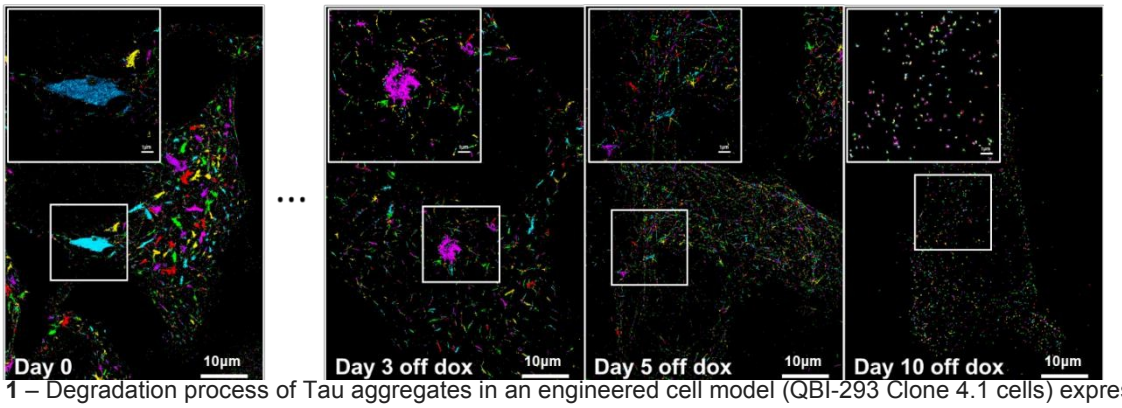


Figure 1 – Degradation process of Tau aggregates in an engineered cell model (QBI-293 Clone 4.1 cells) expressing 4R-P301L-GFP Tau, after removal of doxycycline (Dox). Clusters are pseudo color coded after Voronoi segmentation.

Super-resolution microscopy uses fluorescent labels that stochastically switch between an *on* and *off* state, which allows to image only a sparse subset of the labels at once. These labels were localized (*x,y*-coordinate extraction) over the course of several thousands of frames to visualize the Tau aggregates with a high spatial resolution (up to 20-30 nm) and then clustered using Voronoi tessellation (Fig. 1). To quantify the aggregate compositions after clustering, we developed a point-cloud descriptor method to characterize them using only the raw localizations (e.g., geometric, skeleton, etc. properties), and then used multiple supervised classification algorithms. The different methods were trained and validated on a subsample of the data (ground truth established by a human expert; ~15% of the total data) and we found that Random Forest classification [4] in combination with variable selection (15 out of 65 descriptors) gave the best classification results (1 class 80% accuracy, the others 90+%). Using this model on the remainder of the data, we found that the aggregate composition remains relatively stable early on (up to day 4), but at later time points large aggregates (neurofibrillary tangles and precursors) are rapidly degraded whereas linear fibrils persist and are resistant to degradation.

To sum up, we believe that using this novel shape descriptor method on the raw localizations of super-resolution microscopy images in combination with different chemometric techniques helps in the unbiased quantification of complicated biological processes and have applied it to describe the degradation process of Tau aggregates.

References

[1] I. Grundke-Iqbal et al., *Proc. Natl. Acad. Sci. U.S.A.* **83**, 4913–4917 (1986).
[2] C. Cowan et al., *Acta Neuropathol.* **120**, 593–604 (2010).
[3] J. Guo et al., *J. Biol. Chem.* **291**, 13175–13193 (2016).
[4] L. Breiman, *Mach. Learn.* **45**, 5–32 (2001).

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

CALIBRATION TRANSFER OF NEAR-INFRARED AND RAMAN MODELS WITHOUT USING TRANSFER SAMPLES

E. Tengstrand[1], T. Lintvedt[1], P. Andersen[1], L. Solberg[1], J. Wold[1], N. Afseth[1]
[1]*Nofima, Ås, Norway*
erik.tengstrand@nofima.no

For years, calibration transfer has been an important research area within vibrational spectroscopy. With increasing practical applications, and increasing availability of handheld low-cost spectroscopy systems, the concept of transferring a calibration from one spectrometer to another while retaining model accuracy and precision is gaining significance. Most calibration transfer methods require access to both target and source spectrometers, or at least samples measured on both. In some cases, this is not feasible.
Here we present a calibration transfer method that does not require access, nor samples measured on both spectrometers. Using this approach, we also present results from transferring models between two datasets where traditional transfer methods are not feasible: where the data was taken years apart and the source instrument is no longer available. The two studies both investigate prediction of fatty acid composition in salmon using Near-Infrared (NIR) and Raman spectroscopy, respectively. The first study was performed to investigate the genetic impact on fatty acid composition and to develop an analysis method for bred salmon [1]. The second study was performed to develop a NIR and Raman quality inspection technique for farmed salmon based on fish fed different diets. The two studies were done years apart, with different instruments, with different sample handling, and different sample composition. Despite these differences, a relatively simple calibration transfer was still successful both for Raman and NIR. Figure 1 illustrates the performance of the transfer with predictions plotted against the references for both source and target datasets. In the presentation, the approach and the results will be presented together with a critical discussion on the potential use of the approach.

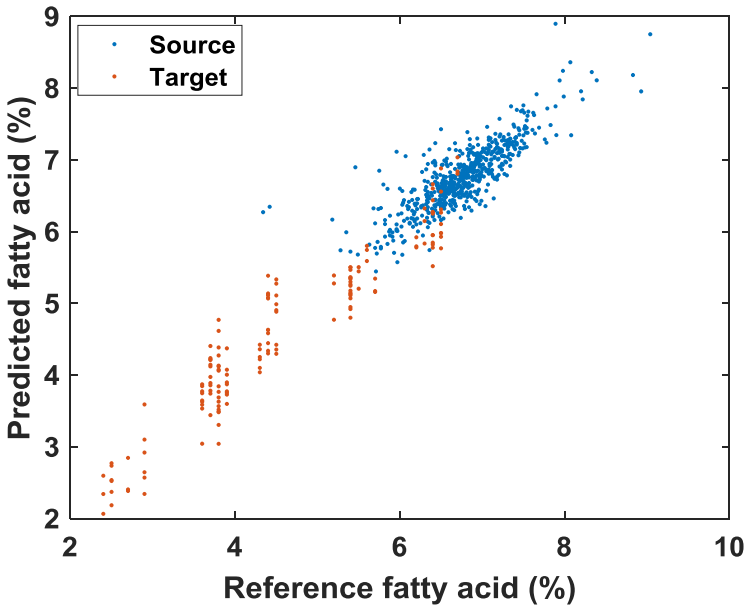


Figure 1 – Raman predictions plotted against the references for both the source and target datasets, using models based on the source dataset. The NIR dataset has similar performance. The source dataset was cross validated.

References

[1] N. Afseth, K. Dankel, P. Andersen, G. Difford, S. Horn, A. Sonesson, B. Hillestad, J. Wold, E. Tengstrand, Foods, 11 (2022) 962.

CHARACTERIZATION OF MICROPLASTICS FROM MARINE ORGANISMS USING NEAR INFRARED HYPERSPECTRAL IMAGING

Alisa Rudnitskaya[1], Catarina Moreirinha[1], Filipa Marques[2,3], Carlos Vale[3], Sara T. Costa[2,3], Maria João Botelho[2,3]

[1] *CESAM, Centre for Environmental and Marine Studies, and Chemistry Department, University of Aveiro, Aveiro, Portugal*

[2] *CIIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto, Portugal*

[3] *IPMA, Portuguese Institute for the Sea and Atmosphere, Lisbon, Portugal*

alisa@ua.pt

Microplastics (MPs), defined as polymeric particles with size below 5 mm, are ubiquitous pollutants encountered in all planet ecosystems and across all trophic levels. Bivalves as filter feeders are capable to accumulate particulate matter, and ingestion of MPs by these organisms is well documented. One of the common saltwater bivalves, mussels, was proposed as an indicator species for MP contamination in seawater.

The objective of this work was feasibility study of the NIR Hyperspectral imaging for identification of particles extracted from bivalve tissue. The mussel *Mytilus spp* and the cockle *Cerastoderma edule* specimens were sampled weekly during winter and summer in the Aveiro lagoon. After soft tissues digestion using alkali (KOH) and filtration, particles that presented no cellular or organic structures visible inside and kept the structure when prodded were counted as potential MPs. NIR hyperspectral images of the visually sorted particles arranged on the microscopy slides were recorded using a benchtop HSI system (FX17e HyperSpectral LabScaner, Specim, Finland) in the wavelength range 900–1700 nm with spatial resolution 17 µm. MPs identification was done using FT-MIR spectroscopy. ROIs corresponding to the individual particles were identified in the pre-processed NIR images by analyzing false color images and using optical microscopic images as a reference. Geometrical parameters and mean spectrum of each particle were computed. The PLS-DA classification models were developed using extracted spectra and MPs identification obtained by FT-MIR as a reference. This study confirmed potential of NIR imaging as a rapid MPs' identification tools highlighting importance of reference materials for the characterization of MPs in the environment.

Acknowledgments. We acknowledge financial support to CESAM by FCT/MCTES (UIDP/50017/2020 + UIDB/50017/2020 + LA/P/0094/2020) and to CIIMAR (UIDB/04423/2020 + UIDP/04423/2020), through national funds, and through the strategic project ALG-01-0145-FEDER-032453.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

SPECTRAL EVALUATION OF FRESH GRAPEVINE ORGANS USING SELF-ORGANIZING MAPS (SOM)

E. van Wyngaard[1], E. Blanquaert[1], H. Nieuwoudt[1], J.L. Aleixandre-Tudo[1,2]

[1]*South African Grape and Wine Research Institute (SAGWRI), Department of Viticulture and Oenology, Stellenbosch University, South Africa*

[2]*Insituto de Ingeniería de Alimentos para el Desarrollo (IIAD), Universitat Politecnica de Valencia, Camino de Vera s/n (Valencia), 46022, Spain*

joaltu@sun.ac.za; joaltu@upvnet.upv.es

High quality wine grapes are fundamental to produce the most appreciated wines. Grapevine nutritional status during the growing and ripening season is therefore of utmost importance. The non-destructive and in-field measurement of the nutritional content will generate real time nutritional data that could be used in the benefit of wine grape production. The combination of infrared spectroscopy with chemometrics provides the means to obtain such information. However, during the growing season chemical, physical and morphological changes take place in the shoots, leaves and berries of the vines. Accurate calibrations to quantify nutritional content are not yet available, but before calibrations are attempted, the changes of the spectral properties occurring during the growing season should be explored. This investigation therefore aimed at exploring the infrared spectral of fresh grapevine shoots, leaves, and berries throughout the grapevine growing season with the use of the self-organising maps (SOM). SOM is a type of neural network applied to large and multi-dimensional datasets, making it particularly suited for spectral data. Mid-infrared (MIR) and near-infrared (NIR), with a solid probe (NIR-SP) and a rotating sphere (NIR-RS), were used for spectral data acquisition. Spectral data from shoots, leaves and berries was obtained monthly from November to March across two vintages (2019-2020 and 2020-2021). Five locations, seven cultivars, and 17 commercial vineyards were included. The unsupervised SOM analysis showed the most considerable clustering based on organ type. Additionally, separation trends based on phenological stage were also observed. Further investigations per organ showed separation based on phenological stage for berries and shoots, and shoots based on lignification. Considering the observed clustering, supervised SOM were examined for classification. The accurate prediction of organ at 90.3% was possible for the NIR-SP. Overlapping of various phenological stages were seen for the grape berry datasets, but prediction improved to 85.6% for NIR-RS when certain phenological stages were grouped together. Accurate predictions of lignified and unlignified shoots were also seen for both NIR techniques at 74.4% and 89.9% respectively. Following variable selection with OPLS-DA and S-plots, the prediction of shoots and leaves improved by 14% for the NIR-RS. The prediction of lignified and unlignified shoots improved considerably to 92.3% for the NIR-SP and 95.9% for the NIR-RS. This study showed the extensive information available in the infrared spectra of fresh grapevine organs and how the information could be used to achieve important clustering and classification objectives and finally to attempt the optimisation of specialized calibrations for nutritional content in grapevines.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Infrared Ion Spectroscopy Peak Matching using Peak Annotation Technique

N. Omidikia[1], J. Jansen[1], J. Oomens[2]

[1]*Department of Analytical Chemistry, Institute for Molecules and Materials (IMM), Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands*

[2]*Institute for Molecules and Materials, FELIX Laboratory, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands.*

nemat.omidikia@ru.nl

Detection of unknown compounds requires accurate structural elucidation which is challenging using only a single analytical techniques. However, infrared ion spectroscopy (IRIS) combines the advantages of infrared spectroscopy and mass spectrometry providing mass-selective IR spectra [1]. The IR features provide structural information to link functional groups into group frequencies. So, IRIS is able to provide detailed structural information with a great potential to the unknown metabilte identification [2]. Finally, for the sake of identifaction, the recorded IR spectra can be compared to reference IR databases measured from standards or predicted by quantum-chemical computations [1-2].

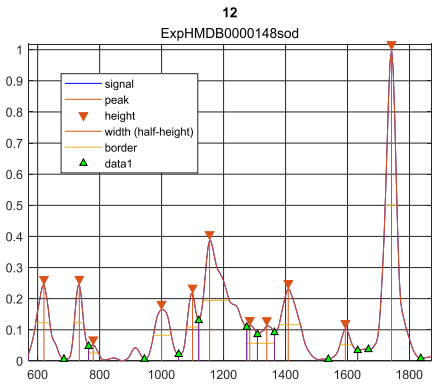


Figure 1 – A summary of peak annotation of a IR spectra from Human metabolite data bank (HMDB).

Such comparison requires a well-defined matching score tolerant to experimental artifacts within the recoded spectra, *e.g.* offset and sloping of baseline as well as variations in the absolute values of peaks and broadening. Meanwhile, analysis of IRIS spectra poses several challenges due to the complexity of the spectra, laser energy effect and high energy states [1-2]. Although peak similarity has been started with functional group matching in the area of analytical chemistry [3], it turned into mathematical indicators like "*angle*" and/or "*distances*" [3]; parameters that do not consider chemical relevance and heavily rely on large parts of the peak, not the details.

Here, we propose a chemically transparent and effective alternative for cosine spectral peak matching. The proposed peak annotation strategy, incrementally calculates the spectral similarity based on the local spectral features (see **fig. 1**). Lastly, the proposed scoring index was exemplified using human metabilte data base.

References

[1] R. E. van Outersterp, U. F.H. Engelke,…, J. Oomens, *Anal. Chem.* **93** (2021) 15340−15348.

[2] F.A.M. G. van Geenen, …, J. Oomens, G. Berden, Anal. Chem. **93** (2021) 2687−2693.

[3] K. Tanabe, S. Saéki, Anal. Chem. **47** (1975) 118-122.

Peak matching across Gas Chromatography – Ion Mobility Spectrometry samples

Sergio Oller-Moreno[1], Celia Mallafré-Muro[1,2], Arnau Blanco[1,2], Meryl Cruz[1,2], Antonio Pardo[1], Luis Fernández[1,2], Santiago Marco[1,2]

[1]*Inst. for Bioengineering of Catalonia (IBEC), The Barcelona Inst. of Science and Tech. Spain*

[2] Department of Electronics and Biomedical Engineering, University of Barcelona, Spain

soller@ibecbarcelona.eu

The analysis of Volatile Organic Compounds (VOCs) found in complex biofluids (volatolome) is crucial for the diagnostic of many medical conditions. Its complexity raises from the high number of VOCs present, ranging in the hundreds [1], and their high variability, as many are exogenous.

The analysis of VOCs is often challenging, due to their usual low concentrations and intrinsic volatility. Several techniques exist, being the one of the most popular the use of Gas Chromatography (GC), coupled to a mass spectrometer (GC-MS). Alternatively, a GC can be coupled to an Ion Mobility Spectrometer (GC-IMS). GC-IMS operates at ambient pressure, and it requires none or very little sample pretreatment. While a GC-IMS does not have the resolution we find on a GC-MS, it's a more affordable instrument providing a fast and reliable response.

A GC-IMS sample can be represented as an intensity matrix, with the GC retention time in the rows and the IMS drift time in the columns (Fig.1). The data analysis pipeline comprises of several steps, including baseline correction, alignment, deconvolution and peak detection and matching, leading to a peak table. Many algorithms already exist to address those signal processing steps.

Still, GC-IMS samples have specific requirements that need tailored signal processing. For instance, compared to GC-MS, GC-IMS peaks tend to be wider and have longer tails in the retention time. This peak shape and a dense peak distribution make the merging of peak lists into a peak table a challenging problem. Some analysis pipelines avoid this peak matching step by modelling the full IMS spectra, missing the interpretability of a peak table, other pipelines use strategies based on the nearest peak [2], or are limited to few manually annotated peaks.

In this work, we will compare and benchmark several peak matching algorithms, applying them to complex human urine samples. The comparison includes k-means, k-medoids and hierarchical clustering methods, provides criteria to determine optimal clustering parameter values and explores the stability of the solutions with respect to the clustering parameters.

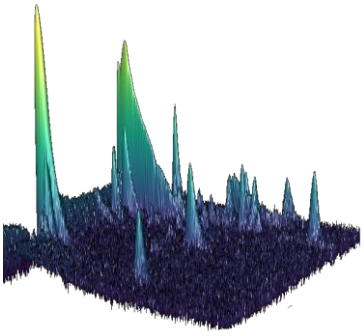


Figure 1 – Example of a GC-IMS urine sample

Acknowledgements

This work has been funded by Spanish MINECO Project TENSOMICS (RTI2018-098577-B-C22).

References

[1] de Lacy Costello, B., Amann, A., Al-Kateb, H., Flynn, C., Filipiak, W., Khalid, T., Osborne, D., and Ratcliffe, N. M. A review of the volatiles from the healthy human body. Journal of Breath Research (2014), 8(1):014001, ISSN: 1752-7155 DOI: 10.1088/1752-7155/8/1/014001

[2] Freire, R.; Fernandez, L.; Mallafré-Muro, C.; Martín-Gómez, A.; Madrid-Gambin, F.; Oliveira, L.; Pardo, A.; Arce, L.; Marco, S. Full Workflows for the Analysis of Gas Chromatography—Ion Mobility Spectrometry in Foodomics: Application to the Analysis of Iberian Ham Aroma. *Sensors* **2021**, *21*, 6156. DOI: 10.3390/s21186156

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Multivariate monitoring and update strategies for calibration models

V. Fonseca Diaz[1], W. Saeys[1]
[1]*KU Leuven Department of Biosystems, MeBioS division, Kasteelpark Arenberg 30, Leuven, Beligum*
wouter.saeys@kuleuven.be

Multivariate calibration models are largely used nowadays for quality control processes that involve chemical quantification of products in an efficient non-destructive way. After models are built and deployed to use them for prediction, the success of multivariate calibration depends on the validity of the models in the long term. Ensuring such validity comprises the study of model maintenance which involves the performance monitoring of the model and its correspondent adaptation when degradation or model drift is detected. Model monitoring has been an active topic of research largely focused on applications where models operate on-line and individual samples are inspected to detect outliers and possible leverage points [1,3]. Likewise, the study of model update has taken place in the context of parameters update with the acquisition of individual samples and their new reference values [1,2]. From a more practical perspective, monitoring strategies have been disseminated in the form of roadmaps to detect not only outliers in new datasets, but also model drift using new reference analyses in a time period after the calibration model was built [3]. In the current work, we have developed a strategy for model monitoring of a multivariate nature where the model drift is quantified based on model comparison and not on distance quantification of new samples. Such a strategy is suitable for use with calibration and test data whose reference analyses are available from the moment the model was built and a separate set of spectral samples from a future time point at the moment of monitoring. The strategy particularly focuses on bias drift of the calibration model and spectral variability drift. In conjunction with the proposed model monitoring strategy, a framework for model update has been developed in which different categories of methods are mapped to the detected model drift. The strategies here presented have been applied to real case studies in the agrofood industry with samples measured in different time frames and multiple times points.

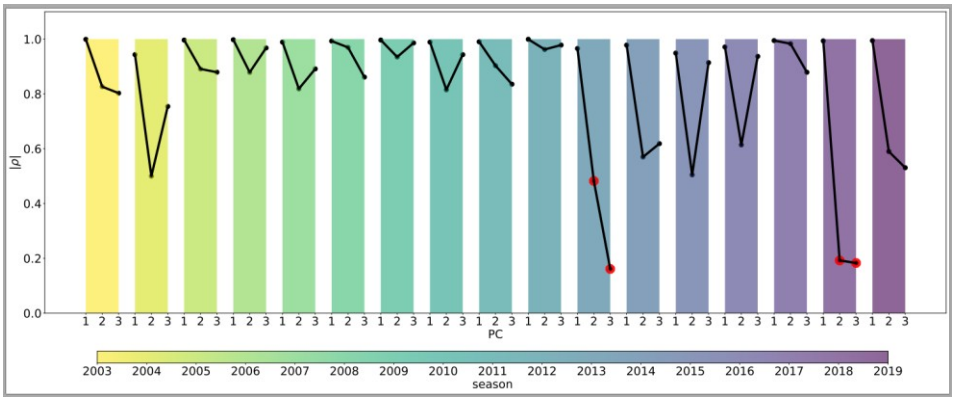


Figure 1 – Multivariate scores monitoring in apples dataset. Monitoring values lower than 0.5 are marked in red

References

[1] Chen, K., Castillo, I., Chiang, L. H., and Yu, J. Soft sensor model maintenance: A case study in industrial processes. IFAC-PapersOnLine 28, 8 (2015), 427–432.
[2] Haavisto, O., and Hyötyniemi, H. Recursive multimodel partial least squares estimation of mineral flotation slurry contents using optical reflectance spectra. Analytica Chimica Acta 642, 1-2 (2009), 102–109.
[3] Wise, B. M., and Roginski, R. T. A calibration model maintenance roadmap. IFAC-PapersOnLine 28, 8 (2015), 260–265.

What's UMAP Doing Anyway?

Barry M. Wise, Manuel A. Palacios, Donal O'Sullivan and Sean Roginski
Eigenvector Research, Inc., Manson, WA USA
bmw@eigenvector.com

Uniform Manifold Approximation for Projection and Dimension Reduction (UMAP) is getting increasing attention in the machine learning community, though it is not without controversy. It is used both as a visualization tool and as a dimension reduction tool, often as a preprocessing step for classification. The mathematics involved in the develop of UMAP is beyond all but the most mathematically savvy. However, it is possible to show how UMAP maps the original data space into its own embedding space. We illustrate this with several examples. This visualization helps users understand what UMAP is doing, even if they don't understand exactly how it gets there. This visualization is also useful going forward when developing classifiers based on UMAP compression.

As an example of the potential of UMAP, consider below the results of a Principal Components Analysis of the Chimiométrie challenge data from 2018 [2]. The data consists of the NIR spectra from 3908 samples of 10 different food stuffs. The PCA scores show considerably overlap among the 10 classes (below, left) while the (unsupervised) UMAP embeddings separate the classes widely and almost perfectly (below, right).

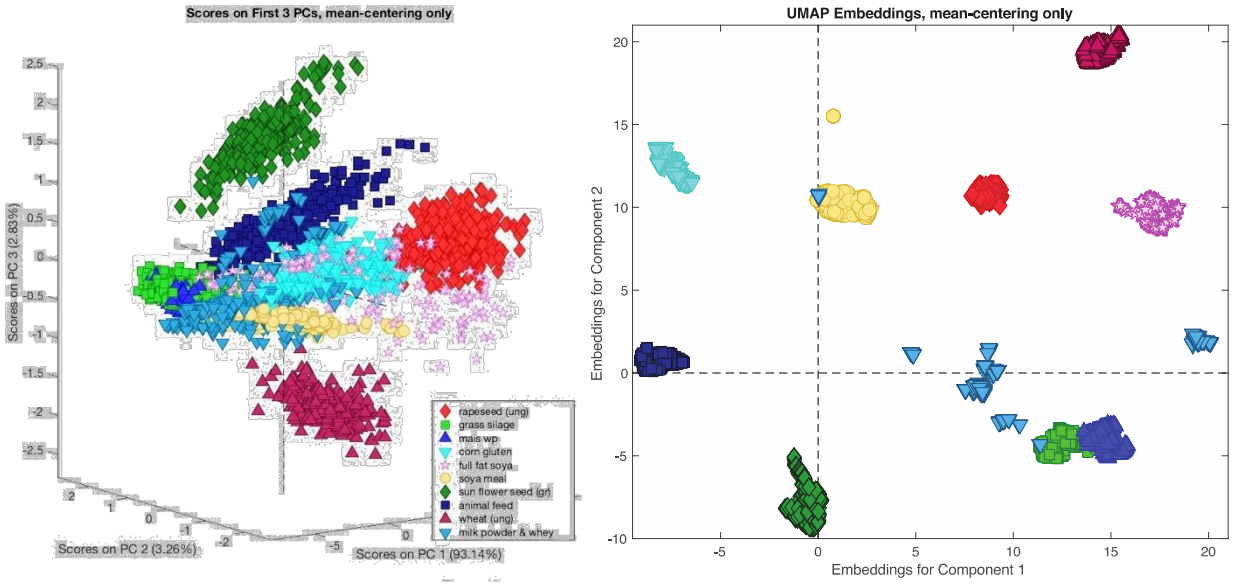


Figure 1 – Principal Components Analysis Scores for Chimiométrie Challenge Data (left) Compared with UMAP Embeddings (right).

While the above example is compelling, the jury is still out on whether UMAP will be a generally useful tool in the chemometrics community. We do, however, have suggestions based on experience so far as to how to best use it in chemical applications.

References

[1] Leland McInnes, John Healy and James Melville, Uniform Manifold Approximation and Projection for Dimension Reduction, https://arxiv.org/abs/1802.03426

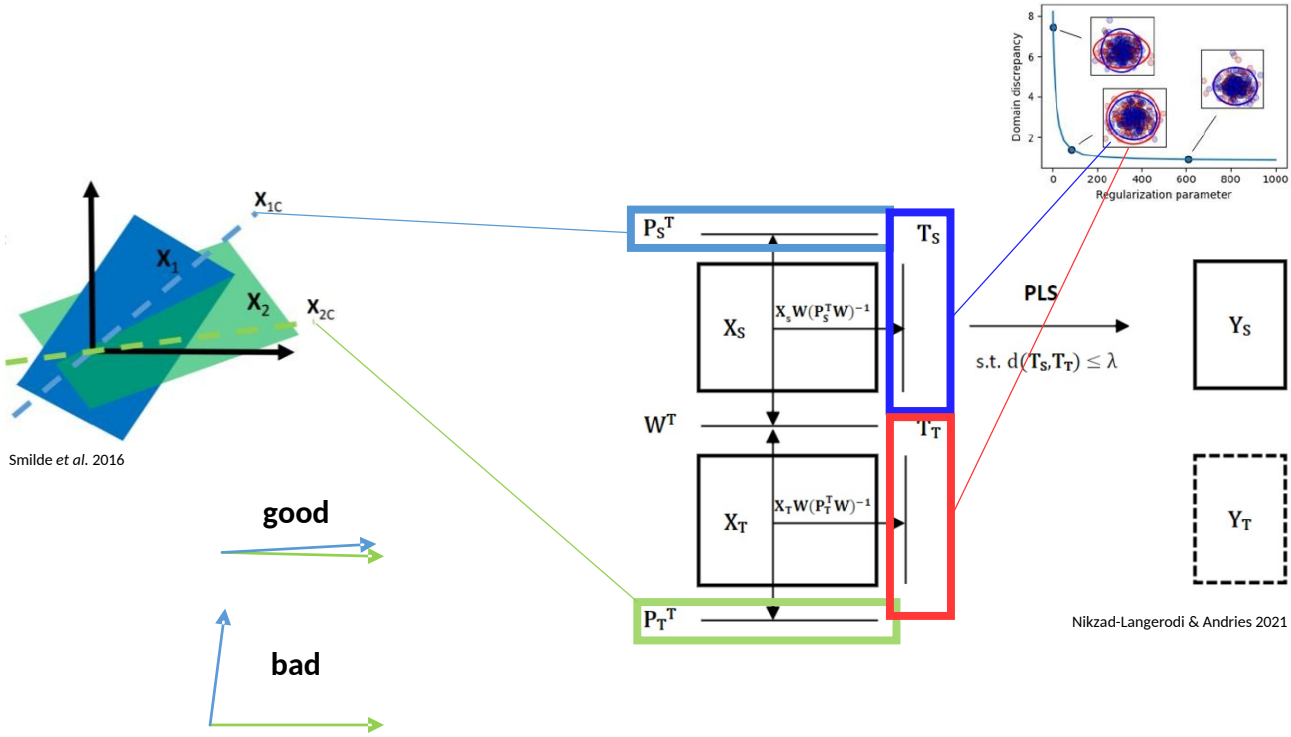
[2] Data provided by Dr Vincent Baeten, director of the Quality and authentication of agricultural products Unit, Walloon Agricultural Research Centre, Gembloux, Belgium
https://www.cra.wallonie.be/en

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Does it Transfer? Assessing model generalization in domain adaptation with data fusion

Ramin Nikzad-Langerodi[1], Valeria Fonseca Diaz[2]
[1] *Software Competence Center Hagenberg, Hagenberg, Austria*
[2] *KU Leuven, Leuven, Belgium*
ramin.nikzad-langerodi@scch.at

Domain adaptation (DA) in general and domain-invariant representation learning in particular, has recently emerged as useful approach to model maintenance and updating without reference standards or labeled data (e.g., spectra + reference measurements) [1,2]. The theory of learning from different domains provides upper bounds for generalization under dataset shift (i.e., when marginal and/or conditional distributions of training and test data are different) [3]. However, these bounds can either not be computed for finite samples or are too loose, which makes it difficult to estimate the generalization error of a (calibration) model in some target domain or to decide if (and how) a model can be updated to restore the required accuracy/precision. In the current contribution, we present a novel measure to assess if a source (or domain-invariant) calibration model will generalize to a target domain. The main underlying hypothesis is that generalization will succeed only if the corresponding features encode both common (in a data fusion sense) and domain-invariant (i.e., in terms of distributional properties) latent information. We have empirically explored (and validated) this hypothesis based on simulated and real-world data.



References

[1] Nikzad-Langerodi, R., Zellinger, W., Lughofer, E., & Saminger-Platz, S. (2018). Domain-invariant partial-least-squares regression. Analytical chemistry, 90(11), 6693-6701.
[2] Nikzad-Langerodi, R., & Andries, E. (2021). A chemometrician's guide to transfer learning. Journal of Chemometrics, 35(11), e3373.
[3] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79(1), 151-175.

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 Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET - Competence Centers for Excellent Technologies program 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).

NEW DEVELOPMENTS AROUND THE VIP INDEX

B. Mahieu E. M. Qannari B. Jaillais
ONIRIS, INRAE, StatSC, Nantes, France

benjamin.mahieu@oniris-nantes.fr

Within the context of PLS regression, VIP index is extensively used to highlight the importance of a given predictor for setting up a PLS model. It was introduced in [1] and further discussed in [2]. Very often, it is used as a means to select meaningful variables by retaining those variable with a VIP index larger that a pre-specified cut-off value. However, VIP index is commonly defined within the context of PLS1 regression (*i.e.*, one response variable). Its extension to the case of a multivariate response variable is simply hinted to in [3] but not formally defined.
The purpose of the present communication is to extend the VIP in several directions:
(*i*) Assessing the variable importance in the case of PLS regression with a multivariate response variable.
(*ii*) Assessing the variable importance of a subset of predictive variables and not merely one predictor at a time. In particular, we may assess the importance of a whole block of variables within the context of multiblock data analysis. In spectroscopy, this may be useful to assess the importance of a whole spectral band in a PLS model.
(*iii*) Adapting VIP index to the framework of Principal components analysis. We believe that VIP index could be as useful in this context as it is in the context of PLS regression.
Furthermore, we set a hypothesis testing framework based on permutations to assess the significance of the VIP values.

Illustrations of the various aspects outlined herein are shown based on real datasets.

References

[1] S. Wold, E. Johansson, M. Cocchi, ESCOM, Leiden, Holland (1993), pp. 523-550.
[2] C. Il-Gyo, J. Chi-Hyuck, Chem. and Intelligent Lab. Systems (2005), 78, 1–2, 103-112.
[3] M. Cocchi, A. Biancolillo, F. Marini, Comprehensive Analytical Chemistry *(2018), 265-299.*

CHEMOMETRICS IN SPENT NUCLEAR FUEL REPROCESSING

D. Kirsanov[1], A. Legin[1], V. Babain[2]
[1]*St. Petersburg University, St. Petersburg, Russia*
[2]*Sensor Systems LLC, St. Petersburg, Russia*
d.kirsanov@gmail.com

Spent nuclear fuel (SNF) reprocessing is intended for recovery of uranium, plutonium and several other important elements in order to employ them further as components for nuclear products. The technological process applied for reprocessing is called PUREX (Plutonium – Uranium Extraction). During this hydrometallurgical process SNF is dissolved in concentrated nitric acid and then the fuel components are separated from each other through the series of extraction/back-extraction steps between aqueous and organic phase to obtain target components. The SNF reprocessing media is very challenging object for chemical analysis due to the complex composition of the media, its' high radioactivity (bulk specific activity up to several dozens Ci/L) and strong acidity (up to 10 M of nitric acid). Still, it requires a thorough analytical control at every stage to ensure the process safety and normal conditions. Currently, there are no methods available for on-line monitoring and process streams are only analyzed with sampling methods after multiple dilutions to decrease radiation burden for personnel and to ensure appropriate functioning of the analytical instrumentation, like inductively coupled plasma atomic emission spectrometry and mass spectrometry. This leads to the fact that the results of element quantification are available only several hours after the sampling and timely process management is not possible. There is a compelling need in the development of on-line methods for analysis of SNF media.

Two different types of analytical instruments are currently being explored for this purpose: optical fiber probe spectrometry and electrochemical sensor arrays. Both of these methods do not provide for sharp selectivity in such a complex media as SNF reprocessing solutions, and thus require thorough and dedicated data processing to extract useful chemical information from unresolved and noisy responses.

This presentation is intended to overview the contribution of chemometrics to the development of on-line tools for SNF reprocessing monitoring. Started about 10 years back from pretty simple model solutions, today chemometrics has allowed developing of the analytical methodologies for UV-Vis spectrometry and potentiometric mutisensor systems that help to quantify plutonium, uranium, thorium and other important analytes in real SNF process streams. The use of such classical chemometric tools as PLS and MCR-ALS yields reliable models capable of providing important information on the composition of SNF media [1,2]. The application of non-linear dimensionality reduction tools (*isomap, SOM*) allows getting useful insights into the process trajectories, while non-linear regression methods (*kernel-regularized least squares, random forests*) yield improved precision in quantification of radionuclides in complex process streams [3]. The advantages and challenges of applying chemometrics for nuclear fuel reprocessing will be discussed in the presentation.

References

[1] B. Debus, D. Kirsanov, C. Ruckebusch, M. Agafonova-Moroz, V. Babain, A. Lumpov, A. Legin, *Chemom. Intell. Lab. Syst.* **146** (2015) 241–249
[2] M. Agafonova- Moroz, J. Savosina, Y. Voroshilov, S. Lukin, A. Lumpov, V. Babain, E. Oleneva, A. Legin, D. Kirsanov, *J. Radioanal. Nucl. Chem.* **323** (2020) 605–612
[3] N. Kravic, J. Savosina, M. Agafonova-Moroz, V. Babain, A. Legin, D. Kirsanov, *Chemosens.* **10** (2022) 90

Monitoring of fermentation processes by gas chromatography-ion mobility spectrometry (GC-IMS)

Joscha Christmann[1,2], Sascha Rohn[2,3], Philipp Weller[1]
[1] *Institute for Instrumental Analytics and Bioanalysis, Mannheim University of Applied Sciences, Paul-Wittsack-Straße 10, 68163 Mannheim, Germany*
[2] *Hamburg School of Food Science, University of Hamburg, Grindelallee 117, 20146 Hamburg, Germany*
[3] *Department of Food Chemistry and Analysis, Institute of Food, Technology and Food Chemistry, Technische Universität Berlin, TIB 4/3-1, Gustav-Meyer-Allee 25, 13355 Berlin, Germany*
j.christmann@hs-mannheim.de

Gas chromatography hyphenated to ion mobility spectrometry (GC-IMS) is a powerful two-dimensional separation and detection technique for volatile organic compounds (VOC). Low detection limits (low ppb to ppt by volume), high selectivity and robust operation characterize it as an ideal tool for non-target screening (NTS) of complex sample materials [1]. Due to the high sensitivity, headspace sampling without enrichment is commonly used. Exhaust gas analysis from fermentations allows to generate profiles of volatile, extracellular metabolites without interference or contamination of the process, as it is easily available. The obtained 3-dimensional GC-IMS data (Figure 1) can then be used to characterize the fermentation and predict target variables that are difficult to be measured directly e.g., the formation of a product or potential deviations in a process.

This talk presents results from an offline proof-of-concept study which demonstrates that *E. coli*, *S. cerevisiae*, *L. brevis* and *P. fluorescens*can be categorized simply by VOC profiling as a first step towards detecting contaminations. Further, the transition to online measurement and data analysis with a new GC-IMS prototype are explained. Data analysis was carried out using the new inhouse-developed Python package *"gc-ims-tools"* for data specific I/O, preprocessing, and visualizations which is available under the BSD 3-clause license at https://github.com/Charisma-Mannheim/gc-ims-tools.

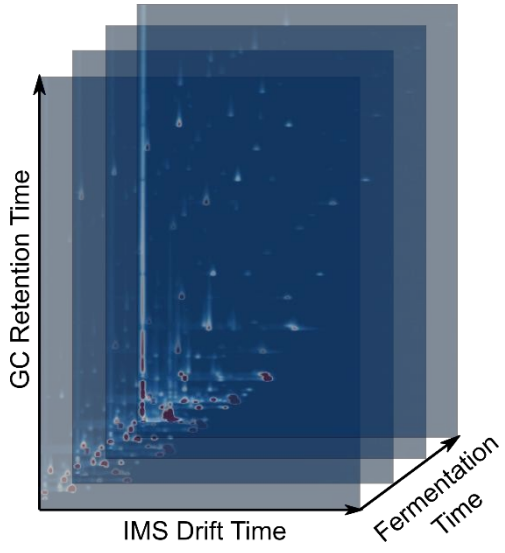


Figure 1 – GC-IMS data cube

References

[1] Capitain, C. and Weller, P. 2021. Non-Targeted Screening Approaches for Profiling of Volatile Organic Compounds Based on Gas Chromatography-Ion Mobility Spectroscopy (GC-IMS) and Machine Learning. *Molecules (Basel, Switzerland)* 26, 18.

Development of an analytical platform for the identification of *Fusarium circinatum* in culture media, using VIS-NIR spectroscopy and chemometric methods

M. Bravo[1,2], E. Sanfuentes[3], J. Ulloa.[4], V. Sandoval [3], A. Navarro[3]. C. Fuentes[1,2], R. Castillo[1,2*]

[1]*Faculty of Pharmacy, University of Concepción, Chile*

[2]*Laboratory of Biospectroscopy and Chemometrics, Biotechnology Center, University of Concepción*

[3]*Laboratory of Phytopathology, Faculty of Forest Science, University of Concepción, Chile* [2]

[4] *Laboratory of Forest Genomics and Molecular Biology, Biotechnology Center, University of Concepción, Chile*

martinbravo@udec.cl

Pitch canker is a disease that aggressively attacks *Pinus radiata* species and is caused by the fungus *Fusarium circinatum*. The current diagnosis is based on PCR molecular technique whose final report takes more than 7 days. The objective of this work was to identify the fungus *F. circinatum* in culture media using various VIS-NIR spectral techniques and multivariate methods. This work allowed discriminating the species *F circinatum*, *F. tricinctum*, and *F. graminearum*, all strains isolated from *P. radiata*. These strains were propagated on 3 different culture media and analyzed after 48, 72 and 96 hours by various VIS-NIR spectral techniques. The data were explored by PCA and then by supervised classification methods, such as k-nearest neighbor (KNN), soft independent modelling of class analogies (SIMCA) and supportt SVM. The obtained models showed correct classification rates higher than 85% for the PDA medium after 72 hours. The chemometric models developed improve the selectivity of the technique used. The results demonstrated the potential of the VIS-NIR spectral technique in early diagnostic support in the identification of *F. circinatum* from culture media. Early detection favors better disease control and decreases nursery losses.

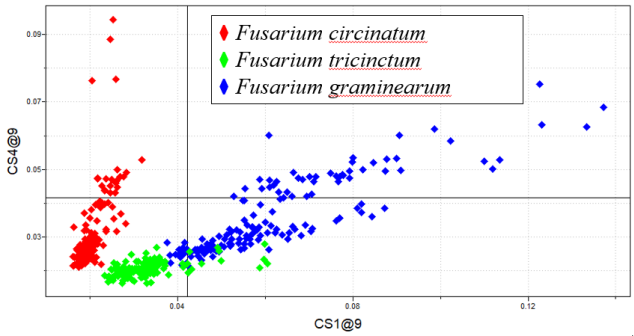


Figure 1 – Cooman's plot of SIMCA model for PDA after 72 h

Acknowledgments:
This work was financed by the FONDECYT 1221387 project, Martín Bravo acknowledges Doctoral Scholarship from the National Agency for Research and Development (ANID) Chile, number 21201971 and the Doctoral Program of Science and Analytical Technology of the University of Concepción, Chile.

References

[1] A. Carrasco, E. Sanfuentes, A. Durán, & S. Valenzuela. *Gayana - Botanica*, *73*(2016), 369–380.
[2] C. Levasseur, L. Pinson-Gadais, D. Kleiber, & O Surel, O. Revue de Medecine Veterinaire, *161*(2010), 438–444.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Retrospective Quality by Design r(QbD) using Historical Process Data and Design of Experiments

T. Offermans[1], L. Galvis[1], C. Bertinetto[1], A. Carnoli[1], E. Karamujić[2], W. Li[2], E. Szymańska[2], L. Buydens[1], J. Jansen[1]
[1]*Radboud University, Nijmegen, The Netherlands*
[2]*FrieslandCampina, Amersfoort, The Netherlands*
t.offermans@science.ru.nl

Quality by Design (QbD) is a popular formal approach for designing, upscaling and optimizing industrial production facilities towards guaranteed quality [1]. To avoid the many costly experiments required for QbD, historical production data may be exploited for optimization instead in what is known as a retrospective QbD (rQbD) study. Current rQbD literature does limitedly discuss data-driven identification of Critical Process Parameters (CPPs) to optimize limited process knowledge availability, and does not cover situations where technical operation limits have not yet been fully explored and/or where parallel equipment (lines) are used [2-3]. This work presents a new rQbD strategy that addresses these challenges by balancing knowledge that can be obtained from statistical analysis of historical data, together with process experts with a carefully designed set of plant-scale experiments within current operational limits. This novel strategy (outlines in Figure 1) utilizes established chemometrics modelling techniques including PLS [4], ASCA [5] and DoE [6], and is demonstrated on a long-running industrial lactose production facility. By digitally and experimentally exploring historical operation variability, we found new operational regimes for this production that may lead to up to 7% product quality improvement, reduced energy consumption and increased process understanding. Although optimizing a specific process by necessity requires a process-specific approach, the way in which we systematically optimize the current process with Hybrid AI (combining available knowledge with new insights from historical data) shows that approaches that are currently used in prospective process upscaling may be modified to be invaluable for optimization of full-scale processes with a long operational history.

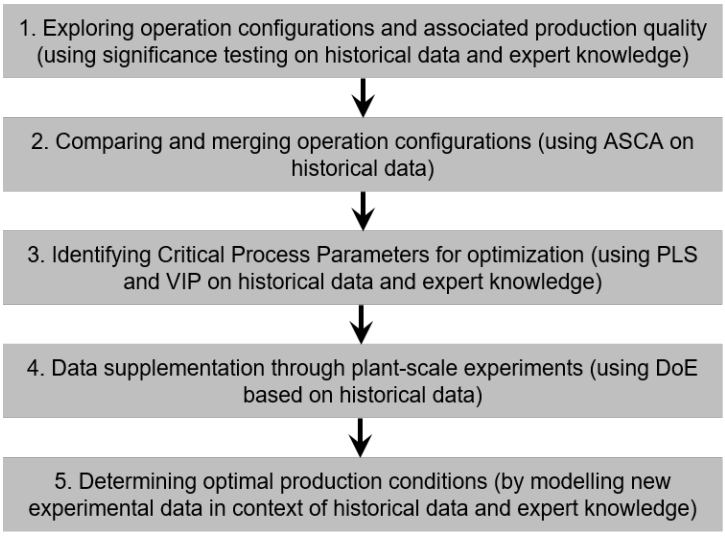
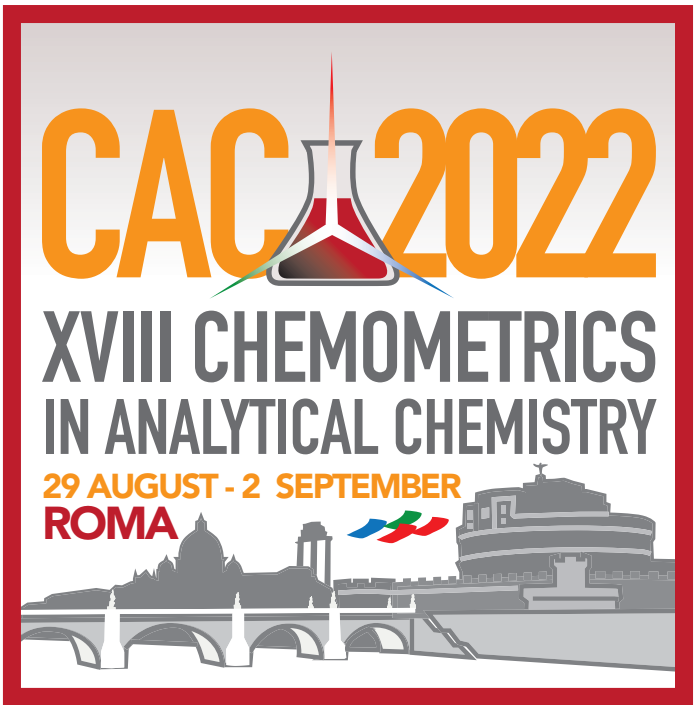


Figure 1 – The 5 steps of the proposed rQbD strategy

References

[1] A. Rathore, H. Winkle, *Nat. Biotech.* **27** (2009) 26-34.
[2] B. Silva et al, *Int. J. Pharm.* **528** (2017) 655-663.
[3] A. Galí et al, *Pharm. 2020*. **12** (2020) 743
[4] A. Boulesteix, K. Strimmer *Brief. Bioinform.* **8** (2007) 32-44
[5] C. Bertinetto, J. Engel, J. Jansen, *Anal. Chim. Acta X.* **6** (2020) 100061
[6] R. Leardi, *Anal. Chim. Acta.* **652** (2009) 161-172


CAC 2022
XVIII CHEMOMETRICS
IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

POSTERS

Authenticity of almond flour using handheld near infrared instruments and one class classifiers

José Marcelino Netto[1], Fernanda A. Honorato[2], Maria Fernanda Pimentel[2]
[1]*Universidade Federal de Pernambuco, Departamento de Química Fundamental, Recife, Brazil*
[2]*Universidade Federal de Pernambuco, Departamento de Engenharia Química, Recife, Brazil*
fernanda.honorato@ufpe.br

Almond is a nutritious and widely consumed seed, mainly found in powder form (flour). Because it is a product of high economic value, it has become the target of fraud with the illicit objective of increasing profit. To adulterate almond flour, offenders add low-cost products, obtaining mixtures with lower nutritional value [1]. Depending on the ingredient added, the identification of the fraud is not trivial. In the present work, fast, reliable and non-destructive methods using near-infrared spectroscopy (NIRS) were developed to identify the authenticity of almond flours. Three different handheld NIR instruments (DLPR NIRscanTM Nano: 900 - 1700nm; MicroNIR™: 950 -1650 nm; and NeoSpectra: 1350 - 2500 nm) were evaluated to verify the authenticity of almond flours and the results were compared with a benchtop FT-NIR (FT-IR Frontier - Perkin Elmer: 900 - 2500 nm). Fifty-four almond flours of different varieties and granulometries were acquired in local markets and adulterated with low-cost flours widely consumed in Brazil, such as cassava flour, oat, peanut and flour mixtures, totaling one hundred and twenty-four samples. Soft independent modelling of class analogies (SIMCA), data-driven SIMCA (DD-SIMCA) and partial least squares of a class (OCPLS) were used as multivariate classification methods, all based on a class strategy to discriminate authentic flours from adulterated [2]. The classification results for all evaluated instruments achieved 100% sensitivity and more than 95% specificity for samples with an adulterant concentration of 5% (w/w) or higher using DD-SIMCA and OCPLS (Figure 1). The results of the classification models indicate that portable NIR instruments are an efficient tool to discriminate between the adulterated and unadulterated samples, enabling in situ analysis, as well as having low cost when compared to benchtop FT-NIR instruments.

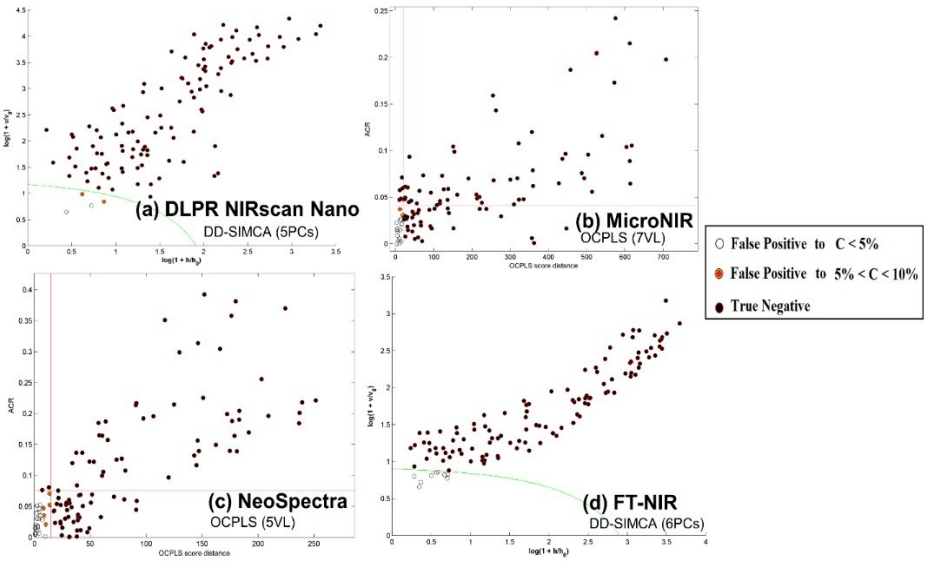


Figure 1 – Prediction results for external set of almond flours adulterated using DD-SIMCA and OCPLS for portable NIRS (a,b,c) and FT-NIR (d).

References

[1] C.S. Tibola, S.A, da Silva, A.A. Dossa, D.I Patrício. Economically Motivated Food Fraud and Adulteration in Brazil: Incidents and Alternatives to Minimize Occurrence. Journal of Food Science, **83** (2018) 2028-2038.
[2] R.G. Brereton. One-class classifiers. J. Chemometrics. **25** (2011) 225-246.

Unsupervised calibration transfer between spectrometer and hyperspectral camera: challenge proposed at the congress "Chimiométrie 2022"

Florent Abdelghafour [1], Maxime Ryckewaert [1,4], Matthieu Lesnoff [2,4], Jean-Michel Roger [1,4], Vincent Baeten[3], Pierre Dardennes [3].

[1] ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France
[2] UMR SELMET, Univ Montpellier, CIRAD, INRA, Institut Agro, Montpellier, France
[3] CRA-W, Gembloux, Belgique
[4] ChemHouse Research Group, Montpellier, France

Abstract:
The annual congress of the French Chemometrics Society, "Chimiometrie 2022" proposed a challenge consisting in characterising samples of wheat flour from NIR data. The peculiarity of the problem is that the calibration data were acquired with a laboratory spectrometer whereas the test data were acquired with a hyperspectral camera. It is then a calibration transfer problem where there are no reference standard between the spectra and Y for the target instrument (data available at https://chimiobrest2022.sciencesconf.org). In addition, the chemical nature of the Y reference value was undisclosed (protein content). The proposed solutions combine various pre-processing and readjustment – realignment methods between the source (spectrometer) and the target (hyperspectral imager). In addition, several unsupervised calibration transfer method were tested. These methods are based on orthogonal projection and are derivation of the EPO-PLS framework [1]. The methods differ from the way the difference matrix is estimated, but all share the purpose of modelling an invariant domain. The first method is based on the difference between processed mean spectra, the second method estimates D by bootstrap sampling, and the last one by unsupervised DOP [2]The proposition also includes a comparison of methods to summarise predictions at the scale of an image, by detecting outlier decisions and weighting pixel decisions.

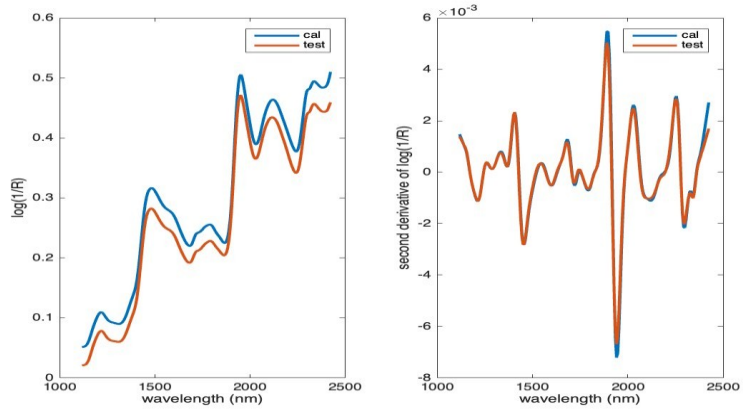


Figure : Realignement of source(specrometer in blue) and target (HSI in red) for raw spectra (left) and derived spectra (right)

References:

[1] .J-M. Roger, F. Chauchard, V. Bellon-Maurel, Epo–pls external parameter orthogonalisation of pls application to temperature-independent measurement of sugar content of intact fruits, Chemometrics and Intelligent Laboratory Systems 66 (2) (2003) 191–204. doi:https://doi.org/10.1016/S0169-7439(03)00051-0
[2] V. Fonseca Diaz, J-M. Roger, W. Saeys, Unsupervised Dynamic Orthogonal Projection. An efficient approach to calibration transfer without standard samples. May 2022. [preprint]

Odor concentration predictive model based on the odor activities of odorants produced by a municipal solid waste odor abatement scrubber

R. Aigotti[1], M. Guercio[2], F. Formigaro[2], F. Dal Bello[1], E. Davoli[3], C. Medana[1]
[1]*Turin University, Molecular Biotechnology and Health Sciences Department, Turin, Italy*
[2]*IREN Laboratori S.P.A., Turin, Italy*
[3] *Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy*
riccardo.aigotti@unito.it

The objective of this research project was the design and development of an integrated model for odor concentration estimation in a Municipal Solid Waste (MSW) odor abatement treatment plant used as a case study. Direct estimation of odor concentration through olfactometry analysis is not always exhaustively applicable [1].
Odorous samples were collected repeatedly at the stack emissions either before and after the scrubber and measured by dynamic olfactometry in accordance with standard EN 13725:2003 [2] by external laboratories. Chemical analysis of volatile organic compounds (VOC) was performed by SPME-GC/MS analysis. Odor sampling was performed filling several bags concurrently to perform simultaneously sensorial and chemical analysis.
The monitoring campaign was divided into two phases. During the first phase, three sampling sessions were carried out; thirty odorous samples were taken and subjected to chemical analysis. Olfactometry analysis was assessed by six different certified laboratories. To establish a correlation between chemical and sensory analysis, it was necessary to transform the VOC concentration into odor activity values (OAV) [3], using the ratio of the odorant concentration to its odor detection thresholds in air predicted from elemental physical properties of odorants [4], which account for the different relative contribution of each compound to the mixture total odor concentration. Forty VOC odorants concentrations were converted to corresponding OAV. Odorants OAVs data showed significant variability in space and magnitude. Multivariate supervised and unsupervised analysis (PCA, GDA) applied to OAVs, allowed to select four key odorants: limonene, toluene, dimethyl disulfide and butanol. The study of the olfactometric analysis response factor (olfactory threshold factory response to butanol [2]) showed variations up to an order of magnitude. A lack of reproducibility with dynamic olfactometry results was observed. Geometrical mean of measured OC was used, while odorant OAVs were ranked in seven chemical families plus the sum of the odor activity value (SOAV). Finally log transformed and mean centered data set of OAVs were then correlated to OC by linear partial least squares (PLS) regressions. Analysis of the quality parameter of the developed PLS model showed a moderate correlation (R^2 (Cal,Val): 0.73,0.57). Prediction was affected by specific error factors consisting in negative offset of the intercept and a significative validation slope bias.
In the second phase of the study the sample population was increased, and control samples were introduced. Control samples were gas mixtures of the key odorant VOC at known concentration. Control samples were prepared at the same time as the real samples are collected then delivered to the sensory and chemical measurement laboratories for simultaneous analysis. The control samples at scalar concentrations were used as reference standards to calibrate the key odorants chemical classes with SPME-GCMS method. Orthogonal PLS data elaboration was applied to control samples used to calibrate the model and the 73 real sample used as validation and test set. The sum of all OAVs yielded linear correlations against OC (R^2(Cal,Val): 0.92,0.90). O-PLS enhanced OC predictions. OAVs explained 90% of the OC variance throughout the PLS model validation. OC values regressed by the PLS model were less underestimated (10%), bias validation slope was compensated. Furthermore, O-PLS model showed good robustness parameters, prediction confidence contains 90% of measured OC. Based on result of PLS analysis the main contribution to OC description is given by terpenes and mercaptans.

References

[1] S. Stöckel, J. Cordes, B.toffels, and D. Wild, Environ Sci Pollut Res Int. 2018; 25: 24787–24797
[2] European Standard EN 13725:2003, 70p. European Committee for Standardization
[3] Wu, C., Liu, J. Zhao, P., Piringer, M. & Schaub, G. Atmos. Environ., 127, pp. 283–292, 2016.
[4] O. Rodríguez, * M. A. Teixeira and A. E. Rodrigues; Flavour Fragr. J. 2011, 26, 421–428

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

ATR-MIR AND MCR-ALS AS A TOOL FOR MONITORING WINE ALCOHOLIC FERMENTATION AND DETECTING BACTERIAL SPOILAGE

J. Cavaglia[1], S. Mas-Garcia[2,3], J.M. Roger[2,3] M. Mestres[1], R. Boque[4]

[1]Universitat Rovira i Virgili. Instrumental Sensometry (iSens), Dept. of Analytical Chemistry and Organic Chemistry, Tarragona, 43007, Catalonia, Spain
[2]ITAP, Univ. Montpellier, INRAE, Institut Agro, 34196 Montpellier, France
[3]ChemHouse Research Group, 34196 Montpellier, France
[4]Universitat Rovira i Virgili. Chemometrics, Qualimetrics and Nanosensors Group, Dept. of Analytical Chemistry and Organic Chemistry, Tarragona, 43007, Catalonia, Spain
ricard.boque@urv.cat

A methodology is proposed to describe the evolution of the main chemical compounds of grape must during wine alcoholic fermentation using Attenuated Total reflectance Mid Infrared (ATR-MIR) spectra in combination with Multivariate Curve Resolution Alternating Least-Squares (MCR-ALS). In addition, we have developed a process control strategy to detect differences between fermentations running under Normal Operation Conditions (NOC) and fermentations intentionally spoiled with lactic acid bacteria at the beginning of alcoholic fermentation (MLF) to promote process deviations from NOC.

MCR-ALS models on these data showed a good data fit (R^2 = 99.95% and lack of fit = 2.31%). It was possible to resolve the pure kinetic and spectral profiles of relevant molecules involved in the normal alcoholic fermentation process and of the molecule related to the promoted bacterial spoilage (lactic acid).

Multivariate Statistical Process Control (MSPC) charts were built based on the concentration profiles obtained from the MCR-ALS models and using T^2 and Q statistics. In this study, we developed MSPC charts based on the contaminated samples, rather than on NOC samples. When a sample is projected onto the model, if it falls under the “control” limits of the charts, it would mean the opposite from a traditional MSPC chart. Instead of being under control, it would mean that lactic acid is being produced and, therefore, the fermentation should be corrected to return to the normal conditions.

This methodology becomes an improvement of the traditional MSPC charts for the control of alcoholic fermentation, as it allows determining what is causing the possible deviation (in this study bacterial spoilage detection). Thus, if a fermentation batch is out of control in a traditional MSPC chart, but in control in this new ‘inverse’ MSPC chart, we could conclude not only that the sample is deviated, but also that the fermentation is deviated because of the production of lactic acid, as shown in the relative concentration profiles of MCR-ALS. By applying this methodology, spoiled wines showed off-limit values for T^2 after 96 hours, making it possible to detect lactic acid bacteria spoilage in early stages of alcoholic fermentation.

Therefore, the use of ATR-MIR spectra and MCR-ALS analysis shows a great potential for a rapid control of the state of the alcoholic fermentation process, making it possible to early detect the appearance of undesired molecules during the process, which allows the winemaker to apply corrective measures on time and obtain a good final product.

Acknowledgements
Project PID2019-104269RR-C33 funded by MCIN/AEI/10.13039/501100011033.

PREDICTION OF BEER SHELF LIFE USING AN HS-MS e-NOSE

Ana Carolina Lima, Laura Aceña, Montserrat Mestres, Ricard Boqué
Universitat Rovira i Virgili. Dept. of Analytical Chemistry and Organic Chemistry, Tarragona, Catalonia, Spain
ricard.boque@urv.cat

The aroma of beer is characterized by a complex mixture of volatile compounds that vary widely in nature and concentration levels. These chemical compounds are extracted from raw materials including water, malt, hops and yeast, during the brewing process. The role of each compound in the sensory perception of the product mainly depends on the ratio between its concentration level and its sensory threshold in the sample, which affects directly its odor activity value [1].

Beer flavor is one of the most determinant factors in its quality. However, during its shelf life, beer is subjected to chemical reactions that can affect flavor composition leading to a decrease in sensory quality. Due to its complexity, beer aging is considered the major quality problem for the brewing industry since the chemical reactions and the rate at which they occur during beer storage are determined by both internal factors (e.g. raw material, brewing techniques, oxygen content, pH, and key odorants precursors contents), and external factors (e.g. packaging, vibration, temperature and light) [2].

Generally, a sensory panel in combination with common analytical tools, such as gas chromatography, are used to evaluate beer flavor. These approaches are expensive and time-consuming and, in the case of sensory analysis, have the disadvantage of assessor fatigue and subjectiveness. We propose a faster and more objective alternative: the use of an electronic nose (e-nose) based on the application of the mass spectrometer to the headspace of the analyzed samples (HS-MS).

The objective of this study was to build a prediction model of beer shelf life by comparing the changes occurred in the volatile matrix of differently packaged samples during one year in optimal storage conditions of light and temperature. Thus, two commercial lager beers, packaged in aluminum cans and glass bottles, were analyzed by HS-MS over 12 months. Different chemometric strategies were applied to find the optimal spectral preprocessing and the best prediction model to relate the mass spectra to the aging months of the beers. The final model showed a good prediction ability and it can be used to predict the freshness of the product during its shelf life under optimal storage conditions despite the type of package used.

References:

1. Ghasemi-Varnamkhasti, M.; Mohtasebi, S.S.; Rodriguez-Mendez, M.L.; Lozano, J.; Razavi, S.H.; Ahmadi, H. Potential application of electronic nose technology in brewery. *Trends Food Sci. Technol.* **2011**, *22*, 165–174, doi:10.1016/j.tifs.2010.12.005.
2. Rodrigues, J.A.; Barros, A.S.; Carvalho, B.; Brandão, T.; Gil, A.M. Probing beer aging chemistry by nuclear magnetic resonance and multivariate analysis. *Anal. Chim. Acta* **2011**, *702*, 178–187, doi:10.1016/j.aca.2011.06.042.

Acknowledgements
Secretaria d'Universitats i Recerca of the Departament d'Empresa i Coneixement de la Generalitat de Catalunya (fellowship 2021FI_B2 00151) for financial support.

The NIR side of lentil

N. Cavallini[1], A. Giraudo[1], M. Sozzi[1], E. Cazzaniga[1], F. Geobaldo[1], F. Savorani[1]
[1]*Department of Applied Science and Technology (DISAT), Polytechnic of Turin, Turin, Italy*
nicola.cavallini@polito.it

The interest towards bioeconomy concepts has been considerably growing during the last years and, in particular, the development of sustainable and renewable bio-based technologies for food production is becoming increasingly important and studied. One of the most interesting applications of bioeconomy in the "food" area is the use of enzymes for the modification of food materials [1], to improve safety and to optimize the overall treatment processes.

In this perspective, the present study was focused on two processes for treating lentil flour: extraction and hydrolyzation, aimed at making protein available in solution. For both processes, an initial extraction phase with $Ca(OH_2)$ at controlled pH = 8 and fixed temperature of 60 °C was performed. Regarding only the hydrolyzation process, 0.2 % of protease enzyme was also added. The two processes were then carried out with two experimental runs each, differing by stirring rate (60 rpm and 120 rpm). A total of 32 samples of the processed solutions were collected at fixed time points in the range 0–300 minutes and stored frozen upon spectroscopic analysis. All samples were then analysed with a FT-NIR spectrometer (MPA by Bruker) and a Visible spectrometer (Carey by Agilent).

The acquired data were imported under MATLAB environment to undergo data quality assessment aimed at removing clear outliers and at choosing the proper preprocessing, specific for each dataset. The NIR data were preprocessed using standard normal variate (SNV) and mean centering, while the Visible data were only mean centered. The noisy regions with also low variance were removed from both datasets.

The datasets were explored by means of PCA, with the aim of obtaining information related to the type of process (extraction only *vs* extraction + hydrolyzation) and its evolution in time.

Curiously, the NIR data provided less clear information compared to the Visible data, with which interesting trends could be more easily identified. For this reason, a low-level data fusion approach was also put in place, by directly joining the two spectral datasets, after proper preprocessing and after the application of block scaling to give both blocks the same variance. The detected trends were confirmed, suggesting that both datasets provided useful information that could be efficiently combined and extracted.

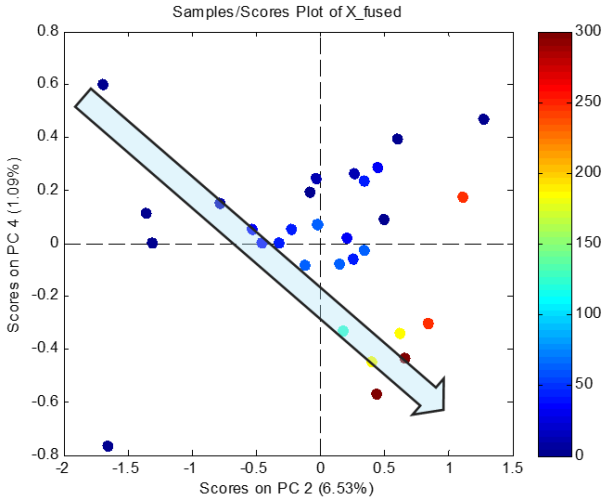


Figure 1 – Interesting time trend detected in the Visible dataset PCA.

These results suggest that Visible spectroscopy could be very useful for monitoring the extraction and hydrolyzation processes, while the application of NIR spectroscopy to this specific process and experimental setup would probably need further investigations.

References

[1] O. L. Tavano, Journal of Molecular Catalysis B: Enzymatic, 2013, 90, 1-11.

EXPLOITING PESTO SAUCE BY SEVERAL ANALYTICAL PLATFORMS: LOOKING FOR MOST EFFICIENT INFORMATION EXTRACTION AND DATA FUSION APPROACH

A. D'Alessandro[1,2], L. Strani[2], C. Durante[2], S. Mas[3], J.M. Roger[3], M. Cocchi[2]
[1]*Research manager at Barilla G. e R. Fratelli, Via Mantova 166, 43122 Parma, Italia*
[2]*Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi 103, 41125 Modena, Italia*
[3]*INRAe ChemHouse, 361 rue Jean-François Breton, 34196 Montpellier, France*
41125 Modena, Italia
alessandro.dalessandro@unimore.it

"Pesto alla Genovese" sauce is well known and appreciated this work regards an extensive characterization of industrial produced Pesto by Barilla. More than twenty samples of Pesto were selected from the whole one-year production, covering three different raw material, i.e. basil, types.
Since company can afford a limited number of quality analysis in routine, the aim of the work was to identify the most effective analytical technique to highlight and describe samples differences with respect to basil category.
In order to fully explore the compositional profile several analytical techniques were used to characterize the aromatic flavor pattern, the less volatile molecules and the physical structure of Pesto.
Flavor profile was determined mainly collecting the volatile organic compounds by head space sampling and then, measuring them by gas chromatography coupled to different detectors: Ion Mobility Spectrometer (HS-GC-IMS); Flame Ionization Detector (HS-GC-FID), namely ultrafast GC based e-nose (Heracles device). The less volatile molecules were studied by HPLC-MS using an untargeted approach.
Pesto physical structure was characterized by viscosity measurements and by calculating the Stability Index by Lumisizer.
Moreover, aiming at a fast characterization for future routine analysis and potentially transferable on-line, near infrared spectroscopy (NIRS) was also employed, at the moment off-line.
The resulting datasets were studied singularly, either by exploratory multivariate data analysis, e.g., Principal Component Analysis or by resolution techniques, e.g., Multivariate Curve Resolution. Finally, different data fusion approaches were applied for data integration.
Results have been evaluated to select the putative markers for pesto differentiation, and consequently the most efficient analytical technique/s, to select in order to perform sample characterization.

Application of chemometric approaches to answer some archeological questions for the study of the Apulian Red-Figure Pottery

T. Forleo[1], L.C. Giannossa[1,2], A. Mangone[1,2]
[1]*Department of Chemistry, University of Bari "Aldo Moro", Bari, Italy*
[2]*Laboratorio di ricerca per la diagnostica dei BBCC, University of Bari "Aldo Moro", Bari, Italy*
tiziana.forleo@uniba.it

One of the most impressive productions of the Magna Grecia is the Southern Italian variant of the famous Attic production, the so-called Apulian Red-Figure Pottery (from 440 BCE to 300 BCE) [1]. The artifacts were in the spotlight of museums and collectors, so much to cause an increase in excavations, sales and on the black market [1,2]. Often the clandestine transport was facilitated by crushing the vessels into small pieces. Once the fragments arrived at their destination, they were reassembled generally with unappropriated methods and sometimes were completed often with non-original pieces [1,2].

From the archeological point of view, these artworks raise many questions involving different areas such as the authenticity, the provenance (import or local production), the production technique [1]. Unsupervised and supervised pattern recognition techniques were used to study and classify several fragments of these artifacts, analyzed by AAS, ICP-OES, ICP-MS and LA-ICP-MS.

Hierarchical non-linear principal component analysis (NL-PCA), based on auto-associative neural networks (ANN), was used as exploratory analysis [3], and Discriminant Analysis [4,5] and Support Vector Machine (SVM) [6] as classification methods.

The chemometric methods demonstrate to be adequate to examine the possible classification of those findings and answer several questions regarding the authenticity, the geographical origin, the workshop location, and the technology used and extract the most discriminant features.

References

[1] L.C. Giannossa, T. Forleo,A. Mangone, *Appl. Sci.* **11** (2021) 3073
[2] Amineddoleh, *Art Antiq. Law* **18** (2013) 227-254
[3] A. Bitetto, A. Mangone, R.M. Mininni, *J.Chemom.* **30** (2016) 405-415
[4] F.Marini, *Curr.Anal.Chem*, **6** (2010) 72-79
[5] D.Ballabio, V.Consonni, *Anal.Methods*, **5** (2013) 3790
[6] J.Luts, F.Ojeda, R.Van de Plas, *Anal.Chim.Acta* **665** (2010) 129-145

LEVERAGING AN INTEGRATED SENSOR ARRAY AND MACHINE LEARNING TO ACCELERATE SENSORY EVALUTION OF COFFEE

G. Gabrieli[1], M. Muszynski[1], E. Thomas[2], D. Labbe[3], P. W. Ruch[1]
[1]*IBM Research Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*
[2]*Nestlé Nespresso S.A., Chaussée de la Guinguette 10, Vevey, 1800, Switzerland*
[3]*Société des Produits Nestlé Nespresso S.A., Nestlé Institute of Material Sciences, Route du Jurat 57, Lausanne, 1000, Switzerland*
gga@zurich.ibm.com

Electronic tongues (ETs) comprising an array of low-selective and cross-sensitive sensors have been proposed to enable rapid analysis of a wide range of food and beverage products [1]. Moreover, by unveiling correlations between sensor array response and descriptive sensory analysis, such technology has been employed in sensory evaluation for food authentication and quality control processes [2]. Instead, the use of ETs to analyze new product formulations has been limitedly explored despite being one application for which such tools could be potentially disruptive in terms of cost and time efficiency. In the current study, we demonstrate how a remarkably simple sensor configuration [3] can be combined with a prescriptive training scheme and machine learning models to provide a quantitative estimate of sensory descriptors associated with coffee samples. In particular, the same sensor array was used to test 33 different coffees and characteristic features were extracted from transient differential voltages arising during the transition of the sensor array from a reference solution to a coffee sample. We propose a pipeline (Fig. 1) based on training of a single regression model to predict principal components of sensory descriptors. The pipeline aims at diminishing the complexity of model training and reconstructing sensory profiles from predicted principal components by using an inverse transformation.

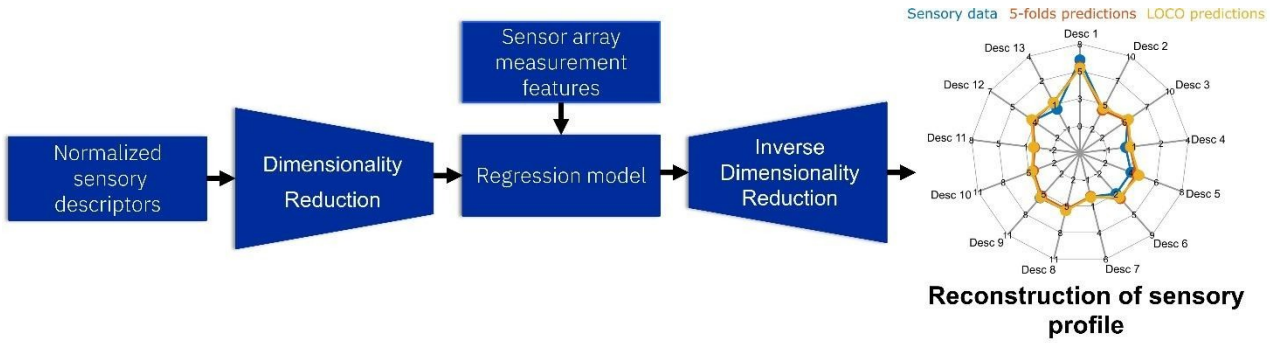


Figure 1 – Data pipeline for simultaneous prediction of 13 coffee sensory descriptors.

A simple artificial neural network (ANN) architecture captured the non-linear response of the proposed sensor array configuration yielded higher performances compared to multiple linear regression (MLR) with an overall RV coefficient of agreement between true and predicted sensory descriptors of 0.78 using a leave-one-coffee-out (LOCO) cross-validation. Leveraging automated testing pipelines allow researchers to estimate the sensory profile of a coffee in less than 2 minutes. Thus, data-driven sensor arrays could also be used to accelerate screening of new product formulations, supporting descriptive analysis performed with a sensory panel.

References

[1] M.Podrazka, E.Báczyńska, M. Kundys, P.S. Jeleń, E.W. Nery, Emilia, *Biosensors* **8** (2017) 1-24.
[2] B. ouad , J. . . aukuu, . al s, . Bodor, . e er, Z. Gillay, G. Bazar, Z. Kovacs, *Sensors* **20** (2020) 1–42.
[3] G. Gabrieli, R. Hu, K. Matsumoto, Y. Temiz, S. Bissig, A. Cox, R. Heller, A. López, J. Barroso, K. Kaneda, Y. Orii, P.W. Ruch, *Anal. Chem.* **93** (2021) 16853-16861.

INSIGHTS INTO MULTIVARIATE DATA ANALYSIS FOR REAL-CASE FERMENTATION PROCESS WITH MINIATURIZED NIR SPECTROSCOPY

Barbara Giussani[1], Alberto Ferrer[2], Giulia Gorla[1]
[1]*Department of Science and High Technology, University of Insubria, Como, Italy*
[2]*Multivariate Statistical Engineering Group, Department of Applied Statistics and Operational Research, and Quality, Valencia Polytechnic University, Valencia, Spain*
barbara.giussani@uninsubria.it

Nowadays, paving the way for real-time monitoring of the process is of rising interest for several application fields. In particular, in recent years growing attention is focused and the use of low cost miniaturized NIR spectrometers [1,2]. One of the main industries that are increasingly devoted to the use of portable near-infrared spectroscopy as a quality assessment method or process analytical tool is the dairy industry [3]. Once the strategy of acquisition is implemented, the coupling between spectra and several multivariate technique approaches are usually employed and investigated to achieve all the information contained in the data and to summarize and illustrate them.

In this study, the strengths and limitations of the different multivariate approaches are highlighted and discussed for a real case example. Kefir fermentation was conducted at different temperatures and mixing conditions. After adding the kefir grains to semi-skimmed milk, spectra were acquired at interval times of 30 minutes during the evolution of the process. Since pH is a usual indicator of the process progress it was also measured.

Initially, spectra features and Principal Component Analysis (PCA) was used to identify the typical trend of the fermentation processes and the main changes that occurred under different condition. Then, Moving Window Principal Component Analysis (MW-PCA) [4] was applied in the spectral range of 1350–2550 nm to detect the main variations in loading caused by the fermentation and the trend of the dissimilarity index was used to identify different phases of the process. At this point, the main wavelength regions related to physical and chemical changes were identified. The investigation of batch-wise and variable-wise unfolded PCA on the array of data and the possibilities offered by Partial Least Squares (PLS) modelling were so investigated [5] .

References

[1] K.B. Beć, J. Grabska, C.W. Huck, Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers, Chem. - A Eur. J. 27 (2021) 1514–1532. https://doi.org/10.1002/chem.202002838.
[2] B. Giussani, G. Gorla, J. Riu, Analytical Chemistry Strategies in the Use of Miniaturised NIR Instruments: An Overview, Crit. Rev. Anal. Chem. (2022) 1–33. https://doi.org/10.1080/10408347.2022.2047607.
[3] Y. Pu, D. Pérez-Marín, N. O'shea, A. Garrido-Varo, Recent advances in portable and handheld NIR spectrometers and applications in milk, cheese and dairy powders, Foods. 10 (2021). https://doi.org/10.3390/foods10102377.
[4] J. Muncan, K. Tei, R. Tsenkova, Real-time monitoring of yogurt fermentation process by aquaphotomics near-infrared spectroscopy, Sensors (Switzerland). 21 (2021) 1–18. https://doi.org/10.3390/s21010177.
[5] S. Wold, N. Kettaneh-Wold, J.F. MacGregor, K.G. Dunn, Batch Process Modeling and MSPC, in: Compr. Chemom. Chem. Biochem. Data Anal. VOLS 1-4, Elsevier, 2009: pp. A163–A197.

Identification of metal ions with the use of quantum dots coupled with excitation-emission matrix fluorescence spectroscopy

K. Głowacz[1], P. Ciosek-Skibińska[1]
[1]*Chair of Medical Biotechnology, Faculty of Chemistry, Warsaw University of Technology, Noakowskiego 3, 00-664 Warsaw, Poland*
kglowacz@ch.pw.edu.pl

Quantum dots (QDs) are semiconductor nanocrystals, which due to their unique photoluminescent properties, have become attractive nanomaterials used in biomedical applications, including bio-labeling, bio-imaging, and bio-targeting. Moreover, as a consequence of their versatile surface chemistry and ligand binding ability, QDs have also been exploited in analytical chemistry, primarily as optical sensors for the detection of small molecules. Since metal ions are an essential analytical target in, e.g., environmental or biological samples, it is not surprising that many QDs-based optical sensors for their detection have been developed over the years [1]. The use of quantum dots for the qualitative/quantitative analysis of metal ions is most often based on the observation of the change in their fluorescence signal under the influence of selective interaction with an analyte. However, applying such an approach for detecting multiple analytes in a sample requires the design of a highly selective receptor element for each analyte, thus representing a major synthetic challenge. To resolve this issue, an alternative sensing strategy might be employed, namely pattern-based sensing, which relies on the design of receptors that differentially interact with various analyte components or, in some cases, on the application of one sensor in combination with detection techniques that provides multidimensional optical information about the composition of the tested sample [2,3]. However, it must be noted that due to the multidimensionality of the data produced by pattern-based sensing systems, chemometric modeling is required to detect an analyte.

In the framework of this study, we propose a simple, pattern-based sensing system employing thiomalic acid (TMA) capped CdTe quantum dots for the identification of metal ions. The presented sensing strategy is based on the fact that selected metal ions exhibit different quenching mechanisms of the QDs' fluorescence, manifested by their diverse influence on the fluorescence spectrum of this nanomaterial (i.e., the degree of the fluorescence quenching, spectral shifts). By utilizing the excitation-emission matrix (EEM) fluorescence spectroscopy as a detection technique, we captured the information on the alterations of the fluorescence signal of QDs caused by metal ions and showed how chemometric modeling of obtained excitation-emission matrices can be used for the identification of selected analytes.

This work was financially supported by National Science Centre (Poland) within the framework of the SONATA BIS project No. UMO-2018/30/E/ST4/0048. Klaudia Głowacz acknowledges financial support from IDUB project (Scholarship Plus programme).

References

[1] Y. Lou, Y. Zhao, J. Chena, Jun-Jie Zhu, *J. Mater. Chem. C*, 2 (2014) 595.
[2] K. Głowacz, M. Drozd, P. Ciosek-Skibińska, *Microchim Acta* 188 (2021), 343.
[3] P. Wu, L.N. Miao, H.F. Wang, X.G. Shao, X.P. Yan, *Angew Chemie - Int Ed*, 50 (2011), 8118–8121.

Evaluation of preprocessing strategies for LCMS data using R

J.Hansen[1], S. Seifert[1]
[1]*University of Hamburg, Hamburg, Germany*
jule.hansen@chemie.uni-hamburg.de

Untargeted LCMS analysis has vast potential to generate specific metabolomic fingerprints that can be exploited in various research fields. For example in food authenticity testing, there is an increased need to constantly find unknown adulterations.
For the processing of the raw data, a wide range of software tools is available. The required steps conducted by those tools are data conversion in an appropriate format, data visualization, peak detection, alignment, correspondence and peak annotation. In addition, to improve the comparability of single measurements, normalization is indispensable.
This poster gives an overview and compares and evaluates packages suitable for LCMS data processing including well-established packages like mzR and Msnbase [1,2] as well as xcms [3] and MAIT [4].
For many applications, however, NMR approaches are used instead of highly sensitive LCMS methods because they are more robust. Therefore, we also discuss different normalization strategies to enhance the robustness of LCMS measurements.

References

[1] L. Gatto, S. Gibb, J. Rainer, J. Proteome Res. 2021, 20, 1063.
[2] L. Gatto, A. Christoforou, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 2014, 1844, 42.
[3] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 2006, 78, 779.
[4] F. Fernández-Albert, R. Llorach, C. Andrés-Lacueva, A. Perera, Bioinformatics 2014, 30, 1937.

Spectral identification of therapeutic allergen products

P. Rani[1,2], K. Tsenova[2], J. Rost[1], F. Führer[1], D. Bartel[1], C.Kamp[1,2]
[1]*Paul-Ehrlich-Instiut, Langen, Germany*
[2]*Goethe University, Frankfurt, Germany*
christel.kamp@pei.de

Raman spectroscopy is widely used to identify chemical compounds and the quality of pharmaceutical products. We explore the potential of this technique for biomedicines like vaccines or therapeutic allergens [1-3] that introduce new challenges in terms of experimental setup and standardisation. Therapeutic allergen products are derived from source material of biological origin with inherent variability between product batches and with excipients strongly contributing to the spectral signal. The differences of interest between spectra of products with different active components or from different manufacturers are often minor compared to those introduced by other sources of variation. This underlines the requirement for careful and standardised procedures including the experimental setup and the subsequent analysis of spectra. We demonstrate this based on a case study conducted on various batches of therapeutic allergen products containing bee and wasp venoms from two different manufacturers. Starting from an overview of common spectral features and differences we explore the impact of spectral pre-processing and subsequent classification of spectra. In particular, we focus on the spectral decomposition and analytic signal reduction of excipient spectra and discuss the potential of variant computational methods to distinguish products with different active components and from different manufacturers.

References

[1] C. Kamp, B. Becker, W. Matheis, V. Öppling, I. Bekeredjian-Ding, Biol. Chem. 402(8) (2021) 1001-1006.
[2] A. Silge, T. Bocklitz, B. Becker, W. Matheis, J. Popp, I. Bekeredjian-Ding, NPJ Vaccines 3:50 (2018)
[3] V. Mahler, R.E. Esch, J. Kleine-Tebbe, W.J. Lavery, G. Plunkett, S. Vieths, D.I. Bernstein, J. Allergy Clin. Immunol. 143(3) (2019) 813 – 828.

Comparison between colloidal and volatile profiles to create a chemometric model to classify different tomato sauce brands

N. Kassouf[1], A. Zappi[1], D. Melucci[1], V. Marassi[1]
[1]*Departiment of chemistry, Bologna, Italy*
nicholas.kassouf@studio.unibo.it

The present work concerns the analysis of the volatile fraction and the colloidal portion of tomato sauce. This preliminary work aims to create an analytical and chemometric method that can group tomato sauces based on the different brands. 19 bottles of tomato sauces pertaining to 6 different brands were collected and analyzed.
The volatile fraction was analyzed by gas-chromatography coupled to ion mobility spectroscopy (GC-IMS), which analyses the headspace (HS) of the samples. After the classic gas chromatographic (GC) column, the volatile molecules are further separated by the ion mobility spectrometer (IMS), which ionizes and separates them by the application of an electric field [1].
The colloidal fraction was instead analyzed by Asymmetric Flow Field-Flow Fractionation (AF4). Protein, polysaccharides, and polyphenols are the main components of the colloidal fraction of the tomato sauce. AF4 is an innovative and high-tech instrument able to analyze colloids, using two different flows: the first one is the carrier-flow, which transports the analytes, and the separation is generated by the cross-flow, which is perpendicular to the carrier-flow, it produces a gradient that can create a separation based on the different colloidal hydrodynamic radius [2].
The datasets obtained with the HS/GC-IMS and AF4 analyses were elaborated by Principal Component Analysis (PCA). The scores made it possible to observe a similar trend in the brand grouping of the two different datasets, allowing to infer that the volatile and the colloidal fraction of tomato sauce carry very similar information.

[1] Wang S, Chen H, Sun B. Recent progress in food flavor analysis using gas chromatography–ion mobility spectrometry (GC–IMS). *Food Chemistry*. 2020;315:126158. doi:10.1016/J.FOODCHEM.2019.126158
[2] Marassi V, Marangon M, Zattoni A, et al. Characterization of red wine native colloids by asymmetrical flow field-flow fractionation with online multidetection. *Food Hydrocolloids*. 2021;110:106204. doi:10.1016/J.FOODHYD.2020.106204

Multivariate analysis of colloidal and volatile profiles for class-modeling of different tomato sauce brands

N. Kassouf[1], A. Zappi[1], F. Gottardi[2], V. Marassi[1], S. Giordani, D. Melucci[1]
[1]*Departiment of chemistry, Bologna, Italy*
[2]*COOP Italia, R&D Laboratory, Bologna, Italy*
nicholas.kassouf@studio.unibo.it

Tomato sauce is a widely available product, with an important market weight, and one of the Italian typical products in the food industry: its classification and grouping in terms of brands, origin, and processing pipelines (organic, untreated…) represent an important goal to ensure quality, avoid adulteration, and recognize local excellencies.
This preliminary work aims to create an analytical and chemometric method that can group tomato sauces based on the different brands, analyzing the volatile fraction and the colloidal portion of tomato sauce. For this study, nineteen bottles of tomato sauces pertaining to six different brands were collected and analyzed.
The volatile fraction was analyzed by headspace (HS) gas-chromatography [1] coupled with ion mobility spectroscopy (GC-IMS). After eluting from the classic gas-chromatographic (GC) column, the volatile molecules were further separated by an ion mobility spectrometer, which ionizes and separates them by applying an electric field [2].
The colloidal fraction was analyzed by Asymmetrical Flow Field-Flow Fractionation (AF4). UV Proteins, polysaccharides, and polyphenols are the main components of the colloidal fraction of the tomato sauce. AF4 is an innovative and high-tech instrument able to analyze colloids, exploiting the combined action of two different flows: the first one is the longitudinal carrier-flow, which transports the analytes, and the separation is generated by the cross-flow, which is perpendicular to the carrier-flow: it produces a gradient that can create a separation based on the different colloidal hydrodynamic radius[3].
The datasets obtained with HS/GC-IMS and AF4 analyses were elaborated by Principal Components Analysis (PCA). The scores made it possible to observe a similar trend in the brand grouping of the two different datasets, corroborating the hypothesis that the volatile and the colloidal fraction of tomato sauce carry very similar information. In particular, given the easiness of operations, greener solvents and reagents required, and faster analysis time, the results obtained from AF4 analysis could represent a more environmentally friendly approach towards chemometric-based tomato grouping.
In perspective, multi-block analysis could be performed, using both colloidal and volatile profiles.

[1] Forleo T, Zappi A, Gottardi F, Melucci D. Rapid discrimination of Italian Prosecco wines by head-space gas-chromatography basing on the volatile profile as a chemometric fingerprint. *European Food Research and Technology*. 2020;246:805-1816. doi: 10.1007/s00217-020-03534-8

[2] Wang S, Chen H, Sun B. Recent progress in food flavor analysis using gas chromatography–ion mobility spectrometry (GC–IMS). *Food Chemistry*. 2020;315:126158. doi: 10.1016/J.FOODCHEM.2019.126158

[3] Marassi V, Marangon M, Zattoni A, et al. Characterization of red wine native colloids by asymmetrical flow field-flow fractionation with online multidetection. *Food Hydrocolloids*. 2021;110:106204. doi: 10.1016/J.FOODHYD.2020.106204

A comparison between artificial neural networks and partial least squares for coffee assessment by high-resolution mass spectrometry

Victor Cardoso[1], Julia Balog[2], Tamas Karancsi[2], Guilherme Sabin[1,3], Leandro Hantao[1]
[1] *University of Campinas, Campinas, Brazil*
[2] *Waters Research Center, Budapest, Hungary*
[3] *OpenScience, Campinas, Brazil*
victor.cardoso@iqm.unicamp.br

Coffee is one of the most consumed beverages worldwide [1,2]. This product presents an important role in the global economy, mainly in Brazil which is the top producer [1,3]. The large consumption of coffee requires efficient quality control to ensure its quality, meeting the wide consumer demands [1-3]. Thus, business intelligence is mandatory in this process, providing coffee assessment through quick analysis and interpretation of large amount of data.

To achieve this objective, a sampler prototype based on laser ablation rapid evaporative mass spectrometry (LA-REIMS) was used to analyze coffee powder by high resolution mass spectrometry (HRMS). This prototype allowed to perform direct analysis of a sample per minute in the same wellplate with minimal sample preparation, seeking to supply the ascending demands of coffee industries.

For this study, 31 types of espresso coffee samples were used, with different sensory properties and processing parameters, such as bitterness, body, acidity, intensity, and roasting level. The dataset contains 54 spectra of each coffee type, being 1674 spectra in total. After suitable preprocessing, the samples were divided into training, validation, and test sets for the modeling process.

Partial least square discriminant analysis (PLS-DA) was applied for data treatment. However, the complexity of this dataset due to the mass spectra with several non-correlated variables and numerous coffee classes makes this classification a hard task [1,4]. Such PLS-DA models usually require many latent variables for a proper prediction, but its choice is not trivial due to under- and overfitting [1,4]. Also, numerous samples with large T^2 and Q residuals are commonly found while modeling this data by PLS-DA [1,5].

As an alternative to PLS-DA, artificial neural networks (ANN) in tandem with Bayesian optimization were applied to model this dataset. This approach presented an automated hyperparameter selection based on probabilistic estimations and classification errors, aiming to find the global optimal solution [5]. Also, ANN can easily deal with outliers, numerous classes and non-correlated variables, and massive data sets [6], which are issues found in this data set.

The comparison between PLS-DA and ANN presents numerous strengths and weaknesses in both. In this application, ANN presented better results to discriminate the coffee types, achieving 95% of accuracy against up 90% in PLS-DA. Regarding outlier samples, they only affected PLS-DA model since the high values of $\bar{T}$ and Q residuals were measured in this algorithm [5], while ANN is capable to deal with non-homogeneous classes. About variables importance, PLS-DA provides this information through variable importance in projection (VIP) score algorithm [5], while it still a limitation in ANN models. In other words, PLS-DA provides simpler model but easily interpretable, whereas ANN provides more complex models with even higher accuracy but coasting the model interpretation [6].

This study proved that the LA-REIMS prototype has provided reliable coffee analysis. The developed method provided quick, clean, and high-throughput analysis compatible with coffee industry demands. A challenge is the data treatment due to the large volume of data, but PLS-DA and ANN provided excellent results according to the different model demands. ANN provided higher accuracies, but PLS-DA provided better interpretability. We believe that in a near future, this method can be adapted and optimized to larger scales and as a routine analysis in coffee industries.

References

[1] V. Cardoso, G. Sabin, L. Hantao, *BrJAC.* **8** (2021) 91-106.
[2] M. Ayseli, H. Kelebek, S. Selli, *Food Chem.* **338** (2021) 127821.
[3] M. Baqueta, A. Coqueiro, P. Valderrama, *J. Food Sci.* **84** (2019) 1247-1255.
[4] V. Cardoso, G. Sabin, L. Hantao, *Anal. Methods.* **14** (2022) 1540-1546.
[5] V. Cardoso, R. Poppi, *Microchem. J.* **164** (2021) 106052.
[6] F. Marini, R. Bucci, A. Magrí, A. Magrí, *Microchem. J.* **88** (2008) 178-185.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Benchmarking Machine Learning approaches for hit detection in High-Content Screening

Erwin Kupczyk[1,2], Kenji Schorpp[3], Kamyar Hadian[3], Sean Lin[3], Dimitrios Tziotis[1], Philippe Schmitt-Kopplin[1,2] & Constanze Mueller[1]

[1]Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, Ingolstaedter Landstr. 1, 85764, Neuherberg, Germany
[2]Comprehensive Foodomics Platform, Chair of Analytical Food Chemistry, TUM School of Life Sciences, Technical University of Munich, Maximus-von-Imhof-Forum 2, 85354 Freising, Germany
[3]Institute for Molecular Toxicology and Pharmacology, Cell Signaling and Chemical Biology, Helmholtz Zentrum München, Ingolstaedter Landstr. 1, 85764, Neuherberg, Germany

erwin.kupczyk@helmholtz-muenchen.de

Abstract

Complex mixtures containing natural products are still an interesting source of novel drug candidates. High content screening (HCS) is a popular tool to screen for such. In particular, multiplexed HCS assays promise comprehensive bioactivity profiles, but generate also high amounts of data. Yet, only some machine learning (ML) applications for data analysis are available and these usually require a profound knowledge of the underlying cell biology. Unfortunately, there are no applications that simply predict if samples are biologically active or not (any kind of bioactivity). Within this work, we benchmark ML algorithms for binary classification, starting with classical ML models, which are the standard classifiers of the scikit-learn library or ensemble models of these classifiers (a total of 92 models tested). Followed by a partial least square regression (PLSR)-based classification (44 tested models in total) and simple artificial neural networks (ANNs) with dense layers (72 tested models in total). In addition, a novelty detection (ND) was examined, which is supposed to handle unknown patterns. For the final analysis the models, with and without upstream ND, were tested with two independent data sets. In our analysis, a stacking model, an ensamble model of class ML algorithms, performed best to predict new and unknown data. ND improved the predictions of the models and was useful to handle unknown patterns. Importantly, the classifier presented here can be easily rebuilt and be adapted to the data and demands of other groups. The hit detector (ND + stacking model) is universal and suitable for a broader application to support the search for new drug candidates.

Robust quantitative analysis in Laser-Induced Breakdown Spectroscopy (LIBS) using artificial neural networks

Qicheng Wu[1], Vincent Motto-Ros[2], Ludovic Duponchel[1],*
[1] *Univ. Lille, CNRS, UMR 8516 – LASIRE – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, 59000, France*
[2] *Institut Lumière Matière, UMR 5306, Université Lyon 1 - CNRS, Université de Lyon 69622 Villeurbanne, France.*
(*): ludovic.duponchel@univ-lille.fr

Laser-induced breakdown spectroscopy, short as LIBS, is an atomic emission spectroscopy technique, which uses a laser as an excitation source to atomize the sample which generates a plasma that exhibits emission lines of atoms and ions present at the sample surface. LIBS is a very fast method with easy sample preparation, it's capable of qualitative and quantitative analysis. However, the existence of matrix effects makes quantitative analysis sometimes a challenge. The characteristic peaks in the spectra of the same element may change in intensity and even disappear, due to matrix effect causing conventional linear quantitative methods to fail. In this work, a non-linear method, artificial neural networks (ANN), is introduced to analyze a mineral rock sample (Figure 1 left). ANN is a non-linear method for non-linear and complex problems and is capable of learning. The sample here has more than 2 million spectra as an image of 1400 by 1700 pixels with the resolution of 20µm is taken. The aim of this dataset is to understand for each pixel which element exists and how much. Simulated LIBS datasets were used to train the ANN, and then the real data was applied as a test dataset for ANN. The Kurucz database[1] was used to generate simulated data, LIBS spectra of 24 elements that commonly exist in rock samples were generated with 7 plasma temperatures(T) and 3 electron densities(Ne). After tuning and selecting netwroks, an ANN with classical structure (figure 1 middle) was applied, and the results of Si are shown below (figure 1 right). Something that needs to be mentioned here, to the right side of the ANN result is the reference map obtained by the conventional integration method, which has very high intensities. But the ANN result is relative, the two can't be compared directly, however, the reference maps are still very helpful to show the right direction for ANN analysis. At the stage of training with simulated data, the root mean squared error (RMSE) and regression coefficient (R) were used as figures of merit. The input of ANN are variables of spectra after normalization, and the output is the concentration in percentage of one element selected. From the results we have so far, the ANN results for 14 main elements are almost all within the range of 0 and 1. The next step would be trying to solve the small part of the ANN results that are outside the range of 0 and 1.

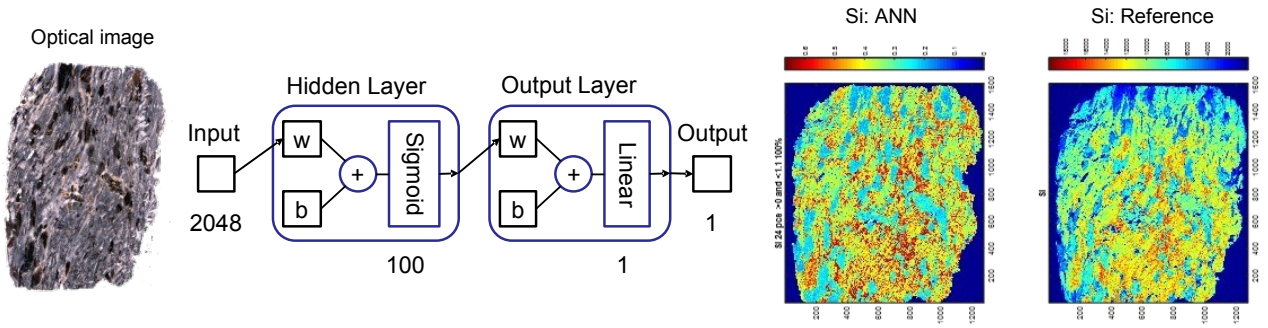


Figure 1 – Left: the optical image of the sample. Middle: the structure of the ANN applied. Right: the result from ANN of Si and the reference obtained from the conventional method.

References

[1] "Atomic Spectral Line Database from CD-ROM 23 of R. L. Kurucz.," accessed February 16, 2021, https://www.cfa.harvard.edu/amp/ampdata/kurucz23/sekur.html.

Analyzing multifactorial designed data from multiple sources with a single model using AComDim

M. de Figueiredo[1], S. Giannoukos[1], C. Wüthrich[1], R. Zenobi[1], D. N. Rutledge[2,3]
[1]*Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland*
[2]*National Wine and Grape Industry Center, Charles Sturt University, Wagga Wagga, Australia*
[3]*ChemHouse Research Group, Montpellier, France*
miguel.defigueiredo@unige.ch

A novel chemometric approach is proposed to analyze high-dimensional data collected from multiple analytical platforms sharing the same multifactorial design of experiments. Although methods combining analysis of variance (ANOVA) with principal components analysis (PCA) or simultaneous components analysis (SCA) efficiently tackle multifactorial designed data, they require the construction and interpretation of one model for each experimental factor and interaction, and each data source. Taking advantage of the multiblock nature of the Common Dimensions (ComDim) method, we propose an extension of the original method ANOVA-ComDim (AComDim) [1] for the analysis of multifactorial designs from one source to multiple sources [2]. As an example, Figure 1 describes the algorithm implementation for two data sources and a 2-fixed effect factors experimental design with 2 and 3 levels, respectively.

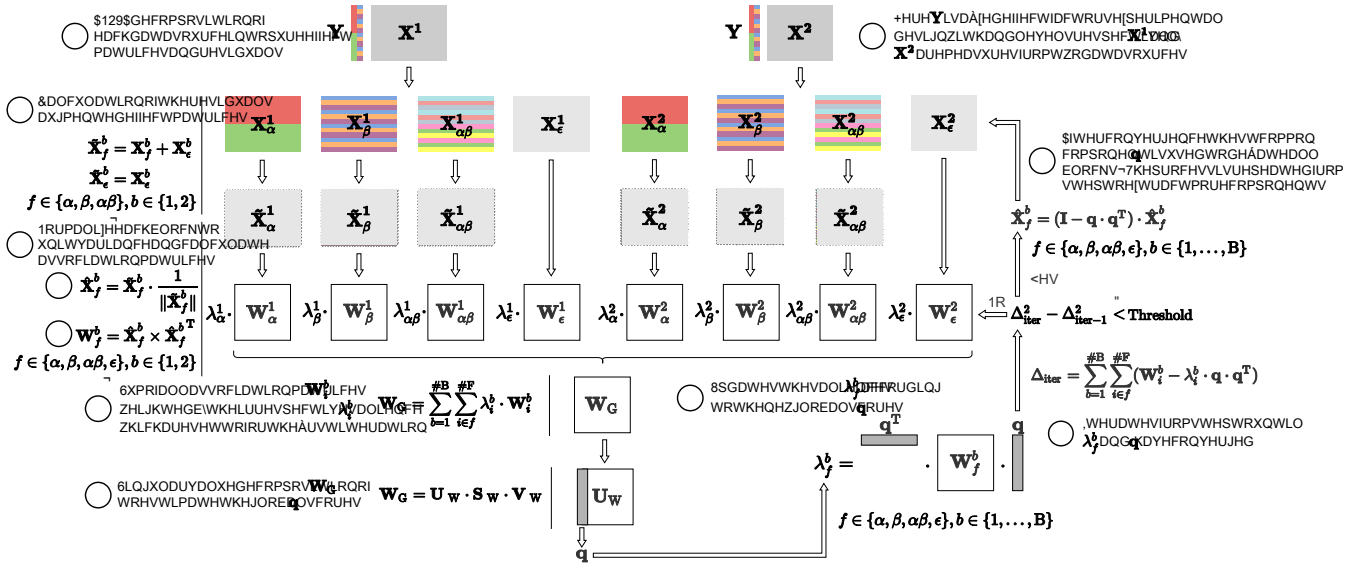


Figure 1 – AComDim algorithm implementation for the analysis of 2 data sources.

AComDim provides a global picture of the main sources of variation in complex data structures produced from multiple sources. To do so, AComDim builds a single model for all main effects and interactions instead of one for each one of them, thus helping the interpretation process. The saliences produced by AComDim are critical for the interpretation because they provide insights into the relationship between the data sources and their contribution to each of the effects.

An example of AComDim applied to multiple sources will be presented. The dataset consists of 8 sources (blocks) from a design with 3-fixed effect factors with 2, 2 and 5 levels, respectively.

References

[1] D. Jouan-Rimbaud Bouveresse, R. C. Pinto, L. M. Schmidtke, N. Locquet, D. N. Rutledge, Chemometr Intell Lab. **106** (2011) 173 – 182.
[2] M. de Figueiredo, S. Giannoukos, C. Wüthrich, R. Zenobi, D. N. Rutledge, J Chemomtr. (2022) e3401.

INVESTIGATING SOURCES OF VARIANCE IN MINIATURIZED NIR SPETROSCOPY: FIND CLUES AND SOLVE THE RIDDLE

Giulia Gorla[1], Paolo Taborelli[1], Barbara Giussani[1]
[1]*Department of Science and High Technology, University of Insubria, Como, Italy*
ggorla@uninsubria.it

Miniaturized NIR instruments have been increasingly used in the last years, and they have become useful tools for many applications on a broad variety of samples [1,2]. In this context, an investigation of the sources of variance that could be encountered while using handheld instrumentation is of interest for several reasons related to possible improvements in setting up experiments and data modelling phases and, on the top of that, to identify the limits of experimental measurements.
In this preliminary study, external reflectance spectra of different type of samples were acquired under different condition with a NeoSpectra Scanner spectrometer (1350 – 2550 nm) and an AvaSpec-Mini-NIR (900 – 1750 nm) with a fiber optic cable.
An experimental design data structure was obtained by considering factor like type of sample, charge condition, replicates and time of background acquisition. ANOVA-Simultaneous Component Analysis (ASCA)[3] was carried out to identify and understand significant type of influences from the different factors. Then, multivariate measurement errors [4] were early studied trying to understand the sources and structure of noise related to the experimental characteristics with raw and preprocessed data. Error covariance matrices (ECMs) were initially calculated using a series of replicates. Afterwards, matrices were decomposed considering a bilinear structure and the profiles obtained were translated to the error types and compared.

References

[1] B. Giussani, G. Gorla, J. Riu, Analytical Chemistry Strategies in the Use of Miniaturised NIR Instruments: An Overview, Crit. Rev. Anal. Chem. (2022) 1–33. https://doi.org/10.1080/10408347.2022.2047607.
[2] K.B. Beć, J. Grabska, C.W. Huck, Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers, Chem. - A Eur. J. 27 (2021) 1514–1532. https://doi.org/10.1002/chem.202002838.
[3] C. Bertinetto, J. Engel, J. Jansen, ANOVA simultaneous component analysis: A tutorial review, Anal. Chim. Acta X. 6 (2020) 100061. https://doi.org/10.1016/j.acax.2020.100061.
[4] M.N. Leger, L. Vega-Montoto, P.D. Wentzell, Methods for systematic investigation of measurement error covariance matrices, Chemom. Intell. Lab. Syst. 77 (2005) 181–205. https://doi.org/10.1016/j.chemolab.2004.09.017.

LOGICAL ANALYSIS OF THE SAMPLE POOLING RESULTS FOR QUALITATIVE ANALYTICAL TESTING: A PROOF-OF-CONCEPT STUDY

L. A. Sarabia[1], O. Valencia[1], M.C. Ortiz[2],
[1]*Departamento de Matemáticas y Computación, Universidad de Burgos, Burgos, España*
[2]*Departamento de Química, Universidad de Burgos, Burgos, España*
lsarabia@ubu.es

When the prevalence of positive samples in a whole population is low, the pooling of samples to detect them has been widely-used for epidemic control since the proposal of Dorfman [1], broadly reviewed by Cela [2]. The SARS-CoV-2 pandemic has spread so rapidly and reached such a large scale that evaluating different pooling strategies for SARS CoV-2 testing has been imperative [3], and some of them have already been approved by regulatory bodies.

With a qualitative (positive/negative) response of the analytical procedure, the supersaturated designs of experiments have been used to get the pooled samples [4,5].

To tackle the problem of identifying positives by sample pooling with qualitative analytical response, this work provides an original proposal that consists of two elements: i) the procedure to make the mixtures ii) the logical resolution, not numerical, to identify the positive samples from the results of the analysis of the pooled samples.

For i) the 'half' of a Plackett-Burman design that includes the mixture of all the individual samples is used. For ii) an algorithm has been built that, from the logical structure of the matrix of the design of the mixtures and the experimental response (0,1) for each of them, determines which individual samples cannot be positive. The complete solution to the problem is found by a hierarchical three-stage method: the first stage consists of pooling all samples, the second one applies the suggested algorithm and, if necessary, a third stage is required for individual identification.

As a proof of concept, the detection of positives by pooling 10 samples is considered. In this case, for a prevalence of positive equal to 0.05, the expected average number of trials is 2.552, which has to be compared to 5.987, the best expected value found in the literature for a hierarchical method.

The procedure of construction of the pooling samples and their analysis has been applied to the detection by polymerase chain reaction (PCR) of the pathogen *Listeria monocytogenes* and the allergen Pistachio. The *Listeria monocytogenes* is regulated by Commission Regulation (EC) No 2073/2005 of 15 November 2005 on microbiological criteria for foodstuffs, whereas the presence of Pistachio by Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on food information provided to consumers. In all cases, contaminated samples have been correctly detected.

Acknowledgments: This work is part of the project with reference BU052P20 financed by Junta de Castilla y León with the aid of European Regional Development Funds.

References

[1] R. Dorfman, *Ann. Math. Stat.* **14**(1943) 436–440.
[2] R. Cela, M. Claeys-Bruno, R. Phan-Tan Luu, Screening Strategies In: Brown S, Tauler R, Walczak R (eds.) Comprehensive Chemometrics, **1** (2009) pp. 251-300, Oxford, Elsevier.
[3] M. Crone, P. Randell, Z. Herm, et al. *Wellcome Open Research* (2021) 6:268.
[4] R. Cela, E. Martínez, A.M. Carro **52** (2000) *Chemom. Intell. Lab. Syst.* 167–182.
[5] R. Cela, E. Martínez, A.M. Carro **57** (2001) *Chemom. Intell. Lab. Syst.* 75-92.

ACETIC OR LACTIC BACTERIA CONTAMINATION? ASCA HAS THE ANSWER

D. Schorn-García[1], J. Ezenarro[1], M. Mestres[1], L. Aceña[1], O. Busto[1], B. Giussani[2], R. Boqué[3]
[1]*iSens Group. QAQO Dpt. Universitat Rovira i Virgili, Tarragona, Spain*
[2]*SmartChemoLab Group. DiSAT Dpt. Università degli Studi Dell'Insubria, Como, Italy*
[3]*GQQiN Group. QAQO Dpt. Universitat Rovira i Virgili, Tarragona, Spain*
daniel.schorn@urv.cat

Alcoholic fermentation is a biochemical process where the main reaction is the transformation of sugars into ethanol, releasing carbon dioxide. It is carried out by yeast, especially *Saccharomyces cerevisae* when it comes to grape must. Like any other bioprocess, it is very complex and there are many factors that influence its course, so to obtain high quality wines, a thorough monitoring of the process is essential [1]. In fact, poor working conditions can lead to sluggish or stuck fermentations and/or the generation of unwanted substances released by microorganisms such as lactic acid bacteria, acetic acid bacteria or even other yeasts.

To carry out analytical control in wineries, temperature, density and pH are daily measured and, in general, an organoleptic evaluation is usually performed. If more information is needed, supplementary analyses are performed in off-site laboratories, which requires more time. This implies the delay in obtaining the results and, therefore, the delay in the application of possible corrective measures. For these reasons, in recent years, different spectroscopic technologies have been studied and used to obtain real-time information when monitoring the alcoholic fermentation process [1,2].

In this work, a strategy is proposed to combine FTIR-ATR spectroscopy and chemometric techniques as a control tool following the PAT (Process Analytical Technologies) guidelines [3]. The proposed approach breaks down the sources of variability that affect spectra and is able to distinguish samples with different bacteria spoilages.

ANOVA-Simultaneous Component Analysis (ASCA) was applied to factorize the alcoholic fermentation variability sources, such as the process evolution and the contamination with lactic acid or acetic acid bacteria. Our previous results showed that there are different spectral pre-processing techniques that would affect ASCA results and that could be used to emphasize specific factors. These techniques were used to study the malolactic subprocess, carried out by unwanted lactic acid bacteria and acetification, carried out by unwanted acetic acid bacteria. Moreover, Simultaneous Component Analysis (SCA) was used to visualize the difference between lactic acid and acetic acid bacteria spoilage.

References

[1] D. Cozzolino, *Appl. Spectrosc. Rev*. **51(4)** (2015) 302-317.
[2] D. Schorn-García, J. Cavaglia, B. Giussani, O. Busto, L. Aceña, M. Mestres, R. Boqué. *Microchem J.* **166** (2021) 106215.
[3] FDA Off. Doc., Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacuring, and Quality Assurance (2004) 16.

Acknowledgements

Project PID2019-104269RR-C33 funded by MCIN/AEI/10.13039/501100011033.

A fluorometric photo-induced four-way calibration method for the determination of multiclass pesticides in citrus fruits

M. Antonio[1], M. R. Alcaraz[1], R. D. Falcone[2], M. J. Culzoni[1]
[1]*LADAQ-FBCB-CONICET, Universidad Nacional del Litoral, Santa Fe, Argentina*
[2]*IDAS-CONICET, Universidad Nacional de Río Cuarto, Río Cuarto, Argentina*
mculzoni@fbcb.unl.edu.ar

Pesticide usage has become an indispensable practice in agriculture production as it positively impacts crop yields and food quality worldwide by reducing diseases and controlling plagues. However, the side effects of their extensive use can be detrimental to the environment and human health, and cause long-term negative effects due to their high stability and bioaccumulation [1]. Hence, stringent regulations have been implemented by different agencies and governments by establishing the maximum residue levels of pesticides (MRLs), which are the residue levels not likely to be exceeded in a specific food or commodity [2,3] when pesticides are applied by their directions for use.

In this work, a four-way multivariate calibration method is presented for the simultaneous determination of 5 pesticides - thiabendazole (TBZ), carbendazim (CBZ), pirimiphos-methyl (PMM), imidacloprid (IMD), and clothianidin (CLT) - in citrus fruits. Third-order data were acquired by registering the photo-induced fluorescence of the analytes as excitation-emission fluorescence matrix (EEM) at different times of UV irradiation. The use of organized media (micelles) was implemented to enhance the fluorescent signal of the compounds.

First, in an attempt to obtain the optimal experimental conditions that yield the best performance of the method, a central composite design was implemented to evaluate the effect of the pH and the surfactant hexadecyltrimethylammonium chloride (HTAC) concentration on the intensity and the reaction velocity of each analyte. HTAC and phosphate buffer were chosen as reagents accordingly to preliminary studies. All the experiments were performed using the same concentration of each analyte, and the acquired third-order data were subjected to PARAFAC resolution. The optimal experimental conditions were finally set as pH 11.5 and 0.032 mol/L HTAC. The total UV-irradiation time was completed at 6 minutes.

The calibration sample sets were built in 5 concentration levels in triplicate for each analyte individually, except for TBZ and CBZ since it has been demonstrated that TBZ presents an inner-filter effect on CBZ [3]. TBZ/CBZ binary samples were prepared by following a random design. Besides, 16 validation samples containing the 5 analyzed compounds were prepared by following a random design at concentration levels different from those used for calibration. Lemon juice samples were pretreated by using a QUECHERS-based methodology.

Next, the quadrilinearity of the corresponding 4-way data arrays comprising calibration/validation samples was evaluated by applying 4-way PARAFAC. For CLT, non-quadrilinearity type 1 was observed due to a lack of reproducibility in the photo-induced reaction mode; thus, U-PLS/RTL and *A*PARAFAC were evaluated, being the former the one which retrieved the best results. In the case of IMD, the implemented chemometric models were not capable of providing satisfactory results shedding light on a possible lack of quadrilinearity due to an unexpected inner-filter phenomenon. For TBZ and CBZ, U-PLS/RTL was implemented as it has been proved that this model can cope with a non-quadrilinearity type 2 given by inner-filter effects.

Except for IMD, all models accomplished satisfactory results in the predictive analysis of the validation samples, with mean recoveries ( ) ranging between 98.6 and 102.9%. In all cases, the lack of significant differences between values and 100% was demonstrated through a hypothesis test. Moreover, the relative error prediction (REP%) values were below 9.1% demonstrating the good predictive performance of the developed method. These results suggest that second- and third-order advantages promote a high likelihood of success in the resolution of samples of high complexity such as citrus fruits.

References

[1] I. El-Nahhal, Y. El-Nahhal, *J. Environ. Manage.* ***299*** *(2021) 113611*.
[2] European Commission Pesticide MRLs-Regulation (EC) No. 396/2005.
[3] G. N. Piccirilli, G. M. Escandar, *Analyst.* ***131*** *(2006) 1012-1020*.

CHEMOMETRICALLY ASSISTED HIGH-THROUGHPUT METHOTREXATE SENSING STRATEGY BASED ON A pH-SWITCHABLE OPTICAL NANOSENSOR

M. Montemurro,[1] D. A. Uriarte,[2] H. C. Goicoechea,[1] S. E. Collins,[3] M. J. Culzoni[1]
[1] *LADAQ-FBCB-CONICET, Universidad Nacional del Litoral, Santa Fe, Argentina.*
[2] *INQUISUR-CONICET, Universidad Nacional del Sur, Bahía Blanca, Argentina*
[3] *INTEC-CONICET, Universidad Nacional del Litoral, Santa Fe, Argentina.*
mculzoni@fbcb.unl.edu.ar

Methotrexate (MTX) is an antineoplastic drug used in high doses for the treatment of different types of cancer. Given the need to carry out therapeutic monitoring in patients undergoing treatment with MTX to minimize the risk of toxicity [1], the development of analytical methods for its determination is of great importance. In this sense, nanotechnology, a multidisciplinary field that focuses on the study and application of materials at the nanoscale level, has emerged as a tool for the development of new analytical methodologies. During the last decade, there has been an accelerated increase in the study of ultra-small metallic nanoclusters (NCs) with luminescent properties, such as AuNCs and AgNCs.

In this work, an analytical method for the quantitation of MTX is presented. The methodology is based on a pH gradient coupled with UV absorbance detection for second-order data generation, and AgNCs as signal enhancement agents. First, AgNCs were synthesized by a chemical reduction method in an alkaline medium at room temperature, as previously reported in the literature [2]. The synthesized AgNCs were characterized by UV-Vis spectroscopy, DLS, and TEM. The pH gradient for MTX sensing was generated using an Agilent 1260 LC instrument, equipped with a diode array detector. The carrier solution consisted in 1000-fold diluted AgNCs in 0.01 mol L^{-1} sodium citrate (pH = 9.0), while the sample buffer was 0.01 mol L^{-1} sodium citrate (pH = 4.0). The carrier solution was pumped through an 800 mm length and 0.5 mm i.d. flexible mixing coil flowing at 0.1 mL min^{-1}, at 25 ºC. UV absorbance spectra were recorded in the range of 220-400 nm, every 2 nm, for 2.50 min. The data matrices registered for each sample were of size 375 × 91, for temporal and spectral modes, respectively. When the sample at acidic pH is injected, the generated pH gradient induces the dispersion of the AgNCs in the carrier solution and the consequent enhancement of the spectroscopic signal of MTX.

Calibration and validation samples, containing MTX, and test samples, containing MTX and three potential interferences (dexamethasone, prednisolone, and vincristine), were analyzed. Data modeling was performed by extended MCR-ALS and U-PLS/RBL algorithms. The predictive ability of the model in the absence and the presence of uncalibrated components (second-order advantage) was assessed. Besides, to evaluate the advantages of the proposed methodology, the same samples were analyzed without including AgNCs in the carrier solution. The validation samples were successfully modeled by extended MCR-ALS. However, the results obtained for the test samples were not satisfactory, as the effect of the interferences on the signal could not be modeled due to high collinearity in both the spectral and concentration modes. On the contrary, these latter samples could be modeled by U-PLS/RBL. The results obtained are presented in Table 1. The developed method showed a significant improvement in both the predictive ability and analytical figures of merit, compared to the AgNCs-free system.

Table 1. *Analytical figures of merit obtained by U-PLS/RBL modeling*

	AgNCs-free system	**AgNCs system**
REP (%)	11.0	7.0
CV (%) [a]	7.1	3.7
$LOD_{min-max}$ (μg L^{-1})	0.6 – 9.4	0.5 – 1.5
$LOQ_{min-max}$ (μg L^{-1})	1.9 – 28.2	1.6 – 4.6

[a] Calculated for five replicates of test samples corresponding to MTX central concentration level

References

[1] Karami, F., Ranjbar, S., Ghasemi, Y., Negahdaripour, M. (2019) Journal of Pharmaceutical Analysis, 9, 373-391.
[2] Ju, L., Lyu, A., Hao, H., Shen. W., Cui, H. (2019) Analytical Chemistry, 91, 9343-9347.

MULTIWAY DATA MODELING FOR ENHANCING CLASSIFICATION PERFORMANCE: FLUORESCENCE DATA AS CASE OF STUDY

S.M. Azcarate[1,2], J. Zaldarriaga Heredia[1,2], M.R. Alcaraz[2,3], J.M. Camiña[1,2], H.C. Goicoechea[2,3]

[1]*Instituto de Ciencias de la Tierra y Ambientales de La Pampa (INCITAP), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa, Santa Rosa, La Pampa, Argentina*
[2]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CABA, Argentina*
[3] *Laboratorio de Desarrollo Analítico y Quimiometría, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Santa Fe, Argentina*
hgoico@fbcb.unl.edu.ar

In the framework of multivariate classification, there is a continuous need for improving methods for identification and characterization. For quantitative purposes, the increase of the order of the data has proved certain benefits concerning the performance of the analytical method as the improvement of selectivity and sensitivity [1]. However, the advantages gained by increasing the order of the data to solve a classification problem have not been deeply studied yet [2].

This work aims to explore the data acquisition, feature extraction, and analysis methods of multi-way data arrays to improve the performance of the method to classify olive oils according to different purposes (variety, extraction process, and origin), as proof of concept. For the analysis, 21 olive oil samples, including virgin olive oils (VOO) and extra virgin olive oils (EVOO) of different commercial brands, were evaluated. Third-order data was obtained by acquiring an excitation-emission matrix (EEM) over the excitation range of 300-600 nm and the emission range of 400-700 nm for different periods of infrared heating (t_0=no IR heating, original sample). With the acquired data, different data arrays were built and subjected to several chemometric models to evaluate the properties and advantages of each data structure, as well as the performance of the modelling: (1) *First-order data analysis* using the emission spectrum registered at 348 nm excitation wavelength of each sample at t_0; (2a) *Second order data analysis* using the EEMs acquired for each sample at t_0; (2b) *Second order data analysis* using an emission-IR heating matrix (the emission spectra - λ_{exc} 348nm – acquired at different IR heating time) obtained for each sample; (2c) *Second order data analysis* using an excitation-IR heating matrix (the excitation spectra – λ_{em} 450 nm – acquired at different IR heating time) obtained for each sample; (3) *Third-order data analysis* of the emission-excitation-IR heating array obtained for each sample.

Two supervised pattern recognition methods were used to classify the studied oils, partial least squares discriminant analysis (PLS-DA) and its multi-way (NPLS-DA) and unfolding (UPLS-DA) extensions. Moreover, for cases (2) and (3), PARAFAC was implemented as a first decomposition model to extract features and scores that were then used for further classification analysis. Classification results were evaluated through global indices, such as average sensitivity, non-error rate, and average precision. The results revealed a high-class error rate when first-order data was used, higher than 30 %. Notwithstanding, different degrees of improvement were observed by the inclusion of an additional mode to the data structure.

The obtained results shed light on the fact that the use of higher-order data is an attractive approach to be explored in the classification field, particularly, in the study of samples with very similar spectral profiles, for which no evident classification patterns are observed. In addition, it is noteworthy to highlight that third-order data modelling profits from the chemical information of the system, in the direction of bettering the performance of the classification analysis.

References

[1] G.M. Escandar, H.C. Goicoechea, A. Muñoz de la Peña, A.C. Olivieri, *Anal. Chim. Acta* **806** (2014) 8-26.
[2] S.M. Azcarate, A. de Araújo Gomes, A. Muñoz de la Peña, H.C. Goicoechea, *Trends Anal. Chem.* **107** (2018) 151-168.

Data fusion approach applied in chemometrics-assisted metabolomics analysis

Martínez Bilesio AR[1]; Puig Castellví F[2]; Tauler R[3]; Rasia R[1]; Burdisso P[1];García-Reiriz AG[4]

[1] *Argentine Platform for Structural Biology and Metabolomics (PLABEM), Institute of Molecular and Cell Biology of Rosario (IBR), National Scientific and Technical Research Council (CONICET) - Rosario, Santa Fe, Argentina*

[2] *Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, Paris, 75005, France*

[3] *Institute of Environmental Assessment and Water Research (IDAEA), Spanish National Research Council (CSIC) - Barcelona, Catalunya, Spain*

[4] *Institute of Chemistry of Rosario (IQUIR), National Scientific and Technical Research Council (CONICET) - Rosario, Santa Fe, Argentina*

martinezbilesio@ibr-conicet.gov.ar

Metabolomics is characterized by the generation of a large amount of information, making data analysis methods essential to obtain relevant knowledge. This is especially evident in the case of untargeted studies, in which the molecular mechanisms of response for a given disease or environmental stress are intended to be established [1].

In this work, different chemometric strategies for the analysis of metabolomics data recorded by nuclear magnetic resonance (NMR) are presented. In order to optimize this analysis, the information from 145 healthy volunteers was processed, including: i) the ^{1}H NMR spectra (metabolic profiles) of serum and urine samples, and ii) the clinical metadata consisting on qualitative parameters and quantitative biochemical determinations.

In the first step, different preprocessing methods were sequentially applied over the ^{1}H NMR spectra, like resonance alignment and normalization [2, 3]. Afterwards, resonance integration by means of a multivariate curve resolution (MCR)-based strategy was performed [4]. In this way, a set of peak integrals for representative metabolites of serum and urine samples was obtained.

In the second step, the metabolites peak integrals profiles and the clinical metadata values of all the volunteers were analyzed together. Two different data fusion strategies were developed. In the first strategy, the direct fusion of the resulting variables from clinical metadata, serum samples and urine samples was performed, analyzing the obtained data matrix through multivariate curve resolution - alternating least square (MCR-ALS). In the second strategy, the data matrices from the clinical metadata, the serum samples and the urine samples were analyzed individually by MCR-ALS, as a data compression stage. Finally, the resulting scores were merged, analyzing them again by MCR-ALS (tandem MCR-based approach). By applying these two strategies, it was possible to elucidate the general metabolic patterns throughout the volunteers.

These results allowed evaluating different analytical strategies based on chemometrics methods, in order to investigate NMR metabolomics datasets and extract relevant biochemical information. It was possible to apply and optimize the processing to obtain metabolites relative concentration matrices in biofluids, from NMR spectra and through the use of MCR-ALS. Two different data fusion strategies for the global analysis of different types of clinical samples were developed. The second one greatly reduces the dimensionality of the data, facilitating its interpretation. Thus, we could verify the potentiality of the data fusion approach to analyze together the information from different biological matrices.

References

[1] Nicholson JK, Lindon JC, Holmes E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29: 1181-1189.

[2] Vu TN, Laukens K. 2013. Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data. *Metabolites* 3: 259-276.

[3] Kohl SM, Klein MS, Hochrein J, Oefner PJ, Spang R, et al. 2012. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* 8: 146-160

[4] Pérez Y, et al. MCR-ALS analysis of 1H NMR spectra by segments to study the zebrafish exposure to acrylamide. *Analytical and Bioanalytical Chemistry*. 2020;412: 5695-5706.

MAPPING CHEMOMETRICS WITH CHEMOMETRICS

N. Cavallini[1], M. Mancini[2], L. Strani[3], A. Tugnolo[4], E. Alladio[5], N. Iaccarino[6], F. Savorani[1]
[1]*Department of Applied Science and Technology (DISAT), Polytechnic of Turin, Turin, Italy*
[2]*Department of Agricultural, Food and Environmental Sciences, Università Politecnica delle Marche, Ancona, Italy*
[3]*Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Modena, Italy*
[4]*Department of Agricultural and Environmental Sciences (DiSAA), University of Milan, Milan, Italy*
[5]*Department of Chemistry, University of Turin, Turin, Italy*
[6]*Department of Pharmacy, University of Naples Federico II, Naples, Italy*
nicola.cavallini@polito.it

During their daily practice of the topic, chemometricians often face the question "what is Chemometrics?", asked by students, laboratory analysts, chemists, and industry customers. Much effort is spent trying to come up with clear and concise answers, also considering the different levels of detail at which the question needs to be answered. Due to the strong link between Chemometrics and experimental real-life problems, the application of statistical and machine learning methods is done from a practical perspective, which often goes well beyond the simple "trust" that a certain method will work. The "one-fits-for-all" approach is however very common in the wide field of data science, which also is, like all other fields of human knowledge and activities, subject to fashion trends. This often leads to posts on social networks (LinkedIn above all) of extremely simplified data analysis workflows, mainly done with the purpose of getting clicks and views, while providing a partial point of view, strongly influenced by the trendiest methods.
Therefore, together with the question "what is Chemometrics", we should probably ask ourselves: "*where is Chemometrics?*". Starting from the considerations proposed by José Amigo in 2021 [1], we believe that a discussion about the relationship between Chemometrics and the wide field of data science (does the latter includes the former, or vice versa?) should be started in our community. This leads to a second important layer of the subject, which is about how to "organize" and compare the methods and topics of Chemometrics, according to their purpose(s) and uses, but also their characteristics. *This would mean describing Chemometrics with the tools of Chemometrics.*

The discussion we would like to stimulate aims at getting a better understanding of our beloved field of research, starting from collecting the opinions of the participants to the CAC 2022 Conference. We propose to use a data collection method from the food sensory field, the *Napping* technique proposed in 2005 by Jérŏme Pagès [2]. This method allows non-expert panellists to evaluate products based on their own personal criteria: the products get placed on a tablecloth according to their perceived similarity (close) or dissimilarity (distant). The result is a perceptual map which can be digitalized and converted into a set of Euclidean distances. All data collected from the participants can be organized as a matrix which can be analyzed by Procrustes Multiple Factor Analysis [3], which provides, among the other outputs, a "consensus configuration", which looks like a map with all evaluated objects on it.

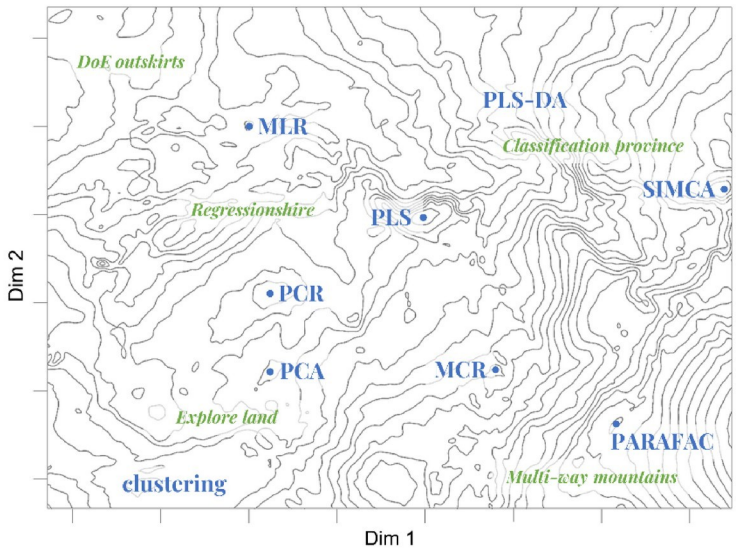


Figure 1 – A "Pirates of Chemometrics" consensus configuration.

References

[1] J.M. Amigo, *BrJAC*, **8**(32), (2021) 22-38
[2] J. Pagès, *FoodQual*, **16**(7), (2005) 642-649
[3] E. Morand, *FoodQual*, **17**(1-2), (2006) 36-42

JOINT FACTORIZATION OF RIGHT-ANGLE AND FRONT-FACE FLUORESCENCE DATA TO IMPROVE PARAFAC PURE PROFILES RECOVERED FROM OIL-IN-WATER EMULSIONS

I. M. A. Viegas*[1], Simon I. Andersen[1], Åsmund Rinnan[2]
[1] *Danish Offshore Technology Center, Technical University of Denmark, Kgs. Lyngby, Denmark*
[2] *Department of Food Science, University of Copenhagen, Frederiksberg, Denmark*
iviegas@dtu.dk

Opaque and/or high-optical density samples are prone to inner-filter effects that interfere with the fluorescence intensity and spectral profiles. Different geometries of sample illumination lead to different path lengths: the longer the optical path, the more molecules along the light beam and emission collection sections, and the more reabsorption effects. The front-face (FF) geometry provides a shorter path length than the traditional right angle (RA) and, therefore, can be helpful to minimize inner-filtering. However, in emulsion samples, such as skin lotions, homogenized milk, pharmaceutical formulations, etc., the concentrations of fluorophores in the "outer layer" hit by light in FF and throughout the optical path in RA are expected to be different. Hence the emission spectra obtained by RA and FF provide complementary information rather than just quenched and/or shifted bands. We propose then the joint factorization of RA and FF excitation-emission matrices (EEM) of crude oil emulsions to verify the presence of additional components that could not be found by each cuvette geometry alone. EEMs of oil-in-water emulsions were recorded with excitation from 300 to 350 nm and emission from 370 to 580 nm, using both geometries, which resulted in two three-mode tensors with dimensions 22 × 2101 × 6 (samples × emission × excitation) each. Since the spectral range measured did not comprise the Rayleigh scattering lines, and Raman scattering did not appear to significantly affect the data, no preprocessing was needed. Each tensor was factorized separately by PARAFAC, and only two components could be extracted with no convergence or two-factor degeneracy issues. On the other hand, the joint factorization of both RA and FF tensors by Advanced Coupled Matrix and Tensor Factorization (ACMTF) enabled the extraction of up to four components depicted in Figure 1. FF data contributed to the four components, while only components 1 and 2 could be extracted from RA data. FF EEM captures emissions mainly from dissolved species, and RA, from both dissolved and dispersed species, so one could expect to recover more factors from RA data than FF. However, RA is more affected by inner filtering, which almost completely masked the emission from the two other variation sources assigned to components 3 and 4. We can therefore conclude that the coupled factorization of RA and FF EEMs has the potential to improve spectral information recovered from oil-in-water emulsions and show underlying features beyond the individual factorization, and therefore it should be thought of as a strategy also for other types of emulsions.

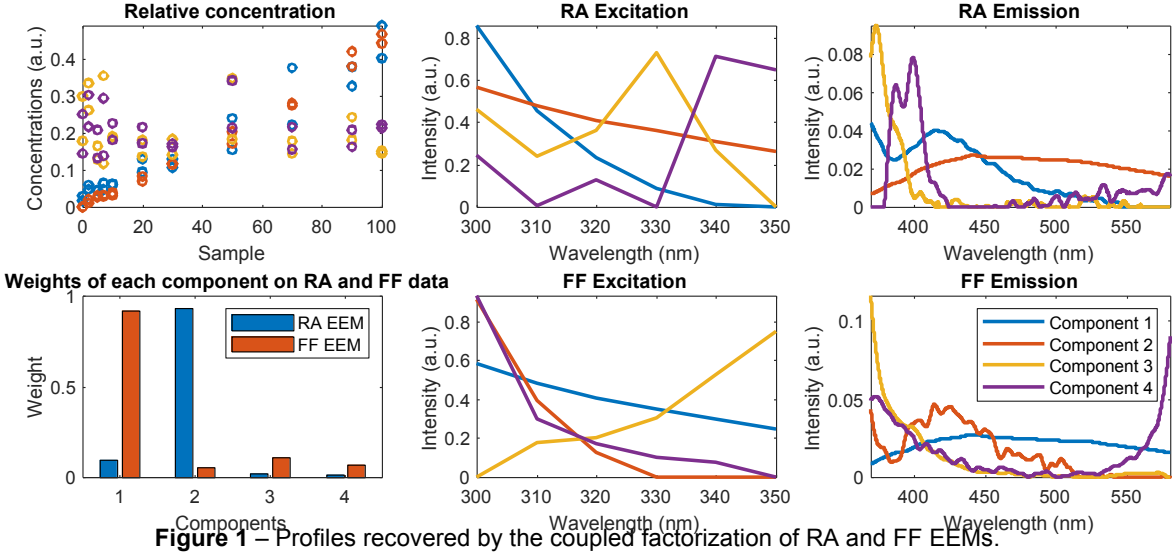


Figure 1 – Profiles recovered by the coupled factorization of RA and FF EEMs.

Reference

[1] E. Acar, E. E. Papalexakis, G. Gürdeniz, M. A. Rasmussen, A. J. Lawaetz, M. Nilsson, R. Bro, BMC Bioinformatics **15** (2014) 239.

MULTIVARIATE CURVE RESOLUTION OF INCOMPLETE AND PARTLY MULTILINEAR MULTI-BLOCK DATA SETS

M. Marín-García[1], R. Tauler[1]
[1] *Department of Environmental Chemistry, IDAEA-CSIC, Barcelona, Spain*
marcmaring@gmail.com

In this work, the analysis of multi-block data sets with the Multivariate Curve Resolution chemometric method is extended to the cases where some of the data blocks are missing (incomplete) and, additionally, to the cases where the application of multilinearity constraints can be applied only partly to some of these blocks [1–3]. Moreover, different ways of releasing the fulfillment of the multilinearity constraint are tested.

The proposed approach is shown in the analysis of three types of multi-block datasets. In the first one, the multi-block data set consisted of the simultaneous analysis of different chemical experiments, all of them monitored with UV spectrometric detection, which includes acid-base equilibria titration experiments, photodegradation kinetic experiments, and liquid chromatography runs. In the latter case of liquid chromatographic experiments, the chromatographic analysis of the same samples was additionally performed using mass spectrometry, allowing the 'fusion' of MS and UV data. The whole data could then be arranged as an incomplete multi-block data set and processed with the proposed extension of the MCR-ALS method, where the trilinearity constraint is applied only for some components in some of the data blocks simultaneously analyzed [4].

In the second case, the multi-block data set consisted of the simultaneous analysis of UV spectrometric and spectrofluorimetric experiments. In this case, both photodegradation kinetic experiments and the further chromatographic analysis of the samples were performed jointly using UV and fluorescence detection. The whole data set was also arranged in an incomplete multi-block data set where some of the data blocks were missing as a result of the setup of the different data sets for their simultaneous analysis. In this case, some parts of the data blocks could be considered multiway due to the fluorescence data behavior. In contrast, only the bilinear model could be applied to other data blocks [5].

Finally, in the third case, the multi-block data set consisted of the simultaneous analysis of two different hyphenated liquid chromatographic experiments, LC-DAD-MS and LC-DAD-FLD. In this case, several samples from the same photodegradation experiment were analyzed using both hyphenated techniques, which were also arranged as an incomplete multi-block data set. In summary, the obtained results show the advantages of using the proposed incomplete multi-block mixed multilinear model approach.

References

[1] M. Alier, R. Tauler, *Chemom. Intell. Lab. Syst.* **127** (2013) 17–28.
[2] M. De Luca, G. Ioele, R. Tauler, G. Ragno, *Anal. Chim. Acta* **837** (2014) 31–37.
[3] A. de Juan, R. Tauler, *Anal. Chim. Acta* **1145** (2021) 59–78.
[4] R. Tauler, *J. Chemom.* **35**(2) (2021) e3279.
[5] X. Zhang, R. Tauler, *Molecules* **27**(7) (2022) 2338.

LOW SIGNAL INTENSITY, MEASUREMENT ERRORS AND BIOLOGICAL SIGNIFICANCE: A MODEL FOR LC-MS PROTEOMICS

S. Keretsu[1,3] and T.K. Karakach[1,2]

[1]*Laboratory of Integrative Multi-Omics Research, Department of Pharmacology, Dalhousie University, Halifax, NS, B3H 4R2, Canada*

[2]*Beatrice Hunter Cancer Research Institute, Halifax, NS, B3H 4R2, Canada*

[3]*Department of Pathology, Dalhousie University, Halifax, NS, B3H 4R2, CANADA*

karakach@dal.ca

Proteomics by LC-MS yields data that are complex and advanced computational tools are required to deconvolve this complexity. Alternatively, test statistics under a generalized least squares framework, are used to model the relationship between the instrumental response (y_i) and the biological question (x_m) under the null hypothesis that the coefficients, $\beta_j's$, are equal to zero *i.e.*, $y_i = \beta_j x_m + \epsilon_i$. Fundamentally, it is assumed that the data are normally distributed while the errors ($\epsilon_i's$) are uniformly distributed. These assumptions are not always true, however, and several approached to transform the data to meet these assumptions have been devised.

We comprehensively characterize measurement uncertainties associated with LS-MS proteomics using an experiment designed to capture contributions of several sources. We show that for a well-designed experiment, the total variance associated with biological are hierarchical such that $\sigma^2_{total} = \sigma^2_{bio} + \sigma^2_{err}$, where $\sigma^2_{bio} \gg \sigma^2_{err}$. We also show the multivariate structure of the noise depicting correlations between variables and devise a method to model this noise allowing errors in LC-MS measurements to be accurately estimated without the need for extensive replication. This estimate is then used to pre-process proteomics data to allow low intensity signals to be modelled with the same relative importance.

QUANTITATIVE EVALUATION OF RED MEATS IN KEBAB LOGHMEH SAMPLES: FOURIER TRANSFORM INFRARED DATA AND CHEMOMETRIC METHODS

Reza Nafari[1], Zahra Nabi[1]
[1]*Islamic Azad University, Shiraz, Iran*
reza_nafari_2010@yahoo.com

Kebab loghmeh is one of the most popular meat products in Iran. Quantitative assessment of the red meats is critical as the most important factor in authentication of this meat product. Other methods of adulteration tracing do not include enough efficiency to quantitatively assess quantities of the red meats in final products. Therefore, the objective of the current study was to quantitatively assess red meats in Kebab loghmeh samples using Fourier transform infrared (FTIR) method and chemometric methods. Samples of industrial Kebab loghmeh containing 70 and 90% of red meats from three various brands were purchased from the local markets and standard formula samples were prepared in meat product factories (total sample number of 36). All samples were transferred to the laboratory under cold conditions. Data from FTIR were analyzed using PCA, PLS-DA and SIMCA methods as chemometric methods. Results of multiple linear regression and chemometric methods with high determination coefficient (R2 = 0. 9999) showed that 67% of the samples did not included information provided on the labels. Analysis of FTIR data using chemometric methods is appropriate for the quantification of red meats in kebab loghmeh samples.

References

[1] H. Hosseini, Kh. Barazandegan Kh, A. Akhondzadeh, B. Shemshadi, HR. Tavakoli, R. Khaksar. *J Food Sci Tech.* **6**, (2009) 95-100.
[2] MP. Callao, I. Ruisánchez. *Food Control.* **86** (2018) 283-293.

Fast and Convenient Authenticity Control of Natural Products using Mass Spectrometry and Chemometrics

Justine Raeber[1], Christian Steuer[1]
[1]*Departement for Chemistry and Applied Biosciences, ETH, Zurich, Switzerland*
justine.raeber@pharma.ethz.ch

Essential oils (EOs) are natural products, which are widely used as flavors and fragrances and as phytomedicine.[1] EOs can consist of only a few or up to more than 100 individual compounds in varying concentrations.[2] Due to production costs and high trading prices, EOs are subject to adulteration. Current quality control of EOs focuses on the analysis of a few selected markers. However, this approach is not sufficient. Due to the broad composition of EOs analytical analysis is challenging.[1] An extended method for the quality control of EOs is based on compound pattern analysis of e.g. terpenes, which has been applied successfully for the geographic allocation of pine oil.[3] A holistic analysis of natural products is essential to guarantee patient safety and correct product labelling in phytomedicine. We therefore present an alternative approach for a fast and convenient analysis of natural products using dielectric barrier discharge ionization-mass spectrometry (DBDI-MS) in combination with chemometrics. DBDI is a soft ionization technique that takes places at atmospheric pressure.[4] Samples can be placed directly in front of the source without pretreatment. Volatile components are ionized and introduced into the MS surface. In this approach samples were analyzed using a triple quadrupole MS (Triple Quad 3500, Sciex) using only the Q_3 scan from a range 50 – 400 Da. This methodology produces characteristic fragment patterns, which can be used for chemometric analysis such as hierarchical cluster analysis (HCA) as presented in Fig. 1.

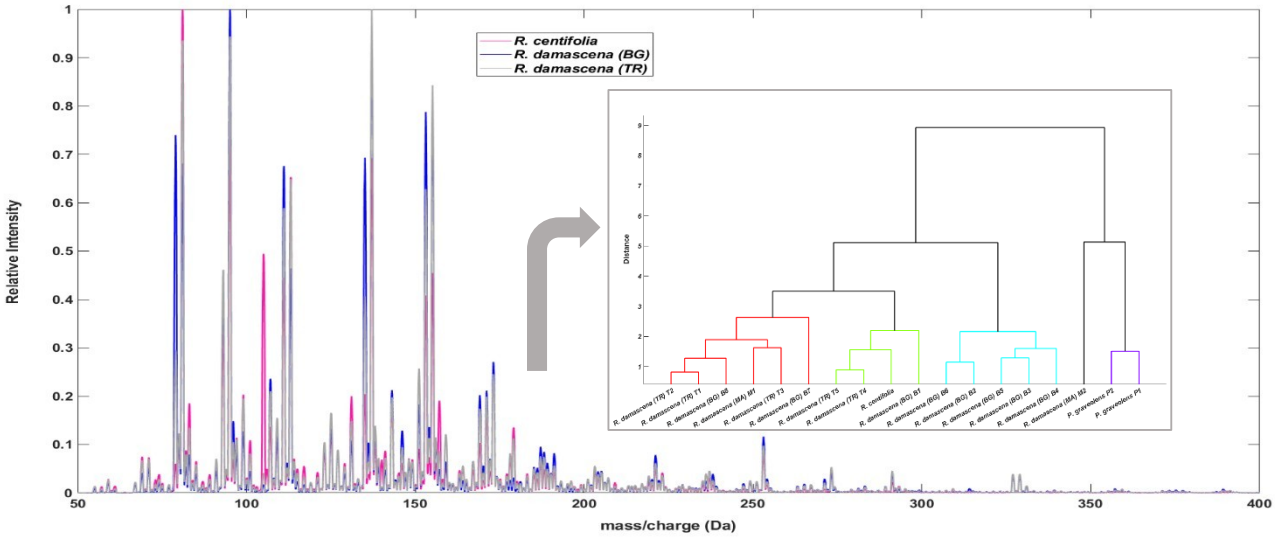


Figure 1 – Fragmentation pattern after Q_3 scan of different rose oil samples and subsequent HCA analysis. Data was processed by conducting background subtraction and normalized by the highest intensity.

In a preliminary experiment Q_3 scan data from rose oil samples was used for HCA. HCA is an unsupervised classification method and calculates the proximity of values and separates these based on their (dis)similarity.[5] Analysis resulted in separation of the origin of rose oil and might be able to reclassify mislabeled samples. Further experiments suggest that the approach can also distinguish natural products based on their variety and plant organ. DBDI-MS in combination with chemometrics seems to be a promising quality control approach with fast and convenient applicability.

References

1. Do TKT, Hadji-Minaglou F, Antoniotti S, Fernandez X, *TrAC*, **66** (2015), 146-157.
2. Sharmeen JB, Mahomoodally FM, Zengin G, Maggi F, *Molecules,* **26** (2021), 666.
3. Allenspach M, Valder C, Flamm D, Grisoni F, Steuer C: *Molecules,* **25** (2020), 2973.
4. Gyr L, Klute FD, Franzke J, Zenobi R: *Anal. Chem,* **91** (2019), 6865-6871.
5. Köhn H-F, Hubert LJ, *Wiley StatsRef: Statistics Reference Online.* edn., (2015), 1-13.

AQUAPHOTOMICS STUDY OF BODY FLUIDS IN CANCER RESEARCH

A. Surkova [1,2]*, E. Boichenko [1], A. Bogomolov [2], V. Artyushenko [3], T. J. Munćan [4], R. Tsenkova [4]
[1] *Institute of Chemistry, Saint Petersburg State University, Saint Petersburg, Russia*
[2] *Department of Analytical and Physical Chemistry, Samara State Technical University, Samara, Russia*
[3] *art photonics GmbH, Berlin, Germany;*
[4] *Kobe University, Kobe, Japan*
melenteva-anastasija@rambler.ru

A modern era of cancer research is characterized by a permanent development of advanced diagnostic and monitoring methods, which enable the transition to a personalized care. Current methods in clinical oncology are mostly invasive and require sophisticated equipment and specifically trained personnel. The extraction of diagnostically-relevant information from body fluids (blood, urine) is a perspective way of low-invasive, fast and cost-effective cancer diagnostic and treatment monitoring.

Near infrared (NIR) spectroscopy is a powerful method for rapid and non-invasive qualitative and quantitative analysis, applicable to a wide variety of samples. NIR spectra of body fluids, such as blood and urine, allow revealing changes in molecular composition of samples containing intact and malignant cells due to the differences in their metabolism. Therefore, NIR spectroscopy has a great potential for noninvasive or minimally invasive and inexpensive cancer diagnostic and treatment monitoring [1]. However, a holistic strategy of NIR spectral data analysis has not been developed yet that hinders a wide application of NIR spectroscopy in clinical practice.

An NIR spectrum contains rich information about the molecular interactions of water. Hydrogen bonds formed between water molecules and their environment reflect molecular changes in biological systems [2]. Water absorption bands in the NIR region are potentially useful for cancer detection, but a specific data processing is needed to extract the relevant information. The aim of this study is to investigate water molecular structure using the NIR spectra of urine, blood plasma and serum by aquaphotomics combined with chemometric methods of data analysis. The samples were collected from patients with diagnosed cancer in different locations before and after cancer surgery. The raw and pre-processed spectra, as well as the loading vectors of the PCA analysis were studied to find and assign the characteristic water absorbance bands that change significantly in response to the surgical treatment. The reported results can be potentially used for the characterization of biological materials and monitoring of cancer treatment.

The reported study was funded by the grant #MK-2192.2021.4.

References

[1] P. Raja, P. Aruna, D. Koteeswaran , S. Ganesan. *Vib. Spectrosc.* **102** (2019) 1–7.
[2] R. Tsenkova, J. Munćan, B. Pollner, Z. Kovacs. *Front. Chem.* **6** (2018) 363.

OPENING THE RANDOM FOREST BLACK BOX OF THE ASPARAGUS METABOLOME

S. Wenck[1], M. Creydt[1], F. Gaerber[1], J. Hansen[1], M. Fischer[1], S. Seifert[1]
[1]*University of Hamburg, Hamburg, Germany*
soeren.wenck@uni-hamburg.de

Misdeclaration of food products regarding their geographical origin and biological identity is a major aspect of food fraud, especially for more expensive foods such as asparagus. Therefore, analytical methods are sought to ensure the determination of these essential characteristics. It has already been shown that Random Forest (RF) analyses of the metabolome of asparagus, obtained by LC-MS studies, are suitable for origin determination [1]. However, in this study, RFs were applied as black box methods, meaning that no further characterization of the metabolome was performed. The goal of the research presented here is to extend RF analysis to both taxonomic identity classification and metabolome characterization of asparagus [2]. For this purpose, we apply different RF based approaches, e.g. the variable selection methods Boruta [3] and Surrogate Minimal Depth (SMD) [4]. SMD considers the mutual impact of the variables on the outcome and is also utilized to identify metabolites that provide similar information for classification. We show that this approach can identify features attributed to fragments and adducts of the same metabolite and features from different metabolites co-occurring in the respective groups. The latter can be used for comprehensive characterization, e.g. for metabolic pathway analyzes.

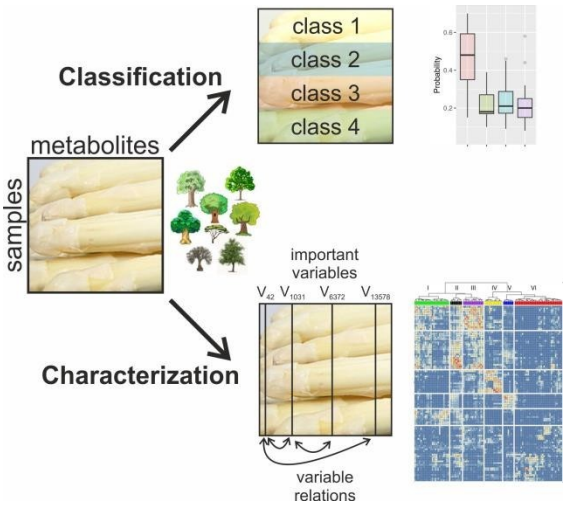


Figure 1 – Graphical summary of the classification and characterization of the LC-MS metabolome of white asparagus with random forest methods (from [2])

References

[1] M. Creydt, D. Hudzik, M. Rurik, M. Fischer, *J. Agric Food Chem.* **66** (2018) 13328-13339.
[2] S. Wenck, M. Creydt, J. Hansen, F. Gaerber, M. Fischer, S. Seifert, *Metabolites.* **12** (2022) 5.
[3] M. B. Kursa, W.R. Rudnicki, *J. Stat. Softw.* **36** (2010) 1-13.
[4] S. Seifert, S. Gundlach, S. Szymczak, *Bioinformatics.* **35** (2019) 3663-3671.

Relationship between cadmium availability and soil properties in cacao farms at Santander - Colombia

C.A. Adarme-Duran[1], P.F.B. Brandão[2], E. Castillo[2]
[1]*Instituto de Biotecnología – Universidad Nacional de Colombia – sede Bogotá, Bogotá, Colombia*
[2]*Departamento de Química –Universidad Nacional de Colombia – sede Bogotá, Bogotá, Colombia*
cadarme@unal.edu.co

Regulation on the content of cadmium (Cd) in chocolate has affected cocoa farmers in different parts of the world, particularly some regions of Latin America where it exceeds the proposed limits [1]. Considering that Cd moves from the soil to the plant, it is important to study its availability and association with other soil factors (including bacterial activities that can modify the metal fractionation on soil through metabolic processes) because this could help to explain or predict Cd content on the cocoa beans, the raw material for chocolate production [2]. Further, usually soil samples are taken without considering that the soil can be divided in two fractions (rhizospheric and non-rhizospheric) that differ in the properties mentioned above due to plant root exudates. In this work, we investigated the correlations between Cd and other physicochemical properties like pH, P, CO, cationic bases, and urease activity, looking for differences among the two soil types, using chemometrics in R software: Spearman correlations and cluster analysis. 102 soil samples (including rhizospheric and non-rhizospheric soils) from two cocoa farms in Santander - Colombia, were evaluated. Rhizospheric soil showed that the correlation between available Cd with the other variables was (from positive to negative): Cd_pseudototal > urease activity > CO > P > Ca > CICE > Mg > K > pH > 0 > Al > Na; while for non-rhizospheric soils it was: Cd_pseudototal > P > CO > pH > Mg > urease activity > Na > K > 0 > Ca > Al > CICE (Fig. 1). According to these results, there is a different correlation between Cd and the other physicochemical properties considering the type of soil, and it is interesting how the urease activity has the second greater Spearman correlation coefficient in the rhizospheric soil but has a lower correlation in the case of non-rhizospheric soil. From the cluster analysis it was observed that the available Cd showed statistical similarity to Cd pseudototal and P, for both non-rhizospheric and rhizospheric soils. This initial approach would be complemented with multivariate analysis. The results presented here are important to understand soil cadmium availability in cocoa farms and contribute to the development of mitigation strategies to reduce Cd content in *Theobroma cacao* L.

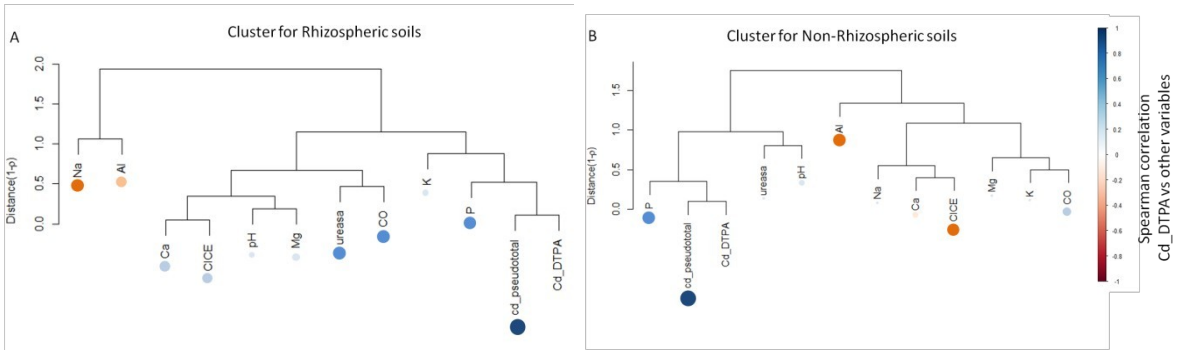


Figure 1 – Cluster and Spearman correlation of Cd and other physicochemical properties for A) Rhizospheric soils and B) Non-Rhizospheric soils from cocoa farms in Santander, Colombia. Sodium (Na), Al (Exchangeable acidity), calcium (Ca), effective cation exchange capacity (CICE), magnesium (Mg), urease activity, pH, organic carbon (CO), potassium (K), phosphorous (P), Cd pseudototal and available cadmium (Cd_DTPA). Circle size and colors varies according to the Spearman correlation coefficient.

Keywords: Cd, soil pollutant, hierarchical clustering, Spearman correlation, *Theobroma cacao* L.

References

[1] D. Argüello, E. Chavez, F. Lauryssen, R. Vanderschueren, E. Smolders, D. Montalvo, *Sci. Total Environ*. **649** (2019) 120-127.
[2] R. Vanderschueren, D. Argüello, H. Blommaert, D. Montalvo, F. Barraza, L. Maurice, E. Schreck, R. Schulin, C. Lewis, J. L. Vazquez, P. Umaharan, E. Chavez, G. Sarret, E. Smolders, *Sci. Total Environ*. **781** (2021) 146779.

Application of multivariate data analysis coupled with spectroscopy to agroalimenaire investigation in Morocco: advancement and challenge

A. Ait sidi mou[1,2], M. Daoudi[3], M. E. A. Ghanjaoui[4], A. El Mchaouri[2]

[1]*Laboratory of Materials Engineering for the Environment & Natural Resources, FST Errachidia, University Moulay Ismail of Meknes, BP 509 Boutalamine, 52000, Errachidia, Morocco*

[2]*Laboratory of Physical Chemistry & Biotechnology of Biomolecules and Materials, Faculty of Sciences and Techniques of Mohammedia (FSTM), Mohammedia, Morocco*

[3]*Laboratory of Solar Energy and Environment, Faculty of Sciences, Mohammed V University, Rabat, B.P. 1014, Rabat, Morocco*

[4]*Laboratory of Water and Environment, Department of Chemistry, Faculty of Sciences, University Chouaïb Doukkali, PO. Box 20, El Jadida 24000, El Jadida, Morocco*

aitsidimou_aziz@yahoo.fr

Abstract

The conservation of agricultural-food, play an important factor in helping people live normally and in enhancing their quality of life [1,2]. Over the past decades, some investigation has been given attention to the adulteration of agricultural-food in Morocco [3]. A many researcher uses the multivariate data analysis coupled with spectroscopy as a tool which contribute to the natural environment protection [4,5]. The multivariate data analysis is an efficient, low-cost and economical tool, which can lead to evaluate the quality and relationship of various variables of the sample. However, a little work published between 2012 and 2022 was carried out in Morocco on the application of chemometric methods toward agricultural-food studies. Hence, the aim of this paper is to clarify to the scientist the advantage, opportunity and challenge can offer the application of multivariate data analysis for enhancing the natural environment protection in Morocco.

Keywords: natural environment protection, agricultural-food, multivariate data analysis, spectroscopy, Morocco.

References

[1] B. L. Allen, *J. Food Webs* **2** (2015) 1–9.
[2] N. Kurashima , L. Fortini , T. Ticktin, *Nat. Sustain.* **2** (2019) 191–199.
[3] I. T. Abdelhedi, S. Z. Zouari, *Knowl Econ.* **11** (2020) 193–210.
[4] H. Zaroual, C. Chèné , E. M. El Hadrami, R. Kaloti, *Food Chem.* **370** (2022) 131009.
[5] A. Bajoub, E. A. Ajal, A. F. Gutiérrez, A. C. Pancorbo, *Food Res. Int.* **84** (2016) 41-51.

CHEMOMETRICS – A CHEMOMETRIC PYTHON PACKAGE

M. Rüdt[1]
[1]*HES-SO Valais-Wallis, Sion, Switzerland*
matthias.rudt@hes-so.ch

In recent years, the field of machine learning has seen a rapid development of new technologies. Some of the most important software libraries such as scikit-learn and TensorFlow use Python as their primary programming interface. While chemometrics may be considered a subset of machine learning, up to date, no comprehensive free and open-source software library exists in Python for chemometric data analysis.

The Python library *chemometrics* tries to close this gap. Currently, the focus of *chemometrics* lies on the chemometric analysis of spectroscopic data (e.g., UV/Vis, NIR, Raman, NMR and MS). To ensure free and open-source usage, *chemometrics* is released under GPL-3.0 license. *chemometrics* builds on the concepts of scikit-learn and extends scikit-learn's functionalities to support chemometric data analysis. The package provides methods for plotting, preprocessing (extended multiplicative scatter correction [EMSC], Whittaker smoother, asymmetric Whittaker), fitting spectroscopic data (PCA, PLS, MCR, indirect hard modeling [IHM]) and a range of routines for analyzing the data and model quality. The code is available on Github [1] and distributed over the Python Package Index (PyPI). It is thoroughly tested by automated unit tests with a code coverage >97%. The documentation is available online [2], covers the complete public code and provides several examples. Due to the close integration with scikit-learn, *chemometrics'* functionality may be easily included in scikit-learn workflows.

While *chemometrics* has been actively developed since April 2020, significant gaps remain. Future steps will focus on extending the implemented functionality as well as on broadening the user basis.

References

[1] https://github.com/maruedt/chemometrics
[2] https://chemometrics.readthedocs.io/

Structuring and generalizing implementations of N-FINDR algorithm for unmixing hyperspectral data

R. Guliev[1], U. Neugebauer[1,2,3], C. Beleites[2]

[1] *Leibniz Institute of Photonic Technology, Jena, Germany*

[2] *Institute of Physical Chemistry and Abbe Center of Photonics, Jena, Germany*

[3] *Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Jena, Germany*

[4] *Chemometrix GmbH, 61200 Wölfersheim, Germany*

rustam.guliev @leibniz-ipht.de

Visualization of hyperspectral images is not a straight-forward process since each pixel contains a vector of spectral values rather than single value. One of the commonly used approaches is determination of some pure component spectra, assigning each of them a different color, and representing each pixel according to the contribution of the pure components. The crucial part of that process is determination of those pure components (generally called unmixing).
N-FINDR is one of the most commonly used for that. Despite its popularity, the tools and algorithms for application are not well established yet. Since the first publication in 1999 [1], multiple works suggested different implementations and optimizations of the algorithm under many different abbreviations: SM N-FINDR, SQ N-FINDR, SC N-FINDR, IN-FINDR, LDU-N-FINDR, LDU-S-N-FINDR, S-N-FINDR, MN-FINDR, MN-FINDR2, and so on. With this and the lack of publicly available implementations, the definition of the algorithm became somewhat blurred. Although there were attempts to classify implementations of the N-FINDR algorithm [2], it happened that the same abbreviation is used for different implementations of the algorithm [3].

In this work, we generalized the N-FINDR replacing the use of abbreviations by two parameters defining the iteration direction and the volume change estimation. Structuring the algorithm this way covered most of the previous implementations and besides that it also allowed us to introduce some previously unknown implementations. The generalized version is implemented in R and publicly available on GitHub (r-hyperspec/unmixR). In addition, each modification of the algorithm was vectorized by replacing loop steps by matrix operations. Finally, benchmarking of each algorithm implementation was performed and recommendation for default algorithm usage is given.

We gratefully acknowledge the contributions to the R package by Conor McManus (2013) and Anton Belov (2016) as well as their funding by Google Summer of Code, and Bryan Hanson's contributions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 861122 (ITN IMAGE-IN).

References

[1] M. Winter, Proc. SPIE 3753, Imaging Spectrometry V, 3753 (1999), 2689–2692
[2] W. Xiong et al., IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 4 (2011), 545–564
[3] S. W. Dowler, R. Takashima, and M. Andrews, IEEE Trans. Image Process., 22 (2013), 2835–2848

VARIABLE REMOVAL BY LOGICAL BLOCKS IN OPLS PREDICTIONS

Erik Johansson[1], Izabella Surowiec[1], Rafael Machleid[2], Kleanthis Mazarakis[3]
[1]*Sartorius Data Analytics, Umeå, Sweden*
[2] *Sartorius Stedim Biotech GmbH, Göttingen, Germany*
[3]*Sartorius Stedim UK Ltd, London, UK*
Erik.Johansson@Sartorius.com

In multivariate modeling of complex data, a major aim is to accomplish fast, accurate, and quantitative predictions of complex responses based on the collected body of X-data. Furthermore, with appropriate data and a workable OPLS model the X variables correlate with the predicted Y which is described by the predicted loading p_1.
We will present an automated methodology removing the logical blocks of p_1 loadings, while retaining the predictive ability of the model [1]. The automated methodology goes through the following loop until all logical blocks are removed

1. Calculate, for each X variable, a ratio of a correlation value to a confidence value of the correlation value
2. Calculate, for each logical block, an average of absolute values of the ratios calculated for the X variables of the process parameter
3. Exclude, from the data set, the logical block having the smallest average among the calculated average

With this methodology we can support the subject matter experts as they are interested to understand how the important X variables correlate with the predicted Y.

Examples from prediction of titer in biopharma and ‘prediction of glucose by Raman spectroscopy will be presented

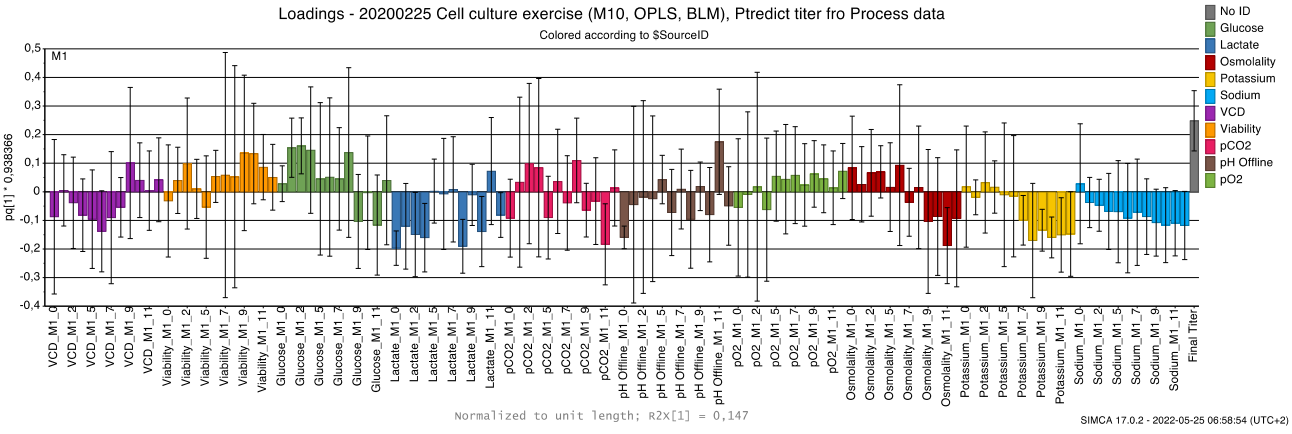


Figure 1 – p_1-loadings and confidence intervals with the local blocks with different colors

References

[1] Erik Johansson, Kleanthis Mazarakis, S14887EU-HB Patent Application, Computer implemented method, computer program product and system for data analysis

RASHOMON EFFECT AND MODEL INTERPRETABILTIY: IS IT POSSIBLE?

J.H. Kalivas[1], R. Spiers[1]
[1]*Department of Chemistry, Idaho State University, Pocatello, Idaho, USA*
kalijohn@isu.edu

A primary goal of chemometric calibration with spectral data is to form an accurate prediction model useable for analysis of new samples. Another goal is to be able to interpret the accurately predicting model. For example, once a regression vector is obtained for a spectral data set, such as with partial least squares (PLS) to predict pulp content of trees, the user may want to interpret the meaning of the regression coefficient values relative to spectral wavelengths and the prediction property. Model interpretation is also a dominant theme in today's machine learning literature [1]. However, due to the complexity of each sample with respective hidden matrix effects, model interpretation may not be possible. Sample dependent matrix effects can stem from an enormous number of possible sources. Leading the matrix effect lists are physiochemical properties such as inter- and intra-molecular interactions dependent on each sample composition (analyte and other species amounts). Biological specimens suffer from additional physiochemical properties due to cellular microenvironment variances in each sample. As with most measurements, temperature has a prevailing effect on spectra. It is thought that a model must be able to account and correct for all matrix effects to be useful. Yet, studies have shown that diverse models can be formed to accurately predict a sample(s). This situation has been labeled the Rashomon effect with the Rashomon set consisting of the collective set of useful models [2]. For example, a variety of modeling methods with proper tuning, such as PLS, deep learning, support vector machine (SVM), or random forests, applied to an appropriate data set will all sufficiently predict the analyte. In this case, a large Rashomon set is said to exist. Recent work connects the Rashomon set of useful models to interpretability [3]. This presentation discusses the Rashomon effect but proposes that strict model interpretability may not be feasible since many models satisfy the requirements to be an interpretable model yet each would be interpreted differently. This presentation discusses the Rashomon effect and its implication in model interpretation. Demonstrating the Rashomon effect are two mathematical approaches from the chemometric literature characterizing the multitude of diverse models that can be achieved and predict accurately [4,5]. Both approaches establish useful models deviating from the orthogonal net analyte signal (NAS) model.

References

[1] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, *Stat. Surveys* **16** (2022) 1-85.
[2] L. Breiman, *Statistical Sci.* **16** (2001) 199-231.
[3] L. Semenova, C. Rudin, R. Parr, https://doi.org/10.48550/arXiv.1908.01755 (v4 2022).
[4] C. Brown, R. Green, *Trends Anal. Chem.* **28** (2009) 506-514.
[5] J.H. Kalivas, J. Ferré, A.J. Tencate, *J. Chemom.* **31** (2017) 1-23.

DIAGNOSTIC PLOTS TO AID FINAL MODEL SELECTION

L. Lawrence[1], B. M. Wise[1],
[1]*Eigenvector Research Inc., Manson, WA, USA*
lawrence@eigenvector.com

Final model selection can be an overwhelming endeavor. The combination of model types, meta-parameters, preprocessing methods, and variable selection lead to a large number of models to consider for deployment. It is useful to consider both the predictive performance of the model and the degree of overfitting. One possibility is to plot $R^2 - Q^2$ versus Q^2 (where Q^2 is the cross-validation equivalent of the calibration R^2) [1]. One problem with this plot is that, unlike the root-mean square error of calibration and cross-validation (RMSEC and RMSECV), it is not in the units of the variable to be predicted, thus it is sometimes hard to assess whether the model is fit for purpose. Furthermore, it has a non-linear relationship with these metrics. Here we present an alternative plot which can aid in model selection, which is to plot the ratio of RMSECV/RMSEC versus RMSECV. This plot makes it easy to find models that are not overfit and still have a small error of cross-validation. We also consider adding contours to the plot to aid in finding the model which is "closest" to the perfect model, which would be a model with RMSECV equal to the reference error and no overfitting, *i.e.* RMSECV/RMSEC = 1. Additional plots to assess model robustness are also considered. Models can be tested to shifts in the wavelength axis and to synthetic interferents. This shows how much the model is sensitive to an unstable instrument or to new minor components in the test samples.

References

[1] K. Mendez, S. Reinke, D. Broadhurst. *Metabolomics.* 15(2019) 150

Bayesian Multivariate Receptor Modeling Software: BNFA and bayesMRM

Eun Sug Park[1], Eun-Kyung Lee[2], Man-Suk Oh[2]

[1] *Texas A&M Transportation Institute,College Station, TX, USA*
[2]*Ewha Womans Univsersity,Seoul, Korea*
msoh @ewha.ac.kr

We present user-friendly software tools to implement Bayesian multivariate receptor modeling in the form of a MATLAB function (BNFA) and an R package (bayesMRM). A basic model and a Markov chain Monte Carlo algorithm underlying BNFA and bayesMRM are given. An example of implementation based on real air pollution data is also provided. Users can freely choose between BNFA and bayesMRM depending on their computing platform. These tools are expected to facilitate implementation of Bayesian multivariate receptor models and/or Bayesian nonnegative factor analysis models and promote their use in chemometrics.

COMPLIANT CLASS-MODELS BASED ON PLS2 TO HANDLE SEVERAL CATEGORIES ENCODED WITH ERROR CORRECTING OUTPUT CODES

M.S. Sánchez[1], M.C. Ortiz[2], S. Ruiz[1], O. Valencia[1], L.A. Sarabia[1]
[1]*Dpto. Matemáticas y Computación, Universidad de Burgos, Burgos, Spain*
[2]*Dpto. Química, Universidad de Burgos, Burgos, Spain*
ssanchez@ubu.es

The work presented here is framed in the characterization problems of several categories, which represent the only ones of interest for the study. It is about simultaneously modeling *K* categories by focusing on the usual sensitivity of each individual class model and the specificity between each two classes [1]. In that sense, as the *K* classes are the unique classes of interest, the *K*-class-models (regarded as a unit) are compliant class-models rather than rigorous class-models (also known as one-class-classifiers) [2].

This means that the K-class-model is evaluated in terms of a sensitivity-specificity matrix, that is, a K× *K* matrix with sensitivities in the main diagonal and pair-wise specificities in the off-diagonal terms. The evaluation of the models is easier to handle if the matrix can be summarized in a global index. The insensitivity of the usual global indices made it necessary to look for a new one, resulting in the Diagonal Modified Confusion Entropy, DMCEN, an entropy-based index which has shown to be more sensitive and competitive [3]. Its computation can be done with the code made available through MATLAB Central File Exchange [4].

In a data-driven strategy, DMCEN is used as a criterion to select the proper codification of the K categories to be fitted with multiresponse PLS (Partial Least Squares), typically denoted as PLS2. Instead of encoding the classes with the usual class indicator variables (also known as One versus All, OVA), the developed methodology includes a new coding system based on optimizing up to five criteria to find optimal coding matrices within the Error Correcting Output Codes (ECOC) encoding.

Irrespective of the encoding, a new decoding system is also developed by using threshold values on the probability to assign objects to class models.

The procedure is applied to several datasets, comparing the performance of the found ECOC encoding with the models computed with the OVA encoding. In all the cases studied, the computed K-class-model with the *ad-hoc* ECOC encoding showed an improvement over the models based on OVA.

References

[1] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, *Anal. Chim. Acta*, **820** (2014) 23–31.
[2] O.Y. Rodionova, P. Oliveri, A.L. Pomerantsev, *Chemometr. Intell. Lab. Syst.*, **159** (2016), 89-96.
[3] O. Valencia, M.C. Ortiz, M.S. Sánchez, L.A. Sarabia, *Chemometr Intell Lab Syst.* **217** (2021) 104423.
[4] M.S. Sánchez, O. Valencia, S. Ruiz, M.C. Ortiz, L.A. Sarabia, (2022) DMCEN a MATLAB function to evaluate the entropy improvement provided by a multivariate k-class-model. MATLAB Central File Exchange in https://www.mathworks.com/matlabcentral/fileexchange/112175-dmcen. Retrieved May 27, 2022.

ARE WE THERE YET? EFFICIENT EXPLORATION AND VISUALIZATION OF MULTIVARIATE DATA WITH SCORXPLOR

Thays R. Gonçalves[1], Peter D. Wentzell[2], Makoto Matsushita[1], Paulo H. Março[3], Patrícia Valderrama[3]

[1] *Universidade Estadual de Maringá, Maringá - Paraná, Brazil*
[2] *Dalhousie University, Halifax – Nova Scotia, Canada*
[3] *Universidade Tecnológica Federal do Paraná (UTFPR), Campo Mourão - Paraná, Brazil*
patriciav@utfpr.edu.br / pativalderrama@gmail.com

Even with the advance of many complex chemometric tools, exploratory data analysis through data visualization remains one of the most widely applied approaches in chemistry and other fields. Data visualization, typically through subspace projection methods, has the advantage of permitting human visualization of the relationships among objects. The evaluation of data analysis has become more challenging as well, with options that range from which signal processing method to use (smoothing, differentiation, multiplicative signal correction, etc), to more complex choices such as which decomposition methods to use (PCA, ICA, etc) or if the data fusion should be considered. The ScorXplor algorithm has the answer. This software enhances the exploratory capability of data analysis through a visualization tool to evaluate the benefit of data fusion and the effects of preprocessing in the different obtained projections. **Fig. 1** illustrates the overall process used in this work.

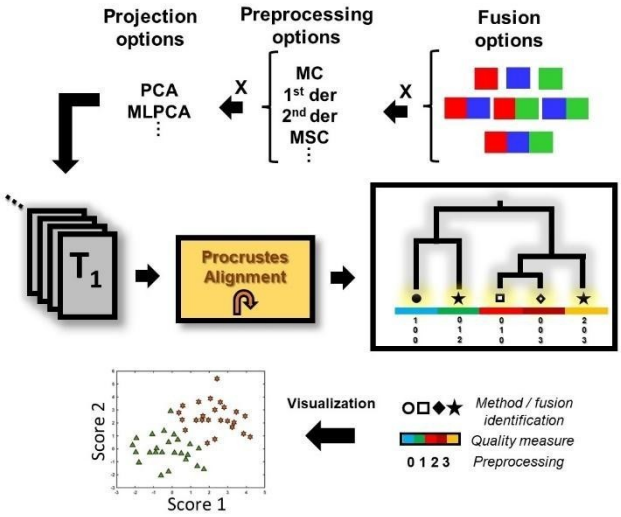


Figure 1. Scheme of the algorithm procedure

Starting with the data in the upper right, the analysis may include multiblock data, which can be combined in various ways. The following steps are designated for preprocessing methods and multivariate projection tools. The result of the collection of projections is presented as the scores matrices, which go through a relational analysis stage, where Procrustes analysis is used to align the projections and assess their similarities. This information is summarized as a dendrogram obtained from hierarchical cluster analysis in the final step. The ScorXplor provides a simple overview of the projection spaces and includes features that allow the rapid identification of the projections and the possibility of cluster quality assessment. In this presentation, the utility of ScorXplor software will be demonstrated using the fusion of three spectral data sets for the exploratory analysis of olive oils with the application of different preprocessing and projections methods.

Acknowledgments
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Finance Code 001 (PDSE-CAPES-88881.362191/2019-01), Fundação Araucária, and Natural Sciences and Engineering Research Council of Canada – NSERC (ref. no. 46316).

Chemical variation of sugar beet subjected to long-term storage by Vis-NIR spectroscopy, Hyperspectral Imaging and chemometric methods

M. Rojas[1,2], C. Fuentes[1,2], M. Oztop[3], A. Ozgur[4], R. Castillo[1,2]
[1]*Biospectroscopy and Chemometrics Laboratory, Biotechnology Center, Universidad de Concepción, Concepción, Chile*
[2]*Instrumental Analysis Department, Faculty of Pharmacy, Universidad de Concepción, 4070386 Concepción, Chile.*
[3]*Food Engineering Department, Middle East Technical University, Ankara, Turkey*
[4]*Kayseri Sugar, Kayseri, Turkey*
macarrojas@udec.cl

Twenty percent of the world's sugar production is obtained from sugar beets, whose manufacturing efficiency depends mainly on the quality of the raw material [1]. Prior to processing, sugar beets are subjected to long-term storage, which promotes the degradation of some compounds, especially inversion of sucrose to glucose and fructose, which reduces their quality [2,3]. Infrared (IR) spectroscopy techniques are powerful analytical tools that allow to quickly characterize samples based on their interaction with radiation. Hyperspectral Imaging (HSI) provides simultaneous spatial and spectral information of the samples, where each pixel contains chemical information [4]. The objective of this research was to analyze the chemical variation of beet subjected to a long-term storage period, through of portable Vis-NIR spectroscopy and Hyperspectral Imaging, enhanced with the application of chemometric techniques. For this purpose, beets were stored in a growth chamber at 20°C and sampled at 1, 28, 40, 84 and 104 days. The Vis-NIR spectra and the hyperspectral images of slices were acquired in the range of 400-1100nm and submitted to pattern recognition methods, for differentiation of storage effect of samples and for the evaluation of intrasample distribution of components, respectively. Principal Component Analysis (PCA), Soft Independent Modelling of Class Analogies (SIMCA), Support Vector Machine Discriminant Analysis (SVM-DA) and Spectral Angle Mapper Classification (SAM) were some of the techniques used. The PCA scores of VIS-NIR spectra showed a clear separation between samples that were subjected and not subjected to long-term storage. The classification models validated by cross-validation (Venetian blind) presented between 98-100% of correctly assigned samples. On the other hand, SAM allowed to observe how the chemical distribution in the beets varied during storage (Figure 1). These results are promissory to monitor storage effects on sugar beets in a rapid way for biomass selection in sugar industry.

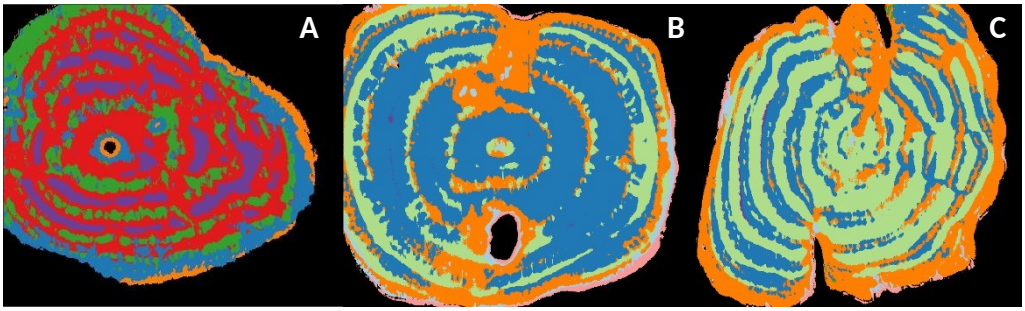


Figure 1 – Spectral Angle Mapper of HSI of sugar beet without storage (A), after 28 (B) and 40 (C) days of storage.

Acknowledgments
The authors are grateful European Union's Horizon 2020 Research and Innovation Programme—MSCA RISE under grant agreement # 101008228 and to ANID Subdirección de Capital Humano/Doctorado Nacional/2021- 21210494.

References
[1] OCDE/FAO. OCDE-FAO Perspectivas Agrícolas 2019-2028 (2019).
[2] C. Kenter, C. Hoffmann. *Int. J. Food Sci. Technol*. **44** (2009) 910–917
[3] K. Schnepel, C. Hoffmann. *SUGAR IND*. **138** (2013) 463-470.
[4] B. Boldrini, W. Kessler, K. Rebner, R. Kessler. *J Near Infrared Spectrosc.* **20** (2012) 438-508.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

The NMR side of lentil: protein extraction and hydrolyzation, and a bit of data fusion

F. Savorani[1], N. Cavallini[1], M. Sozzi[1], E. Cazzaniga[1], F. Geobaldo[1]
[1]*Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino – 10129 Torino, Italia*
francesco.savorani@polito.it

In past years, the interest towards bioeconomy concepts has been considerably growing. In particular, the development of sustainable and renewable bio-based technologies for food production is becoming increasingly important. One of the most interesting applications of bioeconomy in the "food" area is the use of enzymes for the transformation of food ingredients, waste or by-products [1], to improve food safety and optimize the overall food treatment process.
In this perspective, the present study is focused on the optimization of the parameters used for lentil flour treatment, which is known as a "functional food" in the field of food supplements. Batches of lentil ground flour, after an initial extraction process at a fixed pH and temperature, were hydrolyzed changing the most critical process control parameters in accordance with a simple design of experiment (DoE): amount of added protease enzyme, different stirring rate and the effective treatment time. A total of 32 different samples were obtained and analyzed using both UV-Visible, NIR and NMR spectroscopic techniques. Samples were collected at specific timepoints of the process and immediately frozen to avoid chemical changes before analysis. After thawing, they were first analyzed by ^{1}H-NMR spectroscopy and then by NIR and UV-Vis spectroscopies, to compare and merge the outcomes of these different characterization techniques.
All spectra were imported into MATLAB software for the multivariate chemometrics analysis. Using principal component analysis (PCA) we explored similarities among the samples to look for time dependent trends and discrepancies with respect to different factors' levels (Figure 1). With partial least square discriminant analysis (PLS-DA) we also created a model able to clearly distinguish the samples treated with the protease enzyme from those untreated.
Finally, to improve the results and get a better overview, we also performed a low-level data fusion by merging spectroscopic data from NMR, NIR and UV-Visible spectra. This kind of approach helped explaining some odd results observed in the scores plots of the PCA performed on NMR data alone.

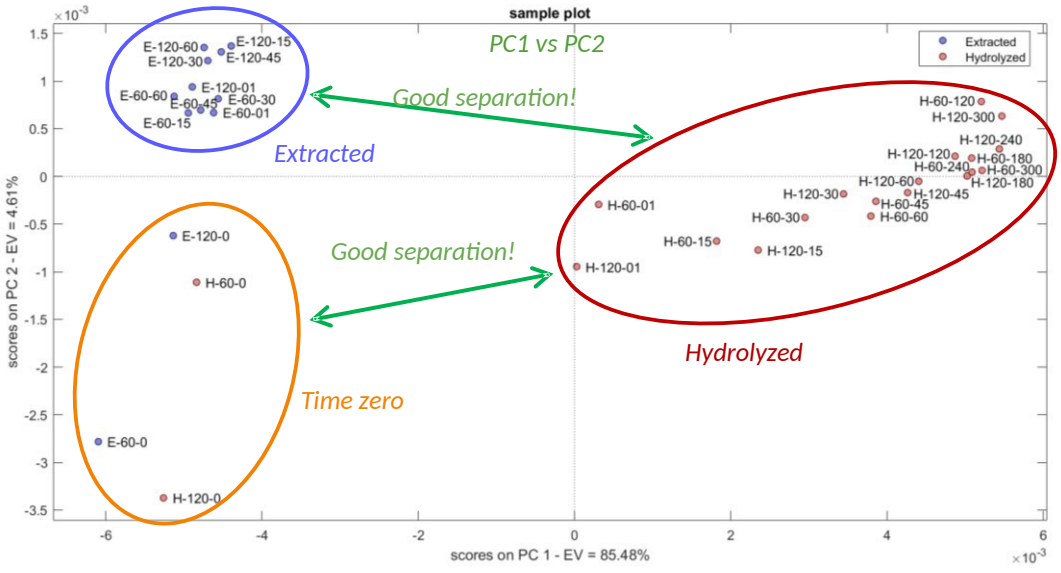


Figure 1 – PCA scores plot (PC1 vs PC2) obtained from all NMR spectra chemometrics analysis

References

[1] O. L. Tavano, *Journal of Molecular Catalysis B: Enzymatic,* **90** (2013) 1-11.

EXPLORATIVE STUDY OF STRAWBERRY JUICE FROM VARIOUS FRUIT VARIETIES USING ABSORBANCE-TRANSMISSION AND FLUORESCENCE EXCITATION-EMISSION MATRIX TECHNIQUE

P. Nowak[1], K. Pawlak-Lemańska[2], E. Sikorska[2], M. Sikorski[1]
[1]*Faculty of Chemistry, Department of Spectroscopy and Magnetism, Adam Mickiewicz University in Poznań, Poznań, Poland*
[2]*Department of Technology and Instrumental Analysis, Institute of Quality Science, Poznań University of Economics and Business, Poznań, Poland*
sikorski@amu.edu.pl

The strawberry (*Fragaria× ananassa*) is a popular fruit with attractive sensory attributes and a high content of nutrient and bioactive non-nutrient components. Polyphenols present in strawberries contribute to their health-promoting effects and attractive sensory attributes such as color, flavor, astringency, and hardness. The subject of this study was to characterize the UV-VIS absorption and fluorescence spectra of juice obtained from various varieties of strawberries and to evaluate the feasibility of using spectral data for identification of fruit variety. An absorbance and transmission excitation-emission matrix (A-TEEM™) technique was used for the measurements of the spectra. This techniques enables rapid simultaneous acquisition of absorption spectra and fluorescence excitation-emission matrices EEMs [1] and enables correction of inner filter effects.

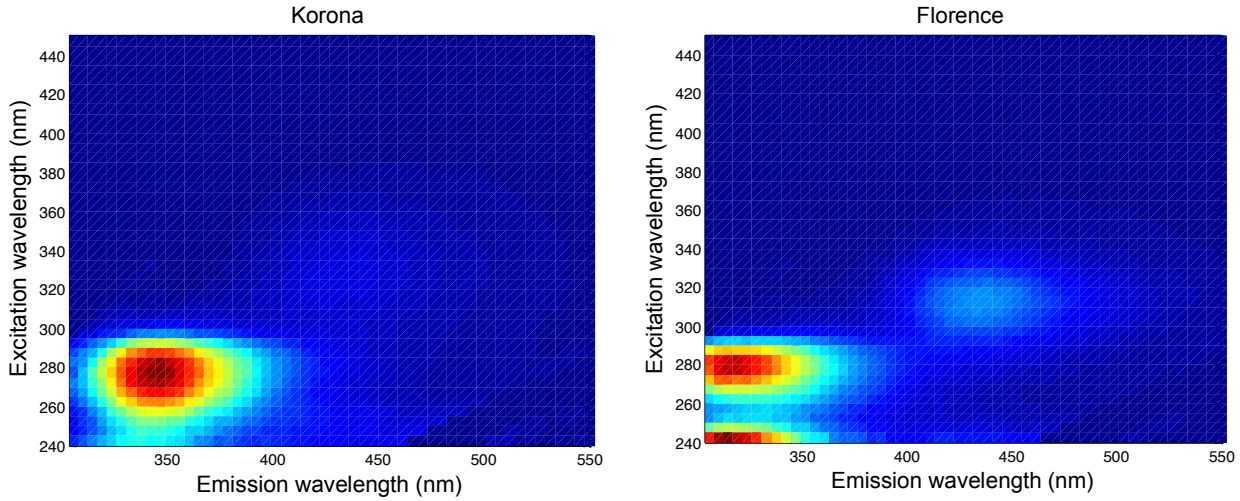


Figure 1 – Fluorescence excitation emission matrices (EEMs) of strawberry juice obtained from fruits of two various varieties Korona and Florence.

The obtained spectra were analyzed using chemometric methods. Principal component analysis (PCA) revealed differences in spectral properties of juices obtained from various fruit varieties. The parallel factor analysis (PARAFAC) was used to characterize excitation and emission profiles and relative contribution of fluorescent components of juices. Partial least squares discriminant analysis (PLS - DA) enabled good discrimination of juices from studied fruit varieties. The variable importance in projection (VIP) was used to identify the spectral regions that contributed to the differentiation of classes of juices.

Acknowledgement
Grants 2016/23/B/NZ9/03591 and 2017/27/B/ST4/02494 from the National Science Centre, Poland, are gratefully acknowledged, project was also supported by the HighChem interdisciplinary grant nr. POWR.03.02.00-00-1020/17.

References
[1] A. Quatela, A. M. Gilmore, K. E. Steege Gall, M. Sandros, K. Csatorday, A. Siemiarczuk, B. Yang, L. Camenen Methods Appl Fluores. 6 (2018), 027002.

Chemsy: Simultaneous feature selection, pre-processing search, model selection, and hyper-parameter optimization in Python

Sin Yong Teng[1], Martijn Dingemans[1], Maria Cairoli[1], Jeroen J. Jansen[1]
[1] *Radboud University, 6525 AJ Nijmegen, the Netherlands*
SinYong.Teng@ru.nl

Chemsy is a chemometrics and machine-learning framework written in Python with a sklearn syntax to allow for flexible usage. The Chemsy framework was designed to provide automated data modelling features with the consideration of full flexibility of the user. Here, we demonstrate the simultaneous capabilities of Chemsy in feature selection, pre-processing search, model selection, and hyperparameter optimization in spectroscopic modelling via examples of regression and classification using standard benchmark dataset. With the Chemsy framework, specific optimization algorithms can also be customized to be used for automatic pipeline search. For example, we demonstrate the use of Chemsy with various different optimization solvers such as brute force search, random search, design of experiment strategy [1], genetic algorithm, particle swarm optimization, etc. The use of Chemsy allows for a simplistic, yet fully customizable chemometrics platform for upcoming applications.

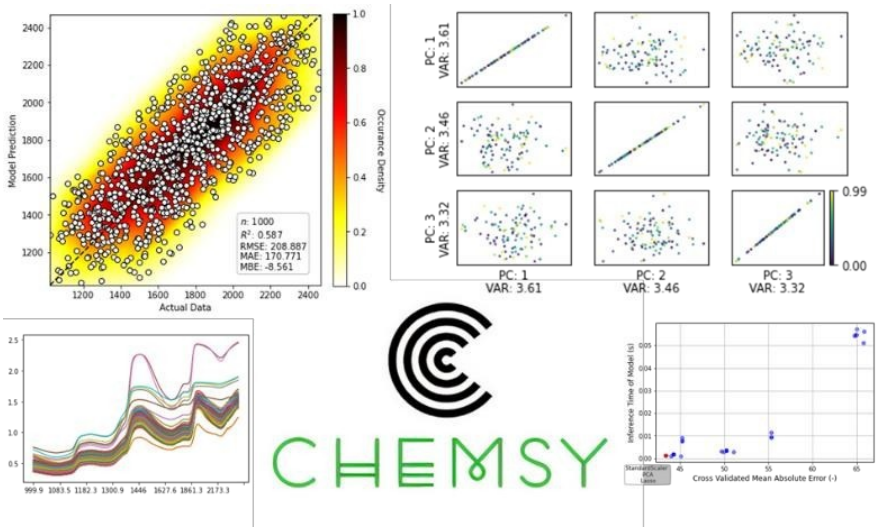


Figure 1 – Illustrative Figure for Chemsy

References

[1] Gerretzen, Jan, Ewa Szymańska, Jeroen J. Jansen, Jacob Bart, Henk-Jan van Manen, Edwin R. van den Heuvel, and Lutgarde MC Buydens. "Simple and effective way for data preprocessing selection based on design of experiments." Analytical Chemistry 87, no. 24 (2015): 12096-12103.

RELIABLE DETERMINATION OF THE LIPIDIC PROFILE OF OILS EXTRACTED FROM FISH BY-PRODUCTS THROUGH NEAR INFRARED SPECTROSCOPY AND CHEMOMETRICS

S. Nieto-Ortega[1], I. Olabarrieta[1], E. Saitua[1], G. Arana[2], G. Foti[1] and Á. Melado-Herreros[1]
[1]*AZTI, Food Research, Basque Research and Technology Alliance (BRTA), Parque Tecnológico de Bizkaia, Astondo Bidea, Edificio 609, 48160 Derio, Spain*
[2]*Department of Analytical Chemistry, University of the Basque Country UPV/EHU, Sarriena S/N, 48940 Leioa, Spain*
snieto@azti.es

The fishing industry produces a large amount of residues. In fish processing, around 50 % of the total fish weight is converted into solid waste and by-products. However, these should not be considered waste materials, since fish by-products usually have a high nutritional value and, therefore, a great potential to be reused in higher-value applications [1]. Fish oils contain high levels of eicosapentanoic acid (EPA 20:5) and docosahexanoic acid (DHA C22:6), two fatty acids (FAs) of great importance in human nutrition. Hence, the analysis of the FAs profile is essential since it will determine the price of the oils, their management and the way they will be reused. Normally, they are analysed by gas chromatography with flame ionization detector (GC-FID). However, this technique is invasive and time consuming, requiring many steps. Thus, developing new tools to determine the lipid profile of the fish oils is of great interest for the industry [2].

In the last years, spectroscopic methods coupled with chemometrics techniques have emerged as an alternative to monitor the quality of many food products. In this work, a handheld near infrared (NIR) spectrometer has been used for the on-site determination of the FAs composition of industrial fish oils derived from fish by-products. Eight fish oil samples, coming from unknown fish by-products, were used to make 269 different mixtures. GC-FID was employed as the reference method and samples were scanned with the MicroNIR OnSite from Viavi, working from 900 to 1650 nm, with a resolution of 6 nm [3].

Several pre-processing methods for NIR signals were compared: multiplicative scatter correction (MSC), standard normal variate with and without detrend (SNVd and SNV), Saviztky–Golay first and second derivatives (using different polynomial orders and windows) and different combinations. In all the cases, the NIR signal and the percentage of FAs (X and Y data, respectively) were mean centered before the multivariate analysis. After that, different partial least square regression (PLSR) models were developed to correlate the spectra with the percentage of saturated fatty acids (SFAs), monounsaturated fatty acids (MUFAs), polyunsaturated fatty acids (PUFAs) and, among them, omega-3 (ω-3) and omega-6 (ω-6) FAs. First, data was divided in two datasets, used for creating and validating the five models. After that, all the data were mixed for building a new dataset (n_c = 269), used for calibration purposes. A random CV with 20 segments was used for testing the models. Then, they were uploaded into the MicroNIR OnSite sensor, to perform onsite an external validation in real time, using an external dataset (n_{test} = 29).

Results regarding the external validation in the prediction of SFAs, MUFAs, PUFAs and ω-3 were good, with $R^2 \geq 0.95$, RMSEP ≤ 1.71 in all the cases, and a bias value of - 0.78%, - 0.12%, - 0.80% and - 0.67%, respectively. However, the model created for the prediction of the ω-6 FAs failed (RMSEP = 2.09% and bias = - 1.76%) due to the low variability between samples. However, this was corrected applying a bias and slope correction (BSC), obtaining a R^2 of 0.95, a RMSEP of 1.09% and a bias value of -0.05%. Future works will be focused on improving this model, adding new samples with more variability regarding the ω-6 content.

References

[1] N. Rubio-Rodríguez, S.M. de Diego, …, J. Rovira, *J. Food Eng.* **109** (2012) 238-248.
[2] J.H. Cheng, D.W. Sun, …, Y.N. Chen, *Food Chem.* **270** (2019) 182-188.
[3] S. Nieto-Ortega, I. Olabarrieta, …, Á. Melado-Herreros, *Foods* **11** (2022) 55.

A MULTIVARIATE APPROACH TO QUANTIFY THE ENHANCEMENT EFFECT ON SURFACE-ENHANCED SPECTROSCOPIES

C. F. Pereira*[1], I. M. A. Viegas[1], I. G. Souza Sobrinha[1], G. Pereira[1], G. A. L. Pereira[1], P. Krebs[2], F. C. S. Trindade[1], B. Mizaikoff[2]

[1] *Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, Brazil*
[2] *Institute of Analytical and Bioanalytical Chemistry, Ulm University,Ulm, Germany*
claudete.fernandes@ufpe.br

Raman and Infrared (IR) spectroscopies are well-established techniques for the unambiguous identification of molecular species. However, these techniques are not sensitive enough to detect small quantities of analytes. This can be mitigated by using surface enhancement strategies. That means the employment of substrate surfaces, which are generally irregular metals, semimetals, semiconductors or polar dielectric nanostructures (i.e., smaller than the wavelength of light), to enhance Raman (a.k.a., Surface-Enhanced Raman Spectroscopy - SERS) and IR (a.k.a., Surface-Enhanced Infrared Absorption Spectroscopy - SEIRA) spectra [1,2]. In this study, we are proposing an alternative approach to quantify the enhancement effect termed 'Multivariate Enhancement Factor' (MEF), which may be more suitable for optimizing the experimental conditions in surface-enhanced studies. For this, attenuated total reflection mid-infrared (MIR) spectra were recorded via: 1) a Bruker Tensor II FT-IR spectrometer equipped with a deuterated triglycine sulfate (DTGS) detector (Bruker Optics, Ettlingen, Germany) and a BioATRII unit (Bruker Optics, Ettlingen, Germany), which has a circular silicon ATR plate providing 8–10 internal reflections; 2) a FTIRSpectrum400 spectrometer (Perkin Elmer) with a universal attenuated total reflectance (UATR) accessory (diamond crystal), where spectra were recorded in the spectral region from 4000 to 650 cm^{-1}. All spectra were preprocessed via a baseline offset correction and mean centered. Silver selenide Quantum Dots (QDs) stabilized via mercaptosuccinic (Ag_2Se–MSA) and mercaptopropionic (Ag_2Se–MPA) acids in aqueous suspension were used for amplifying the IR signature of a variety of dye molecules (auramine, fuchsine, methyl violet 2B, neutral red, rhodamine 6G, and rhodamine B) and atrazine solutions, respectively. IR spectra without and with Ag_2Se QDs composed the data set, and Principal Component Analysis (PCA) models were built. PCA is a suitable way to describe the interpoint distance using as few dimensions as possible. The scores plot carries the main information on the enhancement effect promoted by Ag_2Se QDs vs. the molecules. Therefore, the enhancement effect can be determined by using the interpoint distances from each pair of samples (i.e., with and without Ag_2Se–MSA) along the PC1 and PC2 axes. Usually, the straight line distance between two points in an n-dimensional space with coordinates (x_1, x_2,…, x_n) and (y_1, y_2,…, y_n) is taken as the Euclidian distance [3]. Therefore, the distance between scores for the same dye on PC1 and PC2 axes without and with Ag_2Se QDs determines a novel parameter we call the 'Multivariate Enhancement Factor' (MEF). In addition, the interaction process between auramine and Ag_2Se–MSA QDs, was also investigated using PCA model employing spectral data from 1470 to 1300 cm^{-1} (time-resolved spectra recorded during the entire process of solvent evaporation), which contains spectral information on N–Ph, C–N and carboxylate anions vibrations. As conclusion, the introduced novel metric, which much more realistically quantifies the enhancement effect across the entire spectrum rather than at a single wavelength, is also useful and more suitable for optimizing the experimental conditions during surface-enhanced studies. It is also worth noting that the MEF is considered a generic metric that can also be used to quantify the enhancement effect in SERS.

This work was funded by CAPES-PRINT/UFPE, CNPq/FAPESP/INCTAA and Serrapilheira Institute.

References

[1] A. Hartstein, J.R. Kirtley, J.C. Tsang, Phys. Rev. Lett. 45 (1980) 201–204. https://doi.org/10.1103/PhysRevLett.45.201.

[2] C.F. Pereira, I.M.A. Viegas, I. Souza Sobrinha, G. Pereira, G.A. de L. Pereira, P. Krebs, B. Mizaikoff, J. Mater. Chem. C. (2020).

[3] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, 5th ed., Pearson Education Limited, 2005.

VIRGIN OLIVE OIL EXCITATION-EMISSION MATRICES: EXPLORING THEIR USEFULNESS TO PREDICT TASTE ATTRIBUTES

Beatriz Quintanilla-Casas[1,2], Åsmund Rinnan[3], Agustí Romero[4], Francesc Guardiola[1,2], Alba Tres[1,2], Stefania Vichi[1,2], Rasmus Bro[3]

[1]*Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Santa Coloma de Gramenet, Spain.*

[2]*Institut de Recerca en Nutrició i Seguretat Alimentària (INSA-UB), Universitat de Barcelona, Santa Coloma de Gramenet, Spain.*

[3]*Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958, Frederiksberg C, Copenhagen, Denmark.*

[4]*Institute of Agrifood Research and Technology (IRTA), Fruit Science Program, Olive Growing and Oil Technology Research Team, Mas Bové Ctra, Reus-El Morell Km 3,8 43120-Constantí, Tarragona, Spain.*

beatrizquintanilla@ub.edu

Unlike other food products, virgin olive oils must undergo an organoleptic assessment in order to be graded into a given commercial category, according to their quality grade. Given that the current official method for sensory evaluation is based on a trained human panel, it presents several drawbacks that might affect the efficiency and robustness of the method. For this reason, disposing of instrumental methods that could serve as screening tools to support the sensory panels is of paramount importance. Sensory parameters involved in commercial classification are generally linked with the aroma profile of virgin olive oil (fruitiness and off-flavours), therefore analytical methods based on volatile organic compounds have been developed to become useful screening tools [1]. However, the need for an efficient instrumental method for assessing tasting-related attributes in virgin olive oils has not been resolved yet.

The present work aimed to investigate the application of excitation-emission fluorescence spectroscopy (EEFS) in virgin olive oil to predict bitter and pungent attributes, since both organoleptic properties are known to be related with polar phenolic compounds, which are fluorophores [2]. Bitterness and pungency intensities of 250 samples were provided by an official sensory panel and used to build and compare partial least squares regressions (PLSR) with the excitation-emission matrix (EEM), after proper pre-processing. Both PARAFAC scores and two-way unfolded data led to successful PLSR, given that errors in prediction were always close to the error of the sensory reference method. According to PLS regression vectors, the most relevant PARAFAC scores for both attributes agreed with virgin olive oil phenolic spectra. This fact evidenced that EEFS would be the fit-for-purpose screening tool to support the sensory panel in this regard.

This study has been supported by the Spanish Ministry of Universities predoctoral fellowship (FPU16/01744), with its corresponding short-term mobility grant (EST19/00127), and by the grant RYC-2017-23601 funded by MCIN/AEI/ 10.13039/501100011033 and by "ESF Investing in your future". The authors aknowlege the Catalan cooperatives that provided traceable virgin olive oil samples, as well as the official tasting panel of virgin olive oil of Catalonia.

References

[1] B. Quintanilla-Casas, M. Marin, F. Guardiola, D.L. García-González, S. Barbieri, A. Bendini, T. Gallina Toschi, S. Vichi & A. Tres, *Foods*. **9** (2020), 1–14.

[2] A.M. Inarejos-Garcia, A. Androulaki, M.D. Salvador, G. Fregapane & M.Z. Tsimidou, *Food Res. Int.* 42(2019), 279–284.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Multivariate Data Analysis and PAT in vaccines development: enabling multiple components quantification in complex formulations

Antonino Restivo[1], Lorenzo Di Meola[1], Daniela Pasqui[1], Stephen Luckham[1], Duccio Bianciardi[1], Silvio Colomba[1], Agnese Marcelli[1], Alessio Moriconi[1], Carlo Pergola[1]
[1]GSK, Siena, Italy

antonino.x.restivo@gsk.com

With the advancement of knowledge and new technologies, vaccines with increasingly complex matrices, and different adjuvants, have been developed to improve the efficacy, as well as for dose amount reduction. GlaxoSmithKline (GSK) developed Adjuvant Systems (AS) based on a combination of immuno-stimulants in different adjuvant formulations designed to enhance and modulate immune responses to vaccine antigens.
The current methods used to quantify the components in products containing multi-component adjuvant do not allow to obtain good results in complex matrices since they are not able to discriminate adjuvant, antigen, and excipients/stabilizers content and usually they require sample pre-treatments.
With the ambition to discriminate vaccine components (antigens and excipient/stabilizer) in complex formulations, the current study explores the capabilities of UV/Vis in combination with multivariate data analysis (MVDA) techniques. The main advantage of the UV/Vis method is that no prior knowledge of the refractive index of a sample is necessary, and, at the same time, the method is fast and easy to operate.
For this study, an ad-hoc model matrix was created using various concentrations of multicomponent adjuvant, Ovalbumin (as model antigen) and surfactant (as stabilizer). The corresponding UV/Vis spectra were recorded (wavelength range from 200 to 800 nanometers) and the data were analyzed using chemometrics and MVDA techniques. The study has pursued the ambitious aim of quantifying the three components in matrix: adjuvant, antigen, and stabilizer.
The MVDA based on the partial least square (PLS) model was used to build the calibration model for each of the monitored outcomes. The application of the PLS to the model matrix shows that the UV/Vis method is suitable for the simultaneous quantification of the stabilizer (Figure 1), Ovalbumin (Figure 2) and component A of adjuvant (Figure 3) throughout all sample combinations, as it's clear observing the three Observed vs Predicted plots from the validation sets reported below (color scale identifies the concentrations of the attributes). In addition, the model allows to identify the wavelengths regions in which signals fall in each sample, due to the differing absorption properties of each component.
In the Process Analytical Technology (PAT) and Quality by Design (QbD) era, this study lays the foundation for the development of new real-time analytical methods for the definition of design space and control strategy through the formulation step to improve the production of either existing commercial product or innovative future vaccines.

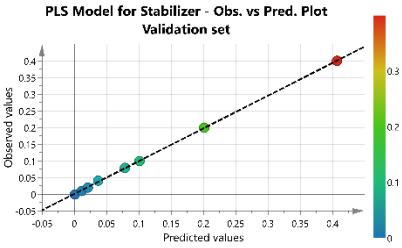


Figure 1 – Stabilizer quantification

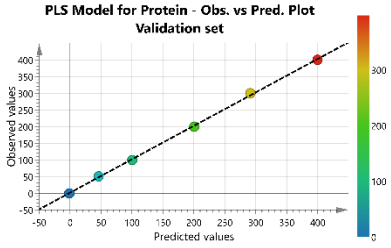


Figure 2 – Protein quantification

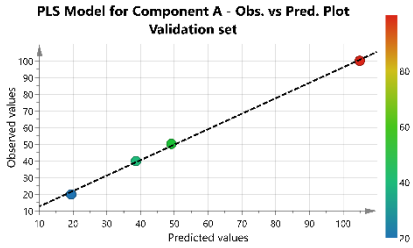


Figure 3 – Component A quantification

OPTIMIZATION OF THE PARAMETERS OF A CONTINUOUS ANNEALING PROCESS IN A STEEL PRODUCING COMPANY BY MULTIVARIATE STATISTICS AND ARTIFICIAL NEURAL NETWORKS

E. Robotti[1], V. Zippo[2], M. Pianezzola[2], S. Maggi[2], P. Fossati[2], R. Langiano[2], E. Marengo[1]
[1]Department of Sciences and Technological Innovation – University of Piemonte Orientale, Viale Michel 11, 15121 Alessandria, Italy
[2]Acciaierie Italia Spa, Strada Bosco Marengo 1, Novi Ligure, Italy
elisa.robotti@uniupo.it

The study was carried out in collaboration with “Acciaierie d'Italia Spa” (Novi Ligure – Italy) with the aim of developing an automatic method for the continuous optimization of the quality parameters of the final product based on the process parameters and on the quality of the input materials.
The continuous annealing process is a complex production plant consisting in several steps: i) continuous feeding and cleaning; ii) annealing and pre-cooling; iii) accelerated cooling, over-aging and final cooling; iv) removal of surface oxides and deposition of the final nickel layer; v) skin passing, trimming, inspection, oiling, and cutting. The result of a step obviously influences the subsequent steps performances. The final aim of the study is to correlate the process parameters and the quality of the raw materials in input with the quality observed on the final product, to provide a predictive model of the product quality and to identify the role played by the input parameters on the quality performances for process/product optimization. The study was developed in different phases:

1) Data collection and synchronization. In collaboration with computer scientists from Acciaierie Italia Spa, data extraction tools have been developed able to extract the process data regarding all the parameters measured online (speeds, voltages, temperatures, etc.), the results of the tests performed in the Quality Lab on the final product and the results of the chemical characterization of the raw materials used in input. The data needed to be carefully aligned and synchronized.
2) Data analysis by pattern recognition. The overall database, consisting in 78 variables and about 6000 collected samples, was then subjected to multivariate analysis. Principal Component Analysis (PCA) was exploited to identify groups of samples nd evaluate the correlation between the variables along the process. Groups of samples emerged from this analysis with different behaviour: these groups were identified and further investigated independently.
3) Construction of models based on the use of artificial neural networks (ANN). The variables identified as the most relevant by PCA were used as input to a back-propagation ANN (BP-ANN), to relate the yield and the degree of aging of the product to the input parameters (process parameters and quality parameters of the raw materials). Different ANNs were trained according to the groups of products identified in 2). Samples were separated in training, test and evaluation set and all the networks were optimized for what regards the architecture and the training parameters.

The study allowed to investigate multiple aspects of the complex chain of the production plant: not only aspects strictly related to the product quality, but also related to process control, how the plant data can be used and disclosed for the improvement of the production process. An automatic tool was developed, able to operate directly on the database, divide the samples by thickness and train the ANNs with good values of R^2 and RSME on the evaluation set for each category of steel.
Aspects such as the aging of the material were also analysed, and the relationship between the latter and the temperature in a particular area of the plant called Over-aging, identified as a critical area of the plant itself, was investigated. In particular, it was possible to modify the temperatures used in the Over-aging area, obtaining an improvement of the process of at least 70% in terms of reduction of the phenomenon. Finally, an internal daily report has been created which reports the assessments of the production to be shared with the production department and the company management.

References

[1] Z. Guo, W. Sha, Comput. Mat. Sci. 29 (2004) 12-28
[2] S.A. Razavi, F. Ashrafizadeh, S. Fooladi, Mat. Sci Engineer. A. 675 (2016) 147-152

Blend uniformity design space development and verification by PAT for minibatch blending

L. Rolinger[1], J. Lamerz, M. Bautista[1]
[1]*Pharma Technical Development, Hoffmann-La Roche, Basel,Switzerland*
laura.rolinger@roche.com

The control strategy for the Mini-batch continuous direct compression line relies on robust and compliant blend assay and blend uniformity for each blended mini-batch. The blend uniformity for each mini-batch is ensured by a parametric control on the blending conditions and the relevant Critical Material Attributes (CMAs).
Therefore, a blending design space has been developed with variable blender fill level, rotation time, rotation speed, and particle size distribution of the API. Two Design of Experiments (DoE) have been performed to develop and verify the design space, where the response blend uniformity has been obtained using Raman-based PAT at-line measurements.
This approach allows ensuring blend uniformity within specification for all mini-batches that are blended within the obtained design space taking measurement uncertainty into account.. The advantage of the presented work is the reduction of measurement time and costs for blend uniformity by PAT and flexibility to adjust for changes in API particle size or fill level while still remaining good blend uniformity.

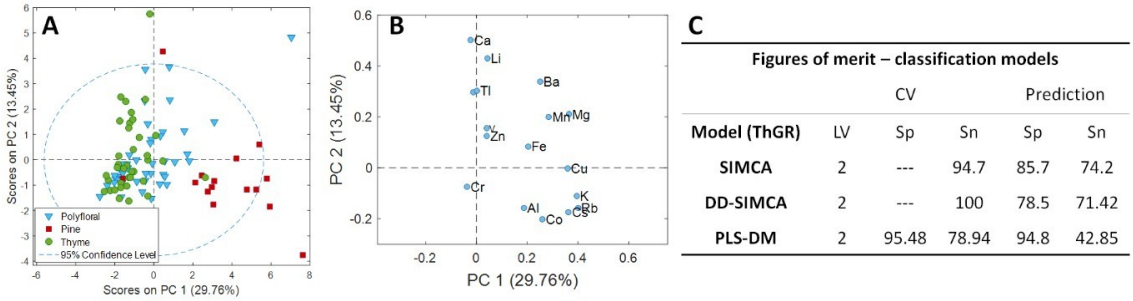
Application of class-modelling approaches for botanical and geographical origins of honey samples based on mineral content

Carolina S. Silva[1,2], Ioannis Gkikopoulos[3], Sotirios Karavoltsos[3], Aikaterini Sakellari[3], Charalampos Proestos[4], Owen Falzon[1], Vasilis Valdramidis[1,4]

[1]Centre for Biomedical Cybernetics, University of Malta; [2]Department of Food Sciences and Nutrition, University of Malta, Msida, Malta; [3]Laboratory of Environmental Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Greece; [4]Laboratory of Food Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Greece

carolina.santos@um.edu.mt

Honey is a valuable natural food product often appearing in the European Commission's food fraud summary as one of the most often adulterated/forged food products in the world [1]. Mislabeling is a major issue related to honey products, and its authentication regarding botanical and geographical origins is challenging due to the uniqueness of each honey chemical profile. However, it is possible to find consistent features that differentiate monofloral honey samples from different botanical sources [2]. The main challenge consists in differentiate certain floral types from polyfloral honeys, especially because there is no consensus in EU regarding labelling regulation. In this initial study, 96 samples of honey from 3 different botanical origins (thyme, pine and polyfloral) from different countries (Greece, Tunisia, Spain, Turkey and Malta) were analyzed by Inductively coupled plasma mass spectrometry (ICP-MS). Principal component analysis was carried out and a 6-PC model explaining 72.92% of data variation showed trends regarding the floral types in the first PC. The scores scatter plot showed a clear difference between the two monofloral types, with the polyfloral samples in between (Figure 1A). According to the data, pine and thyme honeys differ from each other regarding the content of, mainly, Mg, K, Mn, Co, Cu, Rb, Cd and Cs (Figure 1B). Differentiation regarding geographical origin can be observed in the second PC. Extreme values of heavy metals were found in some specific groups of samples. Greek and Tunisian pine showed higher values of Cu, Cd and Ni, while high Pb levels were found in Tunisian thyme, as well as Spanish and Greek polyfloral samples, indicating environmental pollution. One-class classification models were employed attempting to classify the thyme Greek honey samples. Soft independent modelling by class analogy (SIMCA), robust approach of data-driven SIMCA (DD-SIMCA) and partial least squares density modelling were compared. The results observed in Figure 1C shows that the classic approach of SIMCA and DD-SIMCA outperformed PLS-DM. Although the latter showed high values of Sp for the prediction set, maybe due to the fact that the choice for the best model uses information from the non-target class. All models succeeded in differentiate the monofloral samples, being difficult to differentiate thyme from other countries and polyfloral (which also contains thyme pollen in its composition). Those initial results show an important advance in honey botanical and geographical origins identification. Even if 100% of classification was not achieved, no pine samples and most of polyfloral samples were not classified as thyme, avoiding the laborious pollen analysis step.



C

Figures of merit – classification models

		CV		Prediction	
Model (ThGR)	LV	Sp	Sn	Sp	Sn
SIMCA	2	---	94.7	85.7	74.2
DD-SIMCA	2	---	100	78.5	71.42
PLS-DM	2	95.48	78.94	94.8	42.85

Figure 1 Resume of results. PCA (a) scores scatter plot; (b) loadings plot; (c) results for the classification models regarding thyme Greek class.

References

[1] https://knowledge4policy.ec.europa.eu/food-fraud-quality/monthly-food-fraud-summary-reports_en#year2022

[2] Ballabio, D., Robotti, E., Grisoni, F., Quasso, F., Bobba, M., Vercelli, S.,Gosetti, F., Calabrese, G., Sangiorgi, Orlandi, M., Marengo, E. Food Chem., **266** (2018) 79-89.

Application of DoE and multivariate analysis for TXRF method development and data analysis. A case-study from the agri-food sector.

G. Squeo[1], I. Allegretta[1], C. E. Gattullo[1], C. Porfido[1], F. Caponio[1], S. Cesco[2], C. Nicoletto[3], R. Terzano[1]

[1] *Department of Soil, Plant and Food Sciences (DiSSPA), University of Bari Aldo Moro, Bari, Italy*
[2] *Faculty of Science and Technology, Free University of Bozen-Bolzano, Bolzano, Italy*
[3] *Department of Agronomy, Food, Natural Resources, Animals, and Environment (DAFNAE), University of Padova, Padova, Italy*
giacomo.squeo@uniba.it

Total-reflection X-ray fluorescence (TXRF) spectroscopy is a suitable analytical method for the determination of the elemental composition of different kind of samples. In recent years, the technique has been widely used for agri-food products [1]. However, despite the growing application of TXRF, few studies have considered and applied multivariate strategies for improving method performance and usability. In particular, two critical steps could profit from the application of different chemometric tools: i) sample preparation and ii) signal analysis. For the first step, our goal was to find the best sample preparation for an organic material. In literature there is no accordance among the amounts of sample/suspender to be used. Thus, in this work, design of experiment has been used as a rational approach to find suitable conditions of sample preparation (mass of the sample and dispersant volume). A 2^2 factorial design was set up having as responses the recovery of twelve elements. The obtained response surfaces (Figure 1) allowed to identify the region(s) of the experimental domain in which a suitable recovery (80-120%) was reached for most of the elements.

For what concern the second aspect, the output of TXRF is a continuous spectrum which is multivariate in its nature. Nonetheless, in literature these signals are seldom treated directly by multivariate methods, while, most commonly, a quantitation of the single elements is carried out. Thus, we aimed at evaluating the feasibility of TXRF coupled with multivariate data analysis for the discrimination of beans (twenty-four genotypes) from two growing sites comparing this analytical approach with the most common one, *i.e.*, by using the quantified elemental composition for the multivariate analysis. The elemental dataset (144 × 12) and the spectral dataset (144 × 2312) were subjected to different preprocessing methods (according to the different nature of the data), explored by PCA and then different classification models were built and tested. The results showed that good discrimination between the growing sites could be achieved with both the datasets and approaches. In the case of the spectral dataset, the great variability associated to the bean genotypes masked that related to the growing sites which could be highlighted by using the GLSW filter (Figure 2). The practical advantage of the direct use of TXRF signals for classification purposes lied in the possibility to avoid the elemental quantification procedure (and related errors) thus speeding up the analysis and the classification assessment.

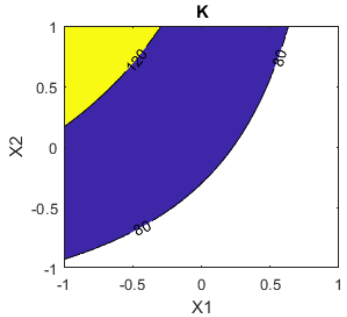


Figure 1 – Contour plot of the recovery (%) for potassium.

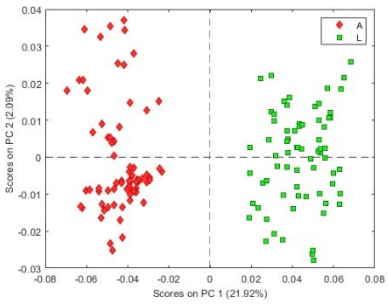


Figure 2 – Score plot of the PCA calculated on the spectral dataset using the GLSW filter.

References

[1] L. Borgese, F. Bilo, R. Dalipi, E. Bontempi, L. E. Depero, *Spectroc. Acta Pt. B-Atom. Spectr.* **113** (2015) 1-15.

A NEW SURVEY FOR MULTICOMPONENT ANALYSIS TO SOLVE PROBLEMS LINKED TO NANO-COMPOUNDS (CASE STUDY).

M. Tomassetti[1,2], R. Pezzilli[3], G. Magna[4], P.G. Medaglia[3], F. Marini[1], C. Di Natale[2].
[1]*Department of Chemistry, University of Rome, 'La Sapienza", Rome, Italy*
[2]*Department of Electronic Engineering, University of Rome ''Tor Vergata'', Rome, Italy*
[3] *Department of Industrial Engineering, University of Rome ''Tor Vergata'', Rome, Italy*
[4]*Dipartimento di Scienze e Tecnologie Chimiche, Università di Roma "Tor Vergata", Rome, Italy*

We have recently synthesized a compound of the Layered Double Hydroxide (LDH) type, that is a nanomaterial with a lamellar structure, such as hydrotalcite [1]. On this compound we were able to immobilize the catalase enzyme, with the aim of building a biosensor [1]. This purpose was actually achieved [1], as evidenced by the characterization of the compound by X-ray diffractometry and FT-IR. However, we still do not have sufficient information on how the catalase enzyme is bound to LDH. It is possible to formulate some speculations, especially taking into account data obtainable by XRD spectroscopy, carried out by us, concerning compounds of the LDH type. First of all, since the charge of the lamellae is positive and the pH of the solution in which the interaction with the enzyme took place is equal to about pH 7, the catalase (with an isoelectric point equal to 5.4) will have a prevalence of COO^- groups in his terminal chains at this pH value, so that the abundance of these negative charges will favor the interactions with the positively charged lamellae. From the X-ray spectrum of the pristine LDH it is possible to derive the interlamellar distances in the pristine LDH compounds synthesized by us, of the type $[Zn^{II}Al^{III}(OH)_2]^+ NO_3^-$, i.e., containing NO_8^- groups, which turns out to be about 5-6 Å, while, after the interaction with the catalase, this calculated distance becomes about 9-10 Å. It is therefore evident that it is impossible for the entire catalase molecule to be housed within the interlamellar space, as the size of the catalase (75-84 kDa) is about 86 Å. More likely, as reported in the literature for synthetic polymeric macromolecules (even if not of the same type), the interlamellar space will be occupied only by a part of the polymer chain, in our case the protein, which will anchor the enzyme to the LDH matrix, most probably by amino acid only. The aim of the difficult research that we have recently undertaken is precisely to try to identify which is the C terminal amino acid, belonging to one of the catalase chains, that has the greatest probability of having entered the interlamellar layer of LDH. To this end, we began to synthesize the same LDH but doped with single amino acids, always contained in the protein chains. These compounds obtained with different synthesis methods, both by coprecipitation, and by calcination and reconstruction, have been characterized, both by X-ray diffraction and by FT-IR spectra. The numerical characteristics, such as the diffraction angle (2θ) and the wavenumber of infrared radiation in (cm^{-1}) are compared with those of pristine LDH and catalase doped LDH.
The result is a data table, destined to expand in the near future, this above all, by increasing the number of different amino acids with which it will be possible to doping LDH. At present consisting of 5 "Objects" (i.e. three LDH compounds doped with three different amino acids, the pristine LDH and LDH with immobilized catalase) and 10 features for each object (consisting of from 4 parameters for each compound, deriving from the XRD spectra, and from 6 parameters, deriving from the FT-IR spectra), on which a multivariate analysis, i.e., PCA was performed. The hope is that, through multivariate analysis, useful information can be obtained, which should provide at least useful clues, about which amino acid, belonging to the catalase chain, is most likely to have interacted with LDH.
In conclusion, we believe that this research undertaken by us, which is destined to expand, may prove useful not only to help a solution to the problem that we have proposed to solve in this note, but that it can also open up new possibilities for applications, for multivariate analysis, in solving problems related to the synthesis of nano-compounds with complex structure, when these interact with materials, especially of the polymeric type, both natural and synthetic.

References

[1] Tomassetti, M.; Pezzilli, R.; Prestopino, G.; Di Natale, C.; Medaglia, P.G. *Microchem. J.* **2021**, *170*, 106700, doi:10.1016/j.microc.2021.106700.

DISCRIMINANT CLASSIFICATION MODELS APPLIED TO HAZELNUT UNSAPONIFIABLE FINGERPRINT FOR GEOGRAPHICAL AND VARIETAL AUTHENTICATION

B. Torres-Cobos[1,2], M. Rovira[3], A. Romero[3], B. Quintanilla-Casas[1,2], F. Guardiola[1,2], A. Tres[1,2], S. Vichi[1,2]

[1] Department of Nutrition, Food Sciences and Gastronomy, Faculty of Pharmacy and Food Science, University of Barcelona, Santa Coloma de Gramenet, Spain.

[2] Institute of Research on Food Nutrition and Safety (INSA-UB), University of Barcelona, Santa Coloma de Gramenet, Spain.

[3] Institute of Agrifood Research and Technology (IRTA), Constantí, Spain.

bertatorres@ub.edu

Hazelnuts are valuable components of the Mediterranean diet for their high nutritional value and sensory attributes, which also make them a fundamental ingredient in the chocolate, confectionery and bakery industries [1,2].

Hazelnuts composition and qualitative characteristics are influenced by cultivar and growing conditions in their country of origin [2-5]. Hence, verifying the hazelnut cultivar and geographical origin is relevant to protect consumers form misleading information.

Hazelnut unsaponifiable fraction contains several secondary metabolites such as hydrocarbons, carotenoids, tocopherols, terpenic alcohols and sterols. Since these compounds arecharacteristic of the hazelnut cultivar and the pedoclimatic conditions [4-7] the fingerprint of the unsaponifiable fraction could be a suitable tool to authenticate hazelnut cultivar and origin.

The aim of the present study is to analyze the unsaponifiable fraction by a fingerprinting approach and chemometrics to distinguish hazelnuts from different geographical origins and cultivars.

In this work the unsaponifiable fraction fingerprint of 267 hazelnut samples from 4 countries and 8 cultivars, produced along 3 harvest years, was analyzed by gas chromatography-mass spectrometry (GC-MS). The chromatographic profiles of 17 specific extracted ions were aligned and used to build two partial least squares discriminant analysis (PLS-DA) classification models, one to discriminate hazelnuts by the geographical origin and another according to the cultivar. The quality of the PLS-DA models was assessed through an external validation with 20% of the samples as test set. To increase the robustness of the validation and to minimize the effect of the sample sets' composition, seven random test sets were evaluated for each model.

The cultivar PLS-DA model, which was a binary model that differentiate Tonda di Giffoni samples from other cultivars, provided a 93% of overall correct classification by external validation. The geographical origin model, that was a multiclass model to distinguish samples according to their country of origin, correctly classified 91% of the samples.

These results show that depending on the variable selected for supervising the pattern recognition analysis, PLS-DA models can successfully classify samples according to their cultivar or country of origin, proving the suitability of the unsaponifiable fraction fingerprint as a hazelnut cultivar and geographical authentication tool.

References

[1] B. Matthaüs, M. M. Özcan, *Eur. J. Lipid Sci. Technol.* **114** (2012) 801–806.
[2] J. S. Amaral, S. Casal, I. Citova, A. Santos, R. M. Seabra, B. P. P. Oliveira, *Eur Food Res Technol.* **222** (2006) 274–280.
[3] K. Król, M. Gantner, *Agriculture.***10** (2020) 375.
[4] M. Özdemir, F. Ackurt, M. Kaplan, M. Yildiz, M. Löker, T. Gürcan, G. Biringen, A. Okay, *Food Chem*. **73** (2001) 411-415.
[5] J. Parcerisa, D. G. Richardson, M. Rafecas, R. Codony, J. Boatella, *J Chromatogr. A.* **805** (1998) 259–268.
[6] G. Lercker, M.T. Rodriguez-Estrada, *J Chromatogr. A.* **881** (2000) 105-129.
[7] S. V. Goriainov, C. A. Esparza, A. R. Borisova, S. V. Orlova, V. V. Vandyshev, Fadi Hajjar, E.A. Platonov, E. P. Chromchenkova, O. O. Novikov, R. S. Borisov, G. A. Kalabin, *J. Anal. Chem.* **76** (2021) 1635-1644.

MULTIVARIATE CONTROL CHART BASED ON PCA/Q RESIDUALS TO EVALUATE *Salmonella* IN MEAT-BONE FLOUR

Thayná K. M. dos Santos[1], Paulo H. Março[1], Patrícia Valderrama[1]
[1]*Universidade Tecnológica Federal do Paraná (UTFPR), Campo Mourão - Paraná, Brazil*
patriciav@utfpr.edu.br / pativalderrama@gmail.com

Meat and bones, considered by-products in the meat industry, are used in flour manufacture for feed production for pets (cats and dogs), chickens, swine, and fish. For quality control, microbiology evaluation concerning *Salmonella* should be done in flours [1]. In this sense, this study aims to develop data-driven decision support tools using near-infrared (NIR) spectroscopy and a multivariate control chart based on principal component analysis (PCA) and Q residuals to verify the presence of *Salmonella* in meat-bone flours.

The Q residuals from PCA are a lack-of-fit statistic that can be used to indicate how well the pre-established model is describing the sample, i.e., the Q statistic shows how well each sample conforms to the model. It measures the residual difference between a sample and its projection into the '*k*' principal components retained in the model [2].

NIR spectra after multiplicative scatter correction and multivariate control chart based on PCA/Q Residuals are presented in Figure 1 and show the potential of coupling NIR and chemometrics to propose a screening method to identify meat-bones flour with/without *Salmonella*.

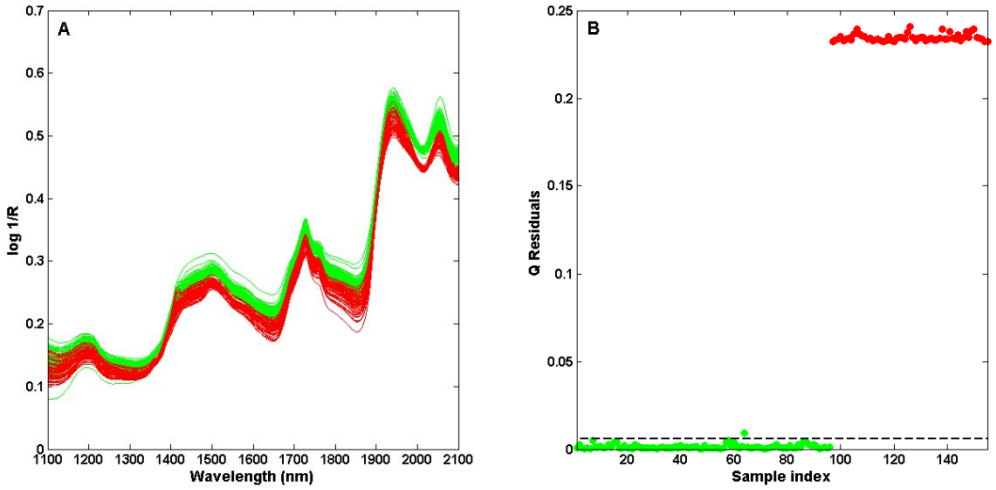


Figure 1 – NIR spectra (A) and Multivariate control chart (B).
Green color = negative for *Salmonella* / Red color = positive for *Salmonella*

NIR spectroscopy coupled with a multivariate control chart based on Q residuals successfully distinguishes meat-bone flours contaminated with *Salmonella* in a nontarget way. From this unsupervised tool, it is possible to implement a criterion for supervision by using Q residuals.

Acknowledgments
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação Araucária.

References
[1] Ministry of Agriculture Livestock and Supply, Brazil, Normative Instruction n °4, January 23, 2007.
[2] L. Valderrama, B. *Demczuk Jr,* P. Valderrama, E. Carasek, *J. Braz. Chem. Soc.* **33** (2022) 401-405.

Fluorescence spectroscopy of wine, a complex food system

H. F. F. Halberg[1,2], Å. Rinnan[2]
[1]*FOSS Analytical A/S, Hillerød, Denmark*
[2]*Department of Food Science, University of Copenhagen, Copenhagen, Denmark*
hehg@foss.dk

Food quality is often not a simple measure, but commonly influenced by several food constituents present in various concentrations. Using wine as an example, process and quality control is commonly ensured by measuring wine parameters (e.g., sugar, acid, and alcohol) by FT-IR. However, these parameters do not describe the wine quality sufficiently. To ensure consistent and high-quality wines, interest also needs to be drawn towards low concentration compounds, such as the phenolics. The phenolics are a large and complex group of compounds associated with different wine quality attributes, such as taste, appearance, mouthfeel, and aging [1,2]. Fluorescence spectroscopy offers a possible solution for rapidly measuring low concentration fluorescent compounds (fluorophores). This is due to the high sensitivity and selectivity [2], achieved by measuring fluorescence as a function of two parameters, excitation and emission, resulting in a fluorescence landscape, known as an excitation-emission matrix (EEM).

Complex food samples, like wine, tend to be heavily affected by inner filter effects (IFEs) caused by the presence of a high concentration of fluorophores or other absorbing species [3]. Hence, one can no longer rely on the intensity of the fluorescence signal to be proportional with the concentration of the fluorophore. Additionally, the presence of many fluorophores increases the chance of overlapping signals in the EEM. The aim of this study is to investigate how to handle these challenges. Wine is here used as example, but these challenges are expected to be common for many food systems and other complex matrices.

A dilution series for wine samples were measured using front-face fluorescence and UV-vis spectroscopy. This study showed both the applicability and limitation of PARAFAC [4] for decomposition of the wine EEMs. The PARAFAC scores proved useful in several aspects of the data analysis: i) to elucidate the occurrence of IFEs, even in front-face fluorescence spectroscopy, and ii) to investigate the optimal dilution factor to avoid IFEs. Comparison with UV-vis spectra was used to obtain an indication of the maximum absorbance that can be tolerated to avoid IFEs in front-face fluorescence spectroscopy. A well-established limit of 0.05 [5] is applied in right-angle fluorescence, but to the best of our knowledge a guideline for the upper absorbance limit is missing for front-face measurements. Additionally, an alternative measure for assessing the PARAFAC model complexity (number of components) was investigated as some of the commonly used measures, like the core-consistency [6], can sometimes indicate too few components, hindering proper analysis of the system.

References

[1] L. Dias-Araujo, W. V. Parr, C. Grose, D. Hedderley, O. Masters, P. A. Kilmartin, D. Valentin, *Food Res. Int.* **149** (2021) 110665.
[2] R. Ferrer-Gallego, J. M. Hernández-Hierro, J. C. Rivas-Gonzalo, M. T. Escribano-Bailón, *Food Res. Int*. **62** (2014) 1100–1107.
[3] M. Bevilacqua, Å. Rinnan, M. N. Lund, *J. Chemom*. **34** (2020) e3286
[4] R. Bro, *Chemom. Intell. Lab. Syst.* **38** (1997) 149-171.
[5] J. Christensen, L. Nørgaard, R. Bro, S. B. Engelsen, *Chem. Rev.* **106** (2006) 1979-1994
[6] R. Bro, H. A, Kiers, *J. Chemom*. **17** (2003) 274-286.

IMAGINE NIR to monitor Pesto sauce industrial production

Daniele Tanzilli[1], Alessandro D'Alessandro[1,2], Lorenzo Strani[1], José M. Amigo[3,4], 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*
[3]*IKERBASQUE, Basque Foundation for the Promotion of Science, Plaza Euskadi 5, P.O. Box 48009, Bilbao, Spain*
[4]*Department of Analytical Chemistry, University of the Basque Country UPV/EHU, P.O. Box 48080 Bilbao, Basque Country, Spain*
daniele.tanzilli@unimore.it

Chemometrics in the industry is increasingly accepted thanks to better handling of process sensors and data. Hence, it is possible to have a better understanding and more efficient industrial process monitoring. This guarantees a constant product quality and consequently a more friendly environment production, thanks to waste reduction.

The case study here concerns the industrial production of Pesto sauce in Barilla food company. Pesto sauce is one of the most popular Italian foods, and it is composed of basil, garlic, extra-virgin olive oil, parmesan cheese and other minor ingredients. In order to achieve good quality control, Barilla has several sensors installed in the plant. The basil is the main ingredient, which mostly influences the final quality. For this reason, it undergoes an accurate control before entering the line with an off-line quality control analysis and real-time control with an RGB Vision System. The intermediate product, a mix of basil, oil and salt, is monitored by an on-line NIR probe. In the end, the Pesto product quality attributes are evaluated by off-line laboratory analysis.

In this work, we evaluated the potentiality of imaging systems to furnish information that could then be integrated with on-line NIR. As a first step, the extraction of features from RGB basil images, such as the percentage of area covered by leaves, branches, and stems respectively, as well as of dark spots eventually present on the leaves, has been considered. These parameters can enter multiblock models to achieve a real-time prediction of quality attributes together with on-line NIR data. In addition, a feasibility study to monitor basil and pesto through multispectral vis-NIR and hyperspectral NIR imaging has been undertaken.

Acknowledgments:

Sensorfint for the Short-Term Scientific Mission grant that allowed me to do the studies on images at University of the Basque Country UPV/EHU

GENOTYPING AND STATISTICAL ANALYSIS OF MARZIPAN WITH DMAS-QPCR

L. F. Voges[1], N. Wax[2], M. Fischer[1,2], S. Seifert[1,2]
[1]*Centre for the Study of Manuscript Cultures (CSMC), University of Hamburg, Warburgstraße 26, 20354, Hamburg, Germany*
[2]*Hamburg School of Food Science - Institute of Food Chemistry, University of Hamburg, Grindelallee 117, 20146, Hamburg, Germany*
lucas.voges@uni-hamburg.de

According to German food guidelines, marzipan raw paste should only contain sweet almonds (*Prunus dulcis*) sugar and water. The proportion of bitter almonds in marzipan is limited to a maximum of 12 wt% while the use of debittered bitter almonds is completely prohibited [1]. However, currently no analytical technique can reliably monitor the compliance with this legal requirement [2].

To detect bitter almond admixtures, we used a genotyping approach based on double-mismatch allele-specific qPCR (DMAS-qPCR) [3]. Since the bitterness of almonds is determined by the maternal genotype [4], we use two prior detected SNPs in *rpoB* and *rps4* genes of the maternal inherited plastid DNA. However, it was found that the bitter genotype of samples from Morocco and Iran could not be observed in samples from Spain or Syria. Assuming that the samples are characterized by the bitter genotype, a regression model was developed to quantify the exact amount of bitter almond admixtures. In addition, statistical tests were applied to verify significant bitter almond admixtures in homogenate samples and to show that exceeding the 12% threshold can also be detected in marzipan. This makes this method very promising for marzipan authentication, especially because debittered bitter almonds can be detected.

In Figure 1 the presented approach is summarized graphically. The methods developed for the statistical analysis are published in the R package *DoubleqpcR* on GitHub (github.com/LucasFVoges/DoubleqpcR).

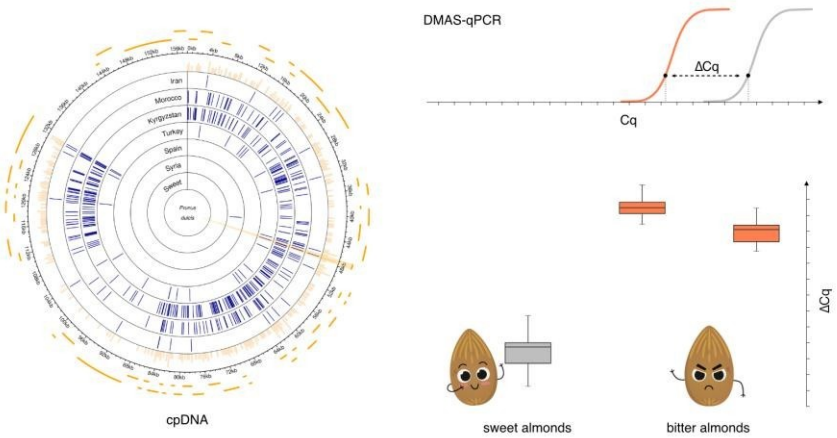


Figure 1 – Overview of the main results obtained in the development of a method for identifying bitter almonds in marzipan. SNP calling with the plastid (cpDNA) reference genome from *Prunus dulcis* showed diverse variations across the genome for bitter almonds from Morocco, Iran, Kyrgyzstan, and Turkey. DMAS-qPCR was used to determine the difference (ΔCq) between amplification of the sweet and bitter genotypes for a selected SNP at the *rps4* locus. The difference is clearly visible in the box plots for sweet almond and bitter almond homogenates from Morocco and Iran.

References

[1] Leitsätze für Ölsamen und daraus hergestellte Massen und Süßwaren *BMEL* (2010).
[2] S. Vichi et al. *Foods* **9** (2020) 747.
[3] S. Lefever et al. *Sci. Rep.* **9** (2019) 2150.
[4] J. Del Cueto et al. *Plant Physiol. Biochem.* **126** (2018) 163–172.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

Alternative approaches to untargeted LC/GC-MS data analysis

Andrea Jr Carnoli[1], Gerjen H. Tinnevelt[1], Francisco Souza[1], Petra Oude Lohuis[2], Gerke Spaling[2], Christina Precht[3], Arend Heerschap[4], Geert Postma[1],Jeroen J. Jansen[1]

[1]*Radboud University, Institute for Molecules and Materials (IMM), The Netherlands;*
[2] *Teijin Aramid, Arnhem, The Netherlands;*
[3] *Vetsuisse Faculty University of Bern, Bern;*
[4]*Radboud University Medical Center, Nijmegen, The Netherlands*
andrea.carnoli@ru.nl

Untargeted LC/GC-MS data are difficult to analyze compared to targeted MS analysis due to data analytical challenges, such as high data dimensionality, data collinearity and data artefacts (also caused by high data variety). Therefore, chemometrics is required to retrieve information regarding similarity among samples and feature importance. For untargeted MS analysis, the response within the measurement may be non-linear with respect to the concentration of the molecules/ions/fragments. Therefore, we explore the application of CMIM feature selection [1] and a combined qualitative [2] and quantitative information analysis [3] to effectively extract and interpret untargeted LC/GC-MS data.

References

[1] F. Fleuret, *J. Mach. Learn. Res.* **5** (2004) 1531-1555.
[2] Y. Song, J. A. Westerhuis, N. Aben, M. Michaut, L. F. A. Wessels, A. G. Smilde, *Brief. Bioinformatics*. **20** (2019) 1-13.
[3] G. Michailidis, J. de Leeuw, *Statist. Sci.* **13** (1998) 307-336.

Acknowledgments

This research was founded by Teijin Aramid

Analysis of Pinot Noir Wines Using UV-Vis Spectroscopy

Cannon Giglio[1], Paul A. Kilmartin[1], Yi Yang[1]
[1]*School of Chemical Science, University of Auckland, Auckland, New Zealand*
Cgig597@aucklanduni.ac.nz

In commercial wines, a wide variety of compounds can influence the perceived taste of the finished product. As a result, winemakers are keenly interested in monitoring the composition of their wines to gain an indication of wine style and even wine quality. Of particular interest is the phenolic family of molecules, including tannins and anthocyanins.

UV-Vis spectrometry has been widely used in wine analysis, and various simple calibrations have been developed, such as using the absorbance at 280 nm to model the total phenolic content, and 520 nm for total anthocyanins. Further methods available include the methyl cellulose precipitation (MCP) assay for tannin content [1]. A major downside of these chemical assays is that they rely on univariate calibration, which cannot account for interfering species and require full selectivity [2]. As a result, these assays are not always fully reproducible, and are rarely validated against more complete phenolic analyses as provided by HPLC. To our knowledge there have only been a dozen or so papers using chemometrics along with UV-Vis spectrometry for wine analysis [3].

106 Pinot noir wines from New Zealand were measured using a UV-Vis spectrophotometer, from 190-600 nm, as shown in Figure 1. Various reference analyses were undertaken using HPLC and several common assays (MCP, Folin-Ciocalteu, and Somers colour indices). PLS calibration models were computed. The results were particularly strong (r^2_{CV} >0.9, %RMSEP<10%) for prediction of total phenolics, tannins, malvidin-3-glucoside, and caftaric acid. The regions of 215-240 nm and 260-290 nm were strongly correlated with tannins and total phenolics, which agree with findings reported by Boulet et al. [4].

An additional external validation set of around 50 wines is being used to assess the robustness of the models. The use of sparse modeling, as well as scaling and normalization, is also being evaluated. Overall, full-spectrum calibration has strong potential as a simple, inexpensive, and accurate method for determining a wide variety of phenolic compounds.

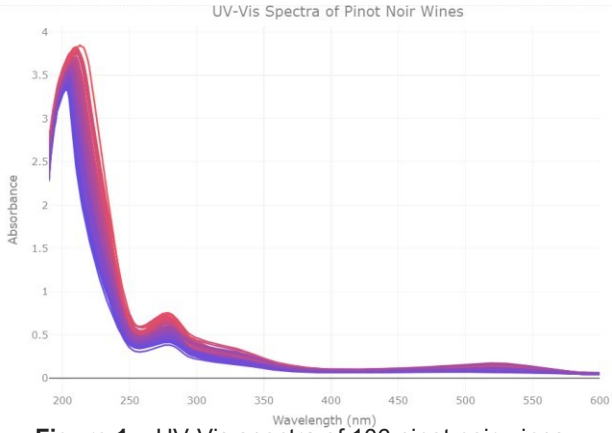


Figure 1 – UV-Vis spectra of 106 pinot noir wines.

References

[1] Sarneckis, C. J.; Dambergs, R.; Jones, P; Mercurio, M; Herderich, M. J.; Smith, P. *Australian Journal of Grape and Wine Research* **2006**, *12*, 39–49.
[2] Booksh, K. S.; Kowalski, B. R. *Anal. Chem.* **1994**, *66* (15), 782–791.
[3] Aleixandre-Tudo, J. L.; du Toit, W. The Role of UV-Visible Spectroscopy for Phenolic Compounds Quantification in Winemaking. In *Frontiers and New Trends in the Science of Fermented Food and Beverages*; Lidia Solís-Oviedo, R., de la Cruz Pech-Canul, Á., Eds.; IntechOpen, **2019**.
[4] Boulet, J. C.; Ducasse, M. A.; Cheynier, V. *Aust. J. Grape Wine Res.* **2017**, *23* (2), 193–199.

EVALUATION OF THE ACCURACY OF NMR PREDICTORS FOR THE PREDICTION OF FATTY ACID SPECTRA

M.N. Chhaganlal[1] and S.A. Mjøs[1],
[1]*Department of Chemistry, University of Bergen, Norway*
milan.chhaganlal@uib.no

Four different NMR predictors have been evaluated for their ability to predict the 1H-NMR spectra of fatty acids and fatty acid methyl esters. The four predictors were three commercial softwares from Chemaxon (www.chemaxon.com), ACD/Labs (www.acdlabs.com) and Mestrelab Research (www.mestrelab.com), and one free online service from NMRdb (www.nmrdb.org)
1H-NMR spectra of 20 common free fatty acids and their corresponding fatty acid methyl esters were acquired on a 600 MHz AVANCE NEO instrument (Bruker). The spectra were collected at 298 K using a spectral width of 30 ppm, a size of FID of 128k, zero dummy scans and eight scans. Experimental and predicted spectra were exported to MATLAB (www.mathworks.com) and resampled to a common scale and resolution before further processing and evaluation.
The NMR predictors were evaluated on two main criteria: 1) their accuracy in direct prediction of the spectra, and 2) the ability to predict the change that occur in the spectra by introduction of a double bond in the fatty acid carbon chain or by esterification of the free fatty acids to methyl esters.
Since the most important information in NMR is where signals occur (chemical shifts), common methods for spectral comparisons, like spectral contrast angle and correlation are unsuitable for NMR spectra. A metrics based of on the position of quantiles of the signals from the functional groups was therefore applied.
The results showed that the predictors from ACD/Labs performed better that the other in the direct prediction (Criterion 1), with lower median deviations and fewer outliers. But there was no clear winner when it came to the ability of predicting how the spectrum change as a result of a change in the molecular structure (Criterion 2).

An IDEL perspective on handling spatial correlation in hyperspectral imaging

M. Ahmad[1, 2], J.M. Amigo[3, 4], 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.
[3] Ikerbasque, Basque Foundation for Science, María Díaz de Haro, 48013 Bilbao, Spain
[4] Department of Analytical Chemistry, University of the Basque Country UPV/EHU, P.O. Box 644, 15 48080 Bilbao, Basque Country, Spain
m.ahmad@live.nl

Hyperspectral imaging is used in many fields of science, for its powerful ability to separate information in both the spatial and spectral domain. Applications range from cell imaging in biology to remote sensing in agricultural sciences. However, in a lot of cases, the spatial correlation is completely disregarded in chemometric analyses, because the spatial dimension is usually unfolded pixelwise to provide a two-dimensional data matrix. These data are then decomposed by workhorse algorithms such as Principal Component Analysis (PCA) or Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS). A very clear example is NIR imaging which is a chemical imaging technique based on reflectance spectroscopy. There is significant spatial-spectral correlation in NIR images, especially coming from the scattering of light that for complex samples (characterized by localized spatial non homogeneities such as morphological/textural variations, object edges or fibres) introduces a non-linear interplay between the spatial and spectral dimension. Usual pre-processing approaches such as multiplicative scatter correction (MSC) that ignore spatial correlations cannot fully correct the measured spectra and if they would, then some information related to the physical nature of the sample will be lost as they assume the spatial variance to be insignificant. We previously investigated a case study of a semen droplet on cotton fabric, analysed with NIR imaging [1]. A novel methodology was utilized, that takes a spatial perspective on data analysis while maintaining spectral relevance, i.e., Image Decomposition, Encoding and Localisation (IDEL) [2]. The data were dominated by scattering effects, however without removing the scattering contributions, the relevant spatial structures and spectral signature were extracted. With this work, we would like to introduce two different problems and how these might be approached with IDEL. We will start with a dataset that has two physically different components, i.e., fabric and stamped flowers [3]. The components show a very similar spectral absorbance and overlap completely in the spatial dimensions. Although the flower pattern is highly obscured by the fabric pattern, the flower shape can still be recovered at distinct spectral channels, namely where the fabric pattern absorbs the least. The second problem is the analysis of bread data [4]. A slice of bread was analysed over six different time points to investigate the staling process. The data are dominated by scattering effects, however, in this work we will show that regardless of scattering, the underlying information within the data can still be extracted. In this case, it is the drying effect that is visualised across the six data sets and highlighted within the results.

The spatial dimension in hyperspectral imaging should not be disregarded without careful consideration of the information that is within the spatial domain. Doing so might either alter or obscure information within the data that is being analysed.

References

[1] Silva C.S.; Pimentel M.F.; Amigo J.M.; Honorato R.S.; Pasquini C.; Tr. in Anal. Chem. 2017, 95, p: 23-35.
[2] Ahmad M.; Vitale R.; Silva C.S.; Ruckebusch C.; Cocchi M.; Analytica Chimica Acta, Volume 1191, 2022, 339285
[3] Khan H.A.; Mihoubi S.; Mathon B.; Thomas J.B.; Hardeberg J.Y.; Sensors, Volume 18, 2045
[4] Amigo J.M.; Alvarez A.d.O.; Engelsen M.M.; Lundkvist H.; Engelsen S.B.; Food Chemistry Volume 208, 2016, 318-325

IDENTIFICATION OF SPECTRAL PATTERNS ASSOCIATED TO DIFFERENT AGGREGATION STATES OF BETA AMYLOID PEPTIDE IN HYPERSPECTRAL IMAGES THROUGH CHEMOMETRIC ANALYSIS

J. Araya[1-2], G. Alvez[1], M. Tiznado-Fariña[1], F. Fuentes-Rabanal[1] D. Mennickent[1-2], A. Rodríguez[2-3], E. Guzmán-Gutiérrez[1-2], J. Fuentealba[1]
[1]Universidad de Concepción, Concepción, Chile
[2]Machine Learning Applied in Biomedicine (MLAB), Chile
[3]Universidad del Bío-Bío, Chillán, Chile.
jarayag@udec.cl

Alzheimer's Disease (AD) is a neurodegenerative disorder that causes deterioration of cognitive functions and loss of memory. Its origin is still unclear, however, it is known that the accumulation of the amyloid beta peptide (Aβ) has an important role due to its neurotoxic effect. The toxicity of Aβ depends on the aggregation state in which the peptide is found, which in turn occurs along with changes in its secondary structure [1,2]. Furthermore, Aβ could be found in different aggregation states at once in different parts of the brain, and since only some of these states are toxic, to understand the real impact of this peptide in AD is challenging.
FT-IR micro-imaging is an ideal tool to study this peptide since it can combine the chemical selectivity of infrared spectroscopy, which is sensitive to secondary structure changes, as well as the spatial resolution of microscopy [3, 4].
The study of hyperspectral images combined with chemometrics analysis allows the extraction of crucial information of the surface in study, enabling the identification of spectral patterns associated with structural changes of Aβ and the localization of said patterns in an image [5].
In this work, we obtained hyperspectral images of the Aβ peptide using attenuated total reflectance - Fourier transform infrared (ATR-FTIR). This technique was coupled with multivariate methods, i.e. Principal Component Analysis (PCA) and Multivariate Curve Resolution with Alternating Least Squares (MCR-ALS) to find spectral patterns associated to different aggregate states of the Aβ peptide and evaluate the distribution of those patterns in the image.
We analyzed ATR-FTIR images of aggregated and non-aggregated Aβ samples which were also corroborated by Thioflavin-T binding assays as well as Western Blot, silver staining and perforated whole-cell patch-clamp as functional physiological control assay. The spectral patterns were found to be coherent with the expected structural changes of the aggregate states of Aβ. Hippocampal lysates samples from J20 mice, a known AD mouse model that overexpresses Aβ, were also analyzed by FT-IR micro-imaging and compared to wildtype mouse brains. Using the spectral patterns resolved by MCR-ALS, it was possible to selectively localize different aggregate states of the Aβ peptide on an image.
This novel analytical platform could allow a selective mapping of the distribution of different Aβ species in a brain with AD, which could provide a better understanding of the pathophysiology of the disease.

References

[1] Glabe, C. G., (2008)., *J. Biol. Chem.*, Structural classification of toxic amyloid oligomers, **283(44)** 29639-29643.
[2] Klementieva, O., Willén, K., Martinsson, I., Israelsson, B., Engdahl, A., Cladera, J., Uvdal, P. & Gouras, G. K. (2017)., *Nat. Commun.,* Pre-plaque conformational changes in Alzheimer's disease-linked Aβ and APP, **8** 14726.
[3] Miller, L. M., (2013)., *BBA–Biomembr.,* FTIR spectroscopic imaging of protein aggregation in living cells., ***1828(10)*** 2339-2346.
[4] Röhr, D., Boon, B. D., Schuler, M., Kremer, K., Hoozemans, J. J., Bouwman, F. H., ... & Gerwert, K. (2020)., *Acta Neuropathol. Commun.,* Label-free vibrational imaging of different Aβ plaque types in Alzheimer's disease reveals sequential events in plaque development, ***8*****(1)** 1-13.
[5] Lewis, E. N., (1995)., *Anal. Chem.,* Fourier transform spectroscopic imaging using an infrared focal-plane array detector., **67(19)** 3377-3381.

SUPERVISED PATTERN RECOGNITION USING NEAR INFRARED SPECTRUM OF SERUM FOR DIAGNOSIS OF GESTATIONAL DIABETES MELLITUS

J. Araya[1-2], D. Mennickent[1-2], A. Rodríguez[2-3], E. Guzmán-Gutiérrez[1-2]
[1]*Universidad de Concepción, Concepción, Chile*
[2]*Machine Learning Applied in Biomedicine (MLAB), Chile*
[3]*Universidad del Bío-Bío, Chillán, Chile.*
jarayaq@udec.cl

Gestational diabetes mellitus (GDM) is a hyperglycemia state that is diagnosed during the second or third trimester of pregnancy, typically by an oral glucose tolerance test (OGTT) [1]. The OGTT is unpleasant, time-consuming and has low reproducibility [2]; moreover, by the time of its use, the fetal phenotype is already altered in GDM pregnancies [3,4,5]. Hence, GDM diagnosis can be improved. Predictive models that combine vibrational spectroscopy with chemometrics classification techniques are an auspicious means to achieve that goal, either as early detection tools or as alternative screening tools [6].

The aim of this study was to develop and evaluate near-infrared (NIR)-based chemometrics models for the prediction of GDM in Chilean pregnant women. Pregnant women with ≤ 12 gestational weeks and without pregestational diabetes were recruited in Concepcion, Chile. GDM diagnosis was performed at 24-28 gestational weeks, with fasting glycemia 100-125 mg/dL or post load glycemia (75 g, 2 h) ≥ 140 mg/dL. Sera were collected during the first ($n_{Control}$ = 71, n_{GDM} = 15) and second ($n_{Control}$ = 40, n_{GDM} = 8) trimester of pregnancy. For each serum sample, 5 NIR spectra (range 10500-4000 cm^{-1}, resolution 4 cm^{-1}) were recorded and averaged. Besides the full NIR wavenumber range, 3 shorter spectral regions were assessed: 10500 to 7600 cm^{-1}, 7600 to 5100 cm^{-1} and 5100-4000 cm^{-1}. For each spectral range, 80 different combinations of transformations (Savitzky-Golay smoothing or first/second derivative with varying filter width, standard normal variate scattering correction, automatic weighted least squares baseline correction, 2-norm normalization) were tested. Data were preprocessed by mean center. For GDM prediction, the chemometrics classification technique partial least squares-discriminant analysis (PLS-DA) was employed. The models' performance was evaluated by their area under the receiver operating characteristic curve (AUC) in calibration and leave-one-out cross-validation.
The best first trimester PLS-DA model was obtained with the 10500-7600 cm^{-1} spectral region, 2-norm normalization and mean center. It got a cross-validation AUC (CV-AUC) of 0.7070. The best second trimester PLS-DA model was obtained with the 5100-4000 cm^{-1} spectral region, Savitzky-Golay first derivative (polynomial order = 2, filter width = 15) and mean center. It achieved a CV-AUC of 0.9313.
The developed models are able to classify GDM and control women with a moderate predictive power in the first trimester, and an excellent performance in the second trimester. Therefore, the latter may be helpful to improve GDM diagnosis, as an alternative screening tool.

References:

[1] I. Tsakiridis, S. Giouleka, A. Mamopoulos, A. Kourtis, A. Athanasiadis, D. Filopoulou, T. Dagklis. Obstet Gynecol Surv. **76** (2021) 367–381.
[2] E.A. Huhn, S.W. Rossi, I. Hoesli, C.S. Göbl. Front Endocrinol. **9 (**2018) 696.
[3] U. Sovio, H.R. Murphy, G.C.S. Smith. Diabetes Care. 39 (2016) 982–987.
[4] H. Venkataraman, U. Ram, S. Craik, A. Arungunasekaran, S. Seshadri, P. Saravanan. Diabetologia.**60** (2017) 399–405.
[5] L. Yovera, M. Zaharia, T. Jachymski, O. Velicu-Scraba, C. Coronel, C. de Paco Matallana, G. Georgiopoulos, K.H. Nicolaides, M. Charakida. Ultrasound Obstet Gynecol. **57** (2021) 607–613.
[6] O. Richards, C. Jenkins, H. Griffiths, E. Paczkowska, P.R. Dunstan, S. Jones, M. Morgan, T. Thomas, J. Bowden, A. Nakimuli, M. Nair, C.A. Thornton. Front Glob Women's Health. **1** (2021) 610582.

SOIL SPECTROSCOPY: USE OF CHEMOMETRICS FOR FINE-TUNING SPECTRA ACQUISITION- CASE OF SCANS NUMBER OPTIMIZATION

I. Barra[1], L. Khiari[2], S. Haefele[3], R. Sakrabani[4], F. Kebede[1]

[1]*Center of Excellence in Soil and Fertilizer Research in Africa (CESFRA) Mohammed VI Polytechnic University (UM6P), Benguerir, Morocco*
[2]*Department of Soils and Food Engineering, Faculty of Agriculture and Food Sciences, Laval University, Quebec, Canada*
[3]*Department of Sustainable Agriculture Sciences, Rothamsted Research, Harpenden, UK*
[4]*School of Water, Energy and Environment, Cranfield University, UK*
Issam.barra@um6p.ma

Vibrational spectroscopy such as Fourier-transform infrared (FTIR), has been used successfully for soil diagnosis owing to its low cost, minimal sample preparation, non-destructive nature. The present study aimed at facilitating the work of soil spectroscopist by optimizing one of the essential settings during the acquisition of FTIR spectra (viz. Scans number) using the standardized moment distance index (SMDI) as a metric that could trap the fine points of the curve and extract optimal spectral fingerprints of the sample. Furthermore, it can be used successfully to assess the spectra resemblance. The study revealed that beyond 50 scans the similarity of the acquisitions has been remarkably improved [1].

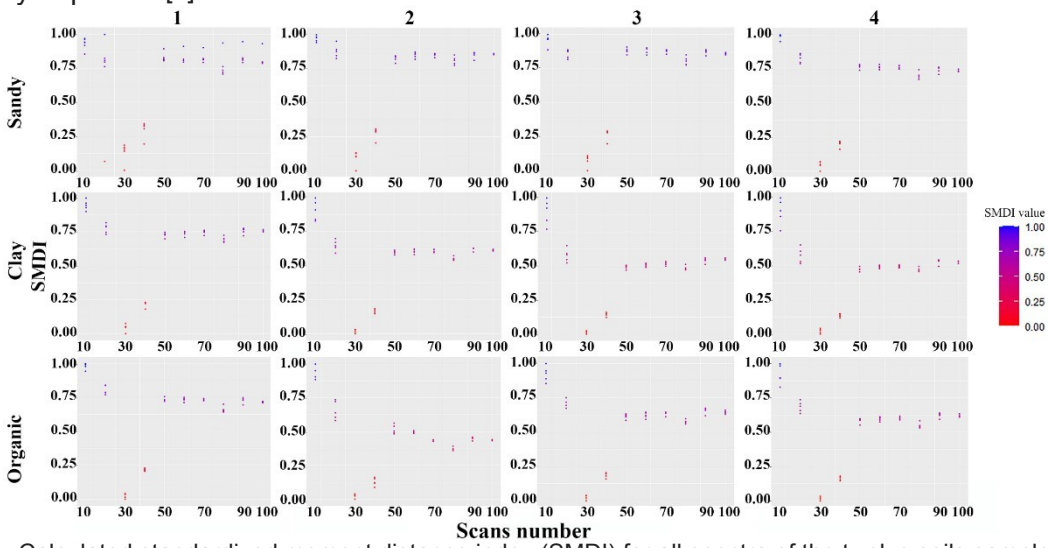


Figure 1 – Calculated standardized moment distance index (SMDI) for all spectra of the twelve soils samples resulting from different scan numbers. The points' (SMDI) rapprochement indicates the improvement of the spectra's similarity

Subsequently, the effect of the number of scans on the predictive ability of the partial least squares regression models for the estimation of five selected soil properties (i.e., water pH, soil organic carbon, total nitrogen, cation exchange capacity and Olsen phosphorus) was assessed, and the results showed a general tendency in improving the correlation coefficient (R^2) as the number of scans increased from 10 to 80. In contrast, the cross-validation error RMSECV decreased with increasing scan number, reflecting an improvement of the predictive quality of the calibrated models with an increasing number of scans.

The final finding of the present study showed that the number of scans has a remarkable effect on spectral stability and represents an important parameter to be taken into consideration when recording FTIR spectra of soil samples for the set-up of predictive models in soil spectroscopy. On the other, the chemometric methodology used as part of this study showed that the recorded spectra's quality (stability and repeatability) was improved by increasing the number of scans.

References

[1] I. Barra, L. Khiari, S.M. Haefele, R. Sakrabani, F. Kebede, Sci. Rep. 11 (2021) 13358.

Classification of Horsetails using Machine Learning Methods on NIR Spectra

K. Beier[1], T.-M. Dutschmann[1], P. M. Puttich[2], M. Lubienski[2], T. Beuerle[2], K. Baumann[1]
[1]Institute of Medicinal and Pharmaceutical Chemistry, TU Braunschweig, Beethovenstr. 55, 38106 Braunschweig, Germany
[2]Institute of Pharmaceutical Biology, TU Braunschweig, Mendelssohnstr. 1, 38106
Katharina.beier@tu-bs.de

Horsetail (*Equisetum arvense L.*), which is native in the northern hemisphere, holds a long tradition in the supportive treatment of numerous diseases [1]. E.g., it is applied externally to support wound healing. Furthermore, due to its diuretic effect, the treatment of infections of the bladder and urinary tract benefits from tea made from dried horsetail [2]. However, a common problem is the confusion with marsh horsetail (*Equisetum palustre*) due to its morphological similarity to *E. arvense*, growing in the same habitats. *E. palustre* contains palustrine, an alkaloid which is potentially toxic [3,4].
To differentiate the desired *E. arvense* from the toxic *E. plaustre*, near-infrared (NIR) spectroscopy, also commonly used in food quality control, can be applied. For this study, over 350 *E. arvense* and *E. palustre* samples originating from all over Germany, consisting of 3 years of harvest, were analyzed using a general-purpose handheld NIR device. After applying dimensionality reduction and clustering techniques (PCA, t-SNE) to the spectroscopic data, a variety of machine learning algorithms (kNN, SVM, RF) were trained to classify the species from the spectrum. In a cross-validation approach, it could be shown that the spectra are sufficient to achieve high classification accuracies. Additionally, the data allow for determining the harvesting season as well. The success of the complete workflow will be further emphasized by assessing its reliability through posterior probabilities for the predicted class labels.

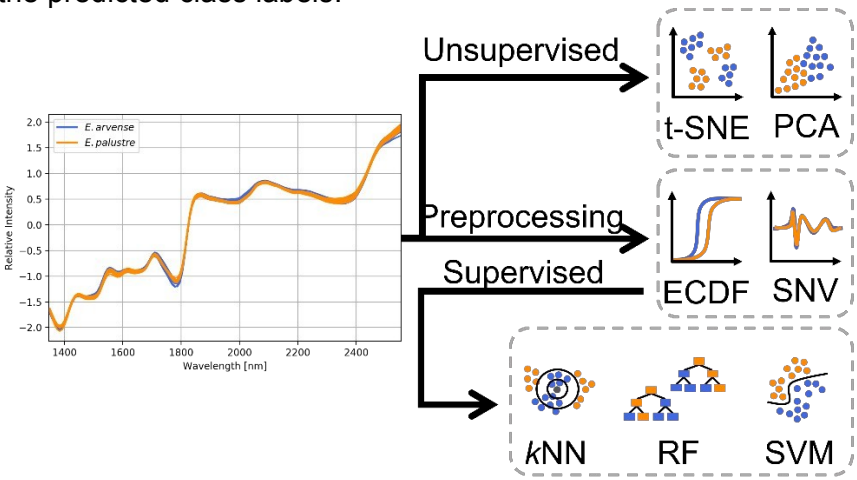


Figure 1 –Workflow from NIR Spectra to Classification.

References

[1] A.-E. Al-Snafi, *IOSR Journal Of Pharmacy*, 7 (2017) 31-42.
[2] A. Waterstradt, M. Winker, A. M. Zimmermann-Klemd, S. Devi, A.-K. Lederer, R. Huber, C. Gründemann, Planta Med, (2021)
[3] M. Nowak, I. Tipke, L. Bücker, K. Franke, M. Lubienski, T. Beuerle, *Planta Med*, **88** (2022) 447-454.
[4] J. Müller, P. M. Puttich, T. Beuerle, *toxins*, **12** (2020), 710-724.

TIME-BASED COLORIMETRIC METHOD FOR THE SIMULTANEOUS DETERMINATION OF CALCIUM AND MAGNESIUM IONS WITH SILVER NANOPARTICLES

I. Berasarte, A. Bordagaray, R. Garcia-Arrona, M. Ostra, M. Vidal[1]
[1]*Department of Applied Chemistry, Faculty of Chemistry, University of the Basque Country (UPV/EHU), Donostia/San Sebastian, Spain*
irati.berasarte@ehu.eus

Over the past decade, the use of metal nanoparticles has increased exponentially. The most studied nanoparticles have been gold and silver NPs because of their unique optical, electrical, and photothermal properties. Noble metal NPs are able to produce quantum effects, a key parameter for naked-eye colorimetric sensing applications, as modifications of their surface charge result in visible color change, which can be rapidly and easily detected. [1].

With that aim, silver and gold nanoparticles have been widely used for the colorimetric determination of different heavy metal ions, such as, Ni^{2+}, Hg^{2+} or Cu^{2+}, as well as other smaller ions as Ca^{2+} and Mg^{2+}. These last two have been mostly detected using gold nanoparticles. However, AuNPs are much more expensive to buy or to synthesize than AgNPs. Taking that into account, simultaneous determination of Ca^{2+} and Mg^{2+} with AgNPs was investigated. Calcium has been previously used for the determination of cysteine, as AgNPs aggregate and undergo a color change from yellow to orange or red in the presence of cysteine and calcium [2,3].

In the present work, AgNPs – amino acid system was used for the colorimetric determination of calcium and magnesium ions. Firstly, different amino acids were analysed in order to find the most selective one. Afterwards, a concentration matrix of 4x4 including both ions was measured by UV-Visible spectrophotometry and Digital Image Analysis (DIA). In the case of UV-Vis, the reaction was carried out for 10 minutes and spectra were recorded every 15 seconds. As for DIA, a 10 minute video was recorded using a smartphone and time variable selection was done in order to optimize obtained models.

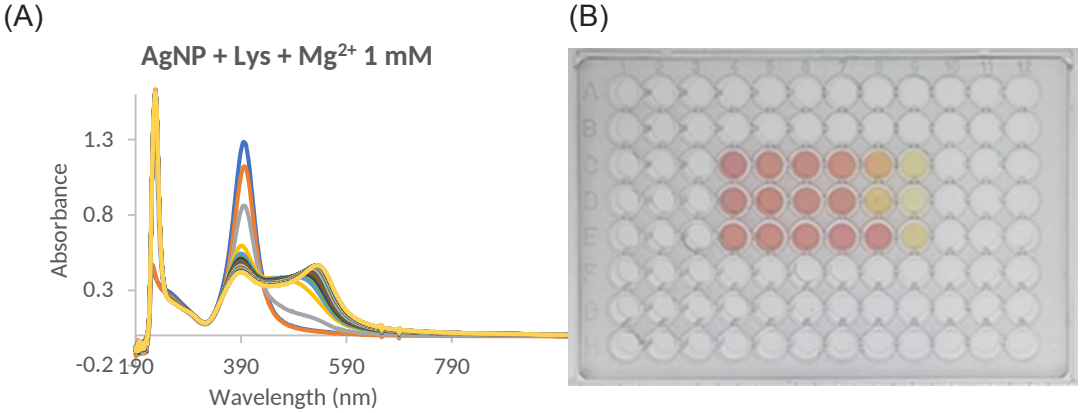


Figure 1 – (A) Example of a kinetic reaction between silver nanoparticles, Lysine amino acid and magnesium 1 mM measured with UV-Vis. (B) Photogram of the reaction recorded with the smartphone.

Taking into account the kinetic nature of the reaction and the complexity of mixtures, multivariate analysis methods were used. Firstly, kinetic profiles were unfolded into two-way data using Matlab®. After, and using the PLS-Toolbox, different pre-processing techniques were analyzed in order to find the best way to treat the data. Then, Principal Component Analysis (PCA) was used for exploratory analysis, and finally, Partial Least Squares Regression (PLS-R) for the calculation of the calibration models for both Ca^{2+} and Mg^{2+}.

References

[1] M. Sabela, S. Balme, M. Bechelany, J.M. Janot, K. Bisetty, *Adv. Eng. Mater.* **19** (2017) 1700270.
[2] S. Hajizadeh, K. Farhadi, M. Forough, R. Molaei, *Anal. Methods.* **4** (2012) 1747-1752.
[3] C. Han, K. Xu, Q. Liu, X. Liu, J. Li, *Sens. Actuators B Chem.* **202** (2014) 574-582.

Investigation of an innovative method for classifying nanostructures based on time series analysis and fuzzy logic in microscopic images

Hooryeh Borhani 1, Sori Haghgoo 2, Keivan Nik Rava 3

1. Department of Physics, Faculty of Science, Azad University, Mashhad, Iran
2. Department of Photonics, Faculty of Sciences and New Technologies, Graduate University of Advanced Technology, Tehran, Iran
3. Department of Physics, Faculty of Science, Azad University, Kordestan, Iran

Sorihaghgoo2010@yahoo.com

The dispersion of nanoparticles in nanostructures is one of the most important indicators that are used to confirm the efficiency of the proposed methods in the synthesis of nanomaterials. Scanning electron microscopic images of nanoparticles have high-resolution structural, chemical and morphological information at the nanometer-scale nanomaterials. In this paper, a new algorithm for classifying nanostructures using these images is presented. For this purpose, first, scanning electron microscopic images of nanoparticles were converted to time series and their characteristics were investigated through time series analysis methods. Then the statistical properties of these series were extracted. The extracted properties were considered as inputs of a fuzzy inference system for classifying microscopic images of nanostructures into three groups: good, medium and bad. This algorithm is applied to 65 microscopic images of nanoparticles with the same dimensions (250 × 250 pixels) and has an accuracy of more than 90%, which is very convenient.

References

1. Veeramsetty V, Deshmukh RJSAS. Electric power load forecasting on a 33/11 kV substation using artificial neural networks. 2020;2(5):1-10.
2. Ding N, Bésanger Y, Wurtz FJEPSR. Next-day MV/LV substation load forecaster using time series method. 2015;119:345-54.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

MULTIVARIATE MODELS FOR PREDICTION QUALITY PARAMETRS OF BERRY BEVERAGES USING FTIR-ATR SPECTROSCOPY

E. Sikorska[1], K. Włodarska[1], K. Pawlak-Lemańska[1], M. Sikorski[2]
[1]*Department of Technology and Instrumental Analysis, Institute of Quality Science, Poznań University of Economics and Business, Poznań, Poland*
[2] *Faculty of Chemistry, Department of Spectroscopy and Magnetism, Adam Mickiewicz University in Poznań, Poznań, Poland*
ewa.sikorska@ue.poznan.pl

The beverages produced from the berry fruit are characterized by high nutritional quality and attractive flavor. They have a high content of macro- and micronutrients, high amounts of dietary fiber, vitamins A, C, E and vitamins of the B group, and phenolic compounds. Therefore, the consumption of red fruit beverages may be an important element of a healthy diet.
The objective of the present study was to explore the feasibility of FTIR-ATR spectroscopy to determine the soluble solids content (SSC), the titratable acidity (TA), and the total polyphenols content (TPC) of red fruit juice.
A beverages commercially available on the Polish market were studied including juices, nectars and syrups produced from blackcurrant (*Ribes nigrum*), chokeberry (*Aronia melanocarpa*), strawberry (*Fragaria* × *ananassa*) and raspberry (*Rubus idaeus*). Juice spectra were measured using a FTIR spectrometer with ATR accessory in the range of 4000-600 cm^{-1}. The soluble solids content of the juices was determined refractometrically. The titratable acidity was determined by means of potentiometric titration. The total phenolic compounds content was determined using Folin-Ciocalteu phenol reagent. Partial least squares (PLS) regression and support vector machine (SVM) regression were used to develop the calibration model between the juices spectra (X matrix) and the chemical parameters (Y matrix). To optimize the models, the spectra were pre-processed using first- and second-order derivatives, multiplicative scatter correction (MSC), and standard normal variate (SNV). The selection of variables was carried out using the forward variant of the interval PLS method (iPLS).

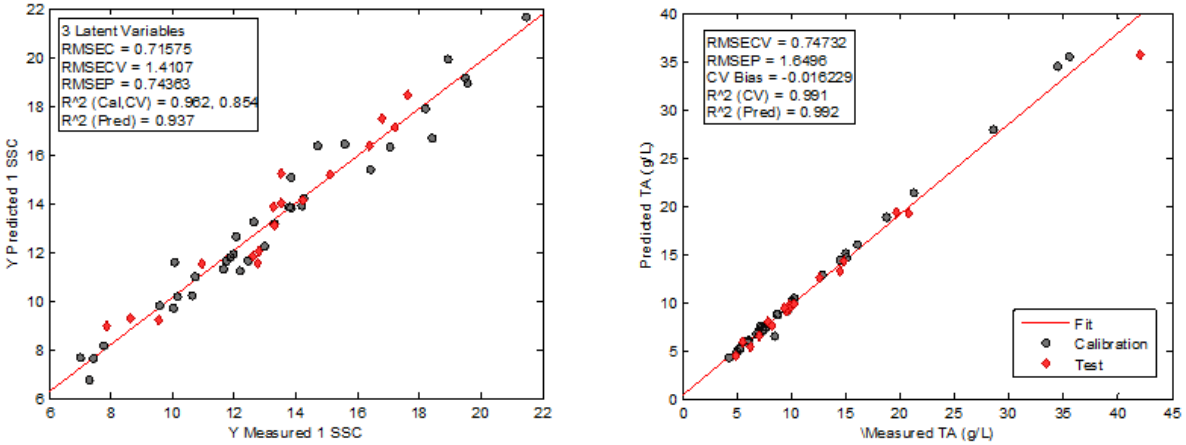


Figure 1. Partial least squares regression models for prediction of soluble solid content and titratable acidity of berry juices

Both PLS and SVM showed similar performance for the prediction quality parameters of beverages. Models with high prediction coefficients (R_P^2) were obtained for the determination of SSC (R_P^2>0.9), TA (R_P^2>0.9) and TPC (R_P^2>0.8). The results demonstrate a good capacity of FTIR-ATR spectroscopy to predict the basic quality parameter of red fruit juices.

Acknowledgement
Grant 2016/23/B/NZ9/03591 from the National Science Centre, Poland, is gratefully acknowledged.

OLIVE RIPENING ASSESSMENT METHODOLOGIES USING DIGITAL IMAGE ANALYSIS

J. Ezenarro[1], A. García-Pizarro[1,2], D. Schorn-García[1], M. Mestres[1], L. Aceña[1], O. Busto[1], R. Boqué[3]
[1]*iSens Group. QAQO Dpt. Universitat Rovira i Virgili, Tarragona, Spain*
[2]*Olive Production and Oil Technology Group. Fruit Production Dpt. IRTA, Constantí, Spain*
[3]*GQQiN Group. QAQO Dpt. Universitat Rovira i Virgili, Tarragona, Spain*
jokin.ezenarro@urv.cat

Analysing the colour of fruits, especially olives, is one of the most basic steps for the evaluation of their ripening and, therefore, to decide when to collect them. Traditionally, this assessment is carried out by experts by determining the colour of each olive on a reference scale and then calculating a global Maturity Index (MI) of the tree/field [1]. Instead, using digital image analysis can be a more objective way to make a quantitative evaluation of the ripeness of olives and other fruits, as colour and ripeness are directly related and only one image of a representative number of olives is needed (Figure 1) [2].

Digital images are based on the decomposition of the colour of each pixel into a 3D space, RGB (Red, Green and Blue) for instance, and the statistical analysis of these values can be used to evaluate the ripening of imaged olives [3].

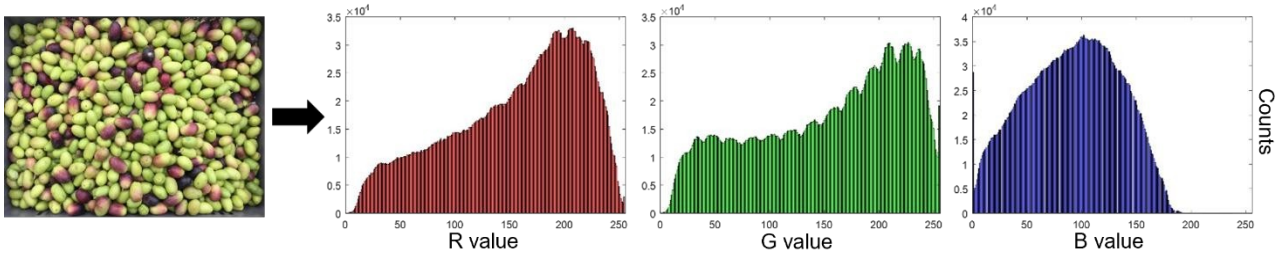


Figure 1 – Decomposition of a digital image of olives into its Red, Green and Blue channel histograms.

In the histograms of the RGB values (Figure 1), mostly in the Green channel, it can be seen that there are two main peaks: one in the lower values (dark pixels) and another one in the higher values (light pixels). These peaks evolve as the olives ripen; the dark peak gets higher as the light one gets lower, resembling the spectra of one specie reacting into another.

In this work, the histograms have been processed as spectral data to study the correlation between sample colour and maturity. For this purpose, several methods have been tried and compared: 1) univariate logarithmic regression between histogram peak means and MI; 2) PCA decomposition of histograms and correlation of scores with MI; 3) PLS regression of histograms and MI; and 4) MCR decomposition of histograms and correlation of concentration profiles with MI. In addition, low-rank data fusion strategies have been used to join and analyse different channel histograms simultaneously.

All methods have offered a good way to correlate colour and maturity of olives: similar correlation coefficients (R^2) and prediction errors (RMSE) have been obtained, showing that when the information of interest is related to the main source of measurements' variability this information can be easily obtained with many different chemometric tools.

References

[1] I. Mínguez-Mosquera, L. Gallardo-Guerrero. *J. Sci. Food Agric.* **69 (1)** (1995) 1–6.
[2] E. Guzmán, V. Baeten, J.A.F. Pierna, *et al*. *J. Food Sci. Technol.* **52** (2015) 1462–1470.
[3] A. Antonelli, M. Cocchi, P. Fava, *et al*. *Anal. Chim. Acta*. **515 (1)** (2004) 3–13.

Acknowledgements

Proyecto PID2019-104269RR-C33 financiado por MCIN/AEI/10.13039/501100011033.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

AN EXPLORATORY STUDY ON MONITORING TOMATO PLANT GROWTH BY NEAR INFRARED PORTABLE DEVICES

Bonifazi G.[1*], Gasbarrone G.[2], Gattabria D.[1], Serranti S.[1]

[1] *Sapienza University of Rome - Department of Chemical Engineering, Materials, Environment (DICMA), via Eudossiana 18, Roma, Italy;*

[2]*Sapienza University of Rome - Ce.R.S.I.Te.S - Research and Service Center for Sustainable Technological Innovation, Sapienza - University of Rome, Via XXIV Maggio 7, Latina (Italy);*

giuseppe.bonifazi@uniroma1.it

A non-destructive method based on infrared spectroscopy, allowing to evaluate the growth of tomato plants, is presented. Different types of tomatoes plants (i.e.*Crusoe*, *Marmarino* and *Red Datterino*) were monitored. They were grown in fertigate experimental greenhouses utilizing the final product of the transformation of the sewage sludge as resulting from a new biokinetic wastewater treatment plant. Main objectives of the study were to monitor the growth of tomato plants, using a non-destructive and non-invasive method based on tomatoes plants leaves reflectance spectra collection by spectrophotometry and to perform the further comparison with those of the same plants grown in conventional greenhouses with traditional fertilizers. A JDSU MicroNIR™ portable spectrophotometer, operating in the near-infrared region (NIR) (950-1650 nm) was utilized. To follow the phenological development of the seedlings, for each plant, grown in the experimental greenhouse, 5 spectra were acquired, every 6 days, for 3 different leaves for a total of 6 time intervals (i.e. T0=0 days to T5: 30 days). To perform the comparison, about 15 reflectance spectra were acquired for the plants grown in conventional conditions and reached the desired final stage of production. Starting from the collected spectra, Principal Component Analysis (PCA) was adopted to carry out an exploratory analysis and to evaluate the possible tomatoes plants growth and production variability. Once the PCA model was developed, a comparison was performed between the data acquired in the experimental (i.e. fertigation) and conventional (i.e. traditional fertilizers) greenhouse where tomatoes plants reached the desired final stage of healthy and production. As a result, it was possible to follow the growth of the plant at different times (i.e. negative space of score plot in PC1: T0, T1, T2 and T3), as well as the production (i.e. positive space of score plot in PC1: T4 and T5) (Figure 1). Furthermore, PCA score plot clearly outlines as the spectral attributes of tomato plants grown in experimental greenhouses (50-72 days after planting) are very similar to those of tomato plants grown with traditional fertilizers (119 days after planting).

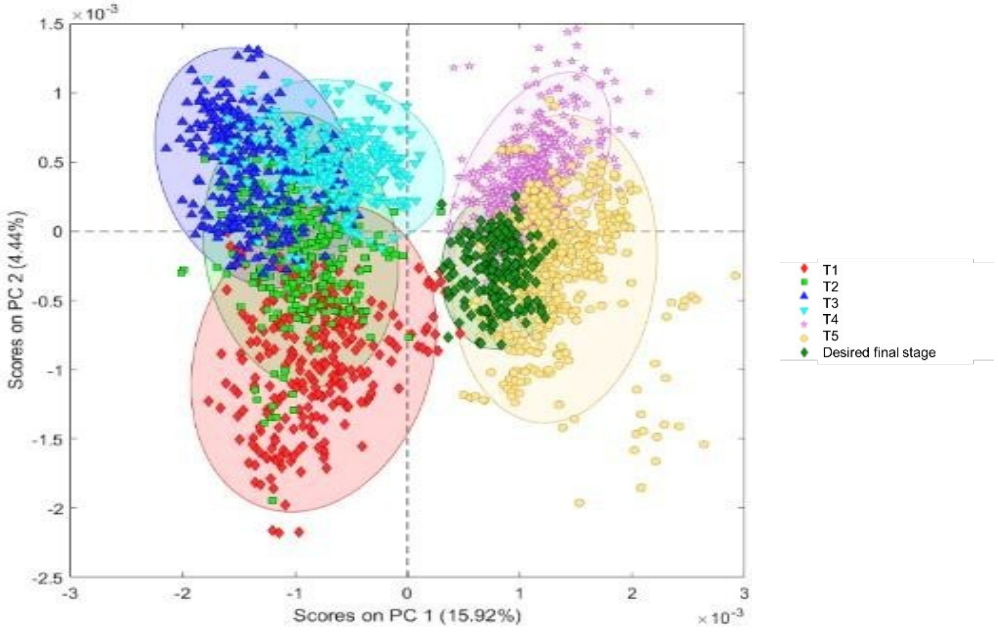


Figure 1 – PCA scores plot of the first two principal components. PCA scores related to the "Desired final stage" (conventional greenhouse in production: 119 days from planting) overlap with T4 and T5 scores (experimental greenhouse: 50-72 days after planting).

CHEMOMETRIC APPROACHES TO ENHANCE THE POTENTIAL OF NEW IR SPECTROSCOPIC TECHNOLOGIES

H. C. Goicoechea[1], M. R. Alcaraz[1], C. K. Akhgar[2], A. Schwaighofer[2], J. Ebner[3], O. Spadiut[3], B. Lendl[2]

[1] *LADAQ-FBCB-CONICET, Universidad Nacional del Litoral, Santa Fe, Argentina*
[2] *Institute of Chemical Technologies and Analytics, TUWien, Vienna, Austria*
[3] *Institute of Chemical, Environmental and Bioscience Engineering, TUWien, Vienna, Austria*

hgoico@fbcb.unl.edu.ar

It is well demonstrated that alternating least square (ALS)-based approaches significantly aid in comprehensively understanding the behaviour of a system under study due to the possibility of obtaining meaningful results with physical and chemical interpretation. This tool has been vastly exploited, particularly in the spectroscopy field, aiming to enhance the efficiency of experimental studies. Notably, applications based on novel instrumental technologies are profiting from its advantages encouraging a mutual broadening of their capacities. Here, the benefits of the usage of ALS-based methodologies in quantum cascade laser-infrared (QCL-IR) spectroscopic-based applications for protein analysis are described.

Mid-IR spectroscopy is a well-established analytical technique routinely employed to study the structure of polypeptides and proteins in a label-free manner. However, the low feasible path lengths needed in conventional Fourier transform (FT)-IR spectrometers for IR transmission measurements of proteins in aqueous solutions are a considerable impairment for the robustness of analysis and impede flow-through measurements and high-throughput applications. In this regard, a new technology based on QCLs emerges as a light source highly interesting for measurements of aqueous samples [1]. Notwithstanding, their applicability in protein analysis in dynamic processes is a hot research topic.

First, a QCL-based setup for mid-IR transmission measurements in the protein amide I region was used for monitoring dynamic changes in the secondary structure of proteins. α-Chymotrypsin acts as a model protein, which gradually forms intermolecular β-sheet aggregates after adopting a non-native α-helical structure induced by exposure to 50% TFE. The effects of varying pH values and protein concentration on the rate of β-aggregation were studied. Extended multivariate curve resolution (MCR)-ALS was employed to obtain pure spectral and concentration profiles of the temporal transition between α-helices and intermolecular β-sheets. The results demonstrated the potential and versatility of the QCL-based IR transmission setup to monitor dynamic changes of protein secondary structure in aqueous solutions coupled with chemometric analysis [1].

Second, a different QCL setup for mid-IR transmission spectroscopy in the amide I and II region was used for monitoring pH-induced changes in the secondary structure of β-lactoglobulin. Chemometric analysis of the dynamic IR spectra was performed by MCR-ALS. Then, a postprocessing procedure based on wavelet analysis was implemented to extract information about protein spectra and spurious signals that may interfere with the result interpretation [2].

Last, a QCL-based flow-through mid-IR spectrometer was placed in-line with a preparative size exclusion chromatography (SEC) system to demonstrate real-time analysis of protein elution with overlapping chromatographic peaks. Chemometric analysis by self-modelling mixture analysis (SMMA) and MCR enabled accurate quantitation and structural fingerprinting across the protein elution transient. The acquired concentration profiles were found to be in agreement with off-line liquid chromatography (HPLC) reference analytics performed on the collected effluent fractions. These results demonstrate that QCL-IR detectors in combination with chemometrics can be used effectively for in-line, real-time analysis of protein elution.

All these works shed light on the fact that EC-QCL-based IR transmission setup combined with chemometric analysis has high potential and versatility for dynamic and flow-through applications.

References

[1] M.R. Alcaraz, A. Schwaighofer, H. Goicoechea, B. Lendl, Anal. Bioanal. Chem. 408(15) (2016) 3933-3941.

[2] A. Schwaighofer, M.R. Alcaraz, L. Lux, B. Lendl, Spectrochim. Acta A Mol. Biomol. Spectrosc. 226 (2020) 117636.

FEASIBILITY OF MCR-ALS TO EXPLOIT THE SECOND-ORDER ADVANTAGE WITH FIRST-ORDER AND NON-BILINEAR SECOND-ORDER DATA. A SYSTEMATIC CHARACTERIZATION

H.C. Goicoechea , F.A. Chiappini , F. Gutiérrez[1,3], A.C. Olivieri[2,3]

[1]*Lab. de Desarrollo Analítico y Quimiometría (LADAQ), Cát. de Química Analítica I, Fac. de Bioq. y Cs. Biológicas (FBCB), Universidad Nacional del Litoral (UNL),Ciudad Universitaria, Santa Fe, Argentina*

[2]*Depto. de Química Analítica, Facultad de Cs. Bioquímicas y Farmacéuticas (FBioyF), Universidad Nacional de Rosario (UNR), Instituto de Química de Rosario (IQUIR-UNR-CONICET),Rosario, Argentina*

[3]*Cons. Nac. de Investigaciones Científicas y Técnicas (CONICET), Godoy Cruz 2290, CABA, Argentina*

hgoico@fbcb.unl.edu.ar

Multivariate curve resolution-alternating least squares (MCR-ALS) continues to be an object of great interest among chemometricians. In the analytical chemistry field, it has led to the development of numerous quantitative applications due to its versatility and ability to exploit the second-order advantage. However, the predictive capacity of MCR-ALS models can be severely affected by rotational ambiguity (RA), which causes lack of solution uniqueness. In recent years, significant contributions have been made in relation to the diagnosis and minimization of RA [1] and the interest of using MCR-ALS to exploit the second-order advantage has spread to other types of data [2], such as first-order and non-bilinear second-order data. In the case of first-order data, although MCR-ALS has already been applied, the second-order advantage has not always been exploited and RA not fully characterized. This latter aspect is critical since it has been theoretically demonstrated that, for first-order calibration in the presence of unmodeled interferences, solution uniqueness is never reached in MCR-ALS modelling [3]. Nevertheless, previous reports reveal that satisfactory analytical results can still be obtained, even in the presence of RA. On the other hand, regarding second-order calibration, it is well-known that most classical chemometric models for second-order data entails the assumption of low rank bilinearity, which cannot be accomplished by all instrumental techniques. This represents a limitation for the development of novel second-order calibration protocols, since methods for comprehensive modelling of non-bilinear second-order data remain only partially explored. In this context, an interesting alternative is to unfold the non-bilinear data matrices obtained for a set of samples, to generate a single bilinear matrix. Therefore, these second-order data can be treated as first-order data and then modelled using MCR-ALS.

In this work, a systematic study was carried out with the aim of tackling the question of when and why MCR-ALS is able to reach the second-order advantage in calibration with first-order data, considering both genuine first-order data [4] and unfolded second-order non-bilinear data [5]. In both scenarios, simulated and experimental calibration and prediction data sets were generated, considering chemical systems of one analyte in the presence of uncalibrated interferents. For each calibration-prediction dataset, a unique bilinear two-way array was submitted to MCR-ALS decomposition, keeping the instrumental signals in the columns (non-augmented mode) and the sample indices in the rows ("augmented-mode"). Purest variables was conducted for ALS initialization, under the restrictions of non-negativity and correspondence between species, and the diagnosis of RA was performed through the grid search methodology. In all cases, the analytical performance of MCR-ALS was interpreted regarding the evaluation of RA, the type of ALS initialization, the signal overlapping patterns and the degree of instrumental noise. As a general conclusion, it was observed that the degree of RA is drastically reduced when the analyte presents, at least, one region of individual contribution (local selectivity) in the non-augmented mode and the ALS algorithm is initialized in the sample direction. Under these conditions, satisfactory predictions can be obtained, which support the ability of MCR-ALS to achieve second-order advantage with first-order data in a single-analyte chemical system.

References

[1] A.C. Olivieri, *Anal. Chim. Acta* **1156** (2021), 338206.
[2] G. Ahmadi, R. Tauler, H. Abdollahi, *Chemom. Intell. Lab. Syst.* **142** (2015) 143-150.
[3] R. Manne, *Chemom. Intell. Lab. Syst.* **27** (1995) 89e94.
[4] F.A. Chiappini, F. Gutierrez, H.C. Goicoechea, A.C. Olivieri, *Anal. Chim. Acta* **1161** (2021) 338465.
[5] F.A. Chiappini, F. Gutierrez, H.C. Goicoechea, A.C. Olivieri, *Anal. Chim. Acta* **1181** (2021) 338911.

APPROXIMATION OF MARTIAN ROCK EMISSION SPECTRA BY MULTIPARAMETRIC OPTIMISATION

I. Krylov[1], S. Zaytsev[1], T. Labutin[1]
[1]*Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia*
ikrylov@laser.chem.msu.ru

Thanks to its versatility, laser-induced breakdown spectroscopy (LIBS) can be used to analyse unique specimens, such as Mars surface, ocean floor, molten metal and radioactive waste. In many cases like these, there are no appropriate certified reference materials. Boltzmann plot is a well-known approach for calibration-free (CF) LIBS; it requires optically thin plasma under local thermodynamic equilibrium. It is frequently impossible to strictly fulfil these requirements; extreme conditions like open space or deep sea can further complicate preliminary checking of these conditions.

Stationary laser-induced plasma can be modelled, synthesising a spectrum for a given elemental composition and other conditions [1]. Using multiparametric optimization, it's possible to fit the model to the experimental data, with relative elemental concentrations and plasma conditions serving as variables. A gradient-free optimisation method has been chosen for reasons related to implementation.

The model was tested on steel samples and aluminium alloys. Various shapes of the loss function have been considered, making it possible to account for analytical lines of very different intensity and highlight different detail of the spectra. The stability of the solution has been verified by repeated runs with random initial parameters.

It has been shown that the homogeneous plasma model is enough to predict several elements with lines in narrow spectral regions, while a multi-zone model with different temperatures and electron densities provides the best results for the full spectra, although in the latter case the accuracy is lower. Finally several Curiosity ChemCam spectra were fitted by our algorithm to predict the composition of the studied object.

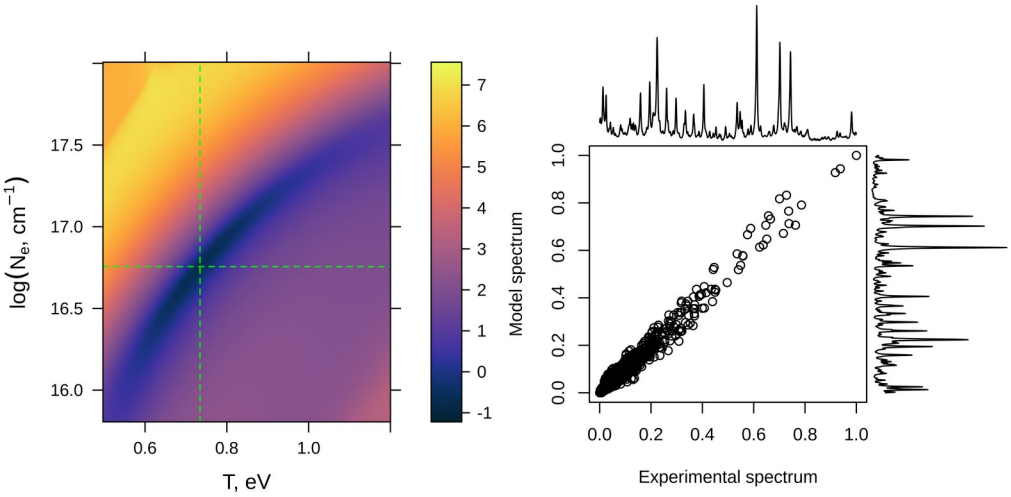


Figure 1 – Left: example surface of the logarithm of the loss function with the plasma temperature and electron density as parameters, the optimal point shown in green dashed line. Right: the optimal model spectrum compared against experimental spectrum of a C9 steel sample.

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

References

[1] S. M. Zaytsev, A. M. Popov, T. A. Labutin, *Spectrochim Acta, Part A.* **158** (2019) 105632.

AN ASSESSMENT OF THE POTENTIAL OF DIFFERENT VIBRATIONAL SPECTROSCOPIC TECHNIQUES IN CLASSIFICATION OF VARIOUS TYPES OF LIQUID MILK BY USING MULTIVARIATE CHEMOMETRIC METHODS

Saeedeh Mohammadi[1], Aoife Gowen*
[1]*University College Dublin, Dublin, Ireland*
[*]*University College Dublin, Dublin, Ireland*
Saeedeh.mohammadi@ucdconnect.ie

Spectroscopic methods such as infrared, Raman, and fluorescence spectroscopy have attracted a lot of attention recently because they provide quick, non-destructive, and simple methods for analyzing dairy products, particularly liquid milk. Since the potential of these spectroscopic techniques for being employed as on-line and at-line tools is also of paramount importance, optimization of experimental set-up for the analysis of liquid milk samples is a critical point in which dairy industries are interested.

In this research, a range of vibrational spectroscopic techniques were compared in the classification of eight types of commercial liquid milk (butter, fresh, heart active, lactose-free, light, slim line, and super milk, supplier: Avonmore, Ireland). Five different spectroscopic instruments were used to measure the three independent samples of each milk type and chemometric methods were applied to the data. Samples were analyzed using, a portable Raman spectrometer and Raman microscope with the laser of 785nm wavelength, near-IR spectrometer and portable and bench-top FT-IR systems. The experimental set-up was optimized separately for each instrument and the spectra were analyzed with principal component analysis (PCA). Application of PCA allowed visual separation of groups in score space (see Figure 1, score plot of PC1 vs PC2 vs PC3). The best visual separation of milk types was achieved with portable Raman and bench-top NIR data with PCA.

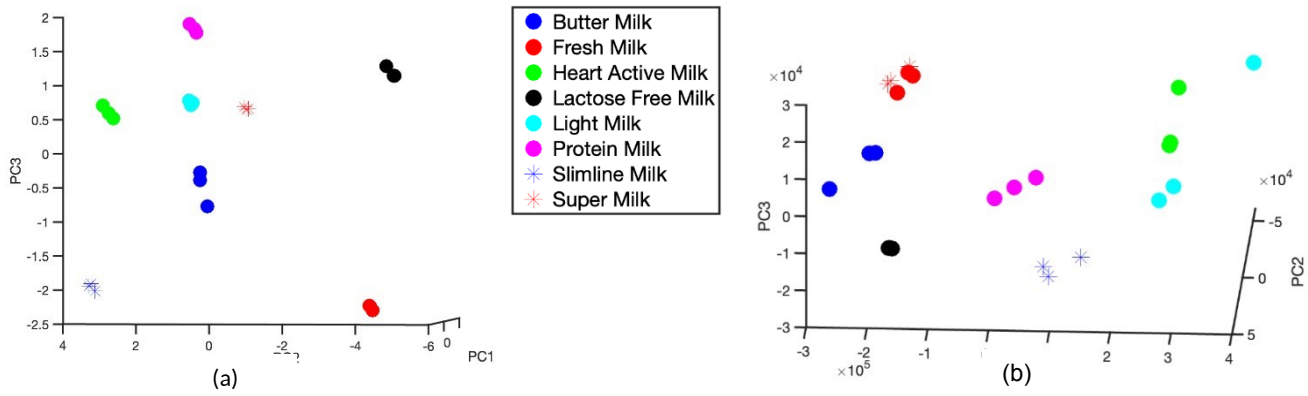


Figure 1 – (a) Score plot of PC1 vs PC2 vs PC3 deniatbo by PCA analysis of the NIR spectra and (b) portable Raman spectra of liquid milk samples

References

[1] R. M. El-Abassy, P. J. Eravuchira, P. Donfack, B. von der Kammer, and A. Materny, Vib Spectrosc, 56 (2011), 3–8,
[2] S. Ehsani, E. M. Dastgerdy, H. Yazdanpanah, and H. Parastar, Journal of Chemometrics, (2022) ,3395

DISCRIMINANT ANALYSIS OF THREE AND FOUR-WAY FLUORESCENCE DATA FOR CLASSIFICATION ISSUES

A. Muñoz de la Peña[1], I. Durán Merás[1], O. Monago-Maraña[2], J. Domínguez Manzano[1]
[1]*Universidad de Extremadura, Departamento de Química Analítica, Badajoz, Spain*
[2]*Universidad Nacional de Educación a Distancia, Las Rozas, Madrid, Spain*
arsenio@unex.es

In this Communication, several methods, intended for different classification issues in the forensic and food areas, and using discriminant analysis of three- and four-way fluorescence data, are presented and discussed.

On the forensic field, the analysis of fibers from clothes is of paramount importance when investigating a crime scene. Non-destructive excitation-emission fluorescence (EEM) microscopy, combined with parallel factor analysis (PARAFAC) supervised by lineal discriminant analysis (LDA) and discriminant unfolded partial least-squares (D-UPLS), allowed the discrimination between visually indistinguishable fibers [1]. The same technique combined with D-UPLS was used to classify pre-dyed textile fibers exposed to weathering and photodegradation. These results provide information on the weathering history of a fiber [2].

On the food area, to discriminate between smoked and non-smoked paprika samples, the full information of EEMs was processed with the aid of D-UPLS, allowing an adequate classification of unknown paprika samples into both categories [3].

On the same context, the detection of extra virgin olive oil (EVOO) adulteration was proposed by using the autofluorescence EEM profiles combined with D-UPLS. The models demonstrated the possibility of detecting adulterations of extra virgin olive oils with percentages of around 15 % and 3% of olive and olive pomace oils, respectively [4]. In addition, front-face fluorescence was used for virgin olive oil monitoring under different photo- and thermal-oxidation procedures. The full EEM information was processed by PARAFAC supervised by LDA. The three -way data models allowed the discrimination between non-irradiated and irradiated EVOO samples. With a temperature of 80 ºC and a heating time of 30 min, the first PARAFAC component showed a remarkable modification in its profile and LDA-PARAFAC allowed the discrimination between non-heated and heated EVOO samples [5].

Finally, the combination of EEM three-way data with LDA-PARAFAC and D-UPLS, allowed the discrimination between the first and the second maturation stages of Tempranillo grapes. The incorporation of an additional mode to the data, achieved by a diethyl ether extraction, gives rise to a four-way excitation-emission-solvent-samples data set. The four-way models, in combination with LDA-PARAFAC and D-UPLS, allowed the classification of Tempranillo grapes according to hydric status [6].

Acknowledgements
The authors are grateful to Grant PID2020-112996GB-I00 funded by MCIN/AEI/ 10.13039/501100011033, and Junta de Extremadura (Ayuda a Grupos GR21048) for financial support.

References

[1] A. Muñoz de la Peña, N. Mujumdar, E. C. Heider, H. C. Goicoechea, D. Muñoz de la Peña, A. D. Campiglia, Anal. Chem., 88 (2016) 2967-2975.
[2] N. Mujumdar, A. Muñoz de la Peña, A. D. Campiglia, Forensic Chemistry, 12 (2019) 25-32.
[3] O. Monago-Maraña, T. Galeano-Díaz, A. Muñoz de la Peña, Food Anal. Methods, 10 (2017) 1128-1137.
[4] I. Durán Merás, J. Domínguez Manzano, D. Airado Rodríguez, A. Muñoz de la Peña, Talanta, 178 (2018) 751-762.
[5] J. Domínguez Manzano, A. Muñoz de la Peña, I. Durán Merás, Food Anal. Methods, 12 (2019) 1399-1411.
[6] M. Cabrera-Bañegil, E. Valdés-Sánchez, A. Muñoz de la Peña, I. Durán-Merás, Talanta, 199 (2019) 652-661.

Hyperspectral imaging data: clustering or spectral unmixing?

A. Olarini[1,2], M. Cocchi[2], L. Duponchel[1], C. Ruckebusch[1]
[1]*LASIRE - Université de Lille, Villeneuve d'Ascq, France*
[2]*Università di Modena e Reggio Emilia, Modena, Italia*
alessandra.olarini@unimore.it

Clustering and spectral unmixing are widely used methods in hyperspectral imaging [1]. The use of either methods strictly depends on the type of objective under investigation. On one hand clustering methods aim at partitioning similar pixels, assuming that the spectral signature measured at one pixel is characteristic of a unique cluster (e.g. case study in Figure 1.A). On the other hand spectral unmixing methods aim at identifying individual sources of spectral variations. Each pixel can be described as a linear mixture of the pure spectra characteristic of those unknown individual sources (e.g. case study in Figure 1.B). However, some situations may be borderline for the use of both methods, which is the case in a mixture of two or more components for which only some compositions of the mixture are observed and data distribution shows clear clusters (Figure 1.C). This is what spectroscopists often face in analyzing biomedical images [2,3].

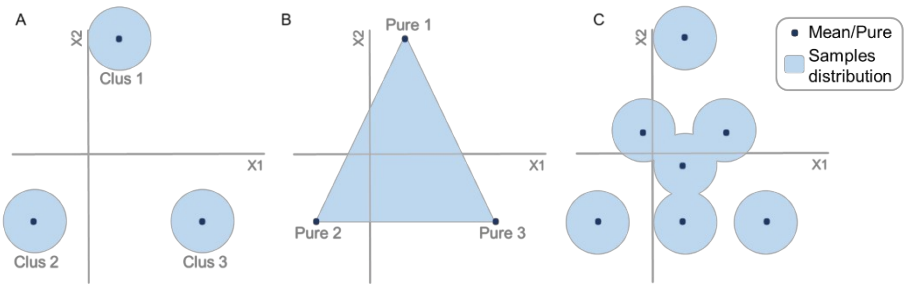


Figure 1 – Normalized scores in the PC subspace.

Advantages and disadvantages of these approaches are evaluated in this work, in relation to the structure of the data, analyzed in both the spectral and spatial mode. Hyperspectral image simulations of three pure components facing different situations are studied to find a general application framework for clustering and spectral unmixing methods. K-means and Multivariate Curve Resolution–Alternating Least Squares (MCR–ALS) approaches are applied to each dataset distinguishing situations with and/or without spectral overlap and with and/or without pure pixels modulating the noise. K-means performance is evaluated for each simulation with and without image background and the results obtained for a different number of components are compared. Simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) is used to generate the initial profiles estimation prior MCR [4]. Results showed that MCR identified the three components in all the simulated datasets belonging to the cases in Figure 1.B and 1.C, including noisy simulations, and the results were highly depended of the added type of noise. K-means succeeded in clustering the data in the situation of Figure 1.A and 1.C. Of course, the clustering technique for scenarios as 1.B, in which the discrete samples are missed by varying the concentration profiles, was not efficient in terms of analysis time and results obtained. The presence of mixed pixels is one limitation because it can result in a wrong data partition. This work is a preliminary work and future development on real datasets of biological tissue aims to investigate pure pixels in challenging scenarios.

References

[1] J. M. Amigo, *Hyperspectral Imaging,* Elsevier **32**(2019).
[2] S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, *Anal. Chim. Acta* **705**(2011), 182-192.
[3] H. Valenta, S. Hugelier, S. Duwé, G. Lo Gerfo, M. Müller, P. Dedecker, W. Vandenberg, *Biophys. Rep.* **1**(2021)
[4] W. Windig, J. Guilment, *Anal. Chem.* **63**(1991), 1425-1432

Update of Transmission Raman Spectroscopy Calibration Models using Dynamic Orthogonal Projection (DOP)

Nicholas I. Pedge[1], Matthieu Papillaud[1], Jean Michel Roger[2]
[1]*AstraZeneca, Pharmaceutical Technology and Development, Macclesfield, UK*
[2]*UMR ITAP INRAE - SupAgro, Montpellier, France*
nicholas.pedge@astrazeneca.com

Transmission Raman Spectroscopy is a relatively new analytical technique for R&D and Quality Control labs that can be used to measure critical quality attributes (CQAs) such as assay or uniformity of unit dosage for pharmaceutical oral solid dosage products such as tablets or capsules.

The maintenance of validated multivariate calibration models for quantitative spectroscopic methods is an important aspect of an analytical method lifecycle. An ongoing issue at the manufacturing site was the impact of the annual preventative maintenance on the performance of the models. This was further exacerbated as a single calibration model was constructed using data acquired using two different instruments (Agilent TRS100), but the impact of the maintenance for each instrument was different. For this reason, simple post-processing such as slope/bias correction was not effective because it could not simultaneously correct the predictions from two different instruments that are exhibiting a prediction bias in opposite directions.

Dynamic Orthogonal Projection (DOP) is a procedure that was well suited to this calibration update application. The function of DOP is to eliminate the additional spectral variation present in new samples that was not captured in the original training set. To implement DOP, a set of new, labelled samples taken from several representative production batches were measured after the instrument service was completed. The measured reference values for each sample were used to estimate an "expected" spectrum for that sample from the original training set space. The difference between the "expected" and the "measured" spectra were collated into a difference matrix, **D**.

By applying the DOP orthogonalization (using difference matrix **D**) as an additional pre-processing step, it was possible to orthogonalize the original training set spectra to the new variation introduced by the instrument service. Recalculation of a new PLS model using the original (orthogonalized) training set successfully eliminated the new spectral variation and restored the performance of the models.

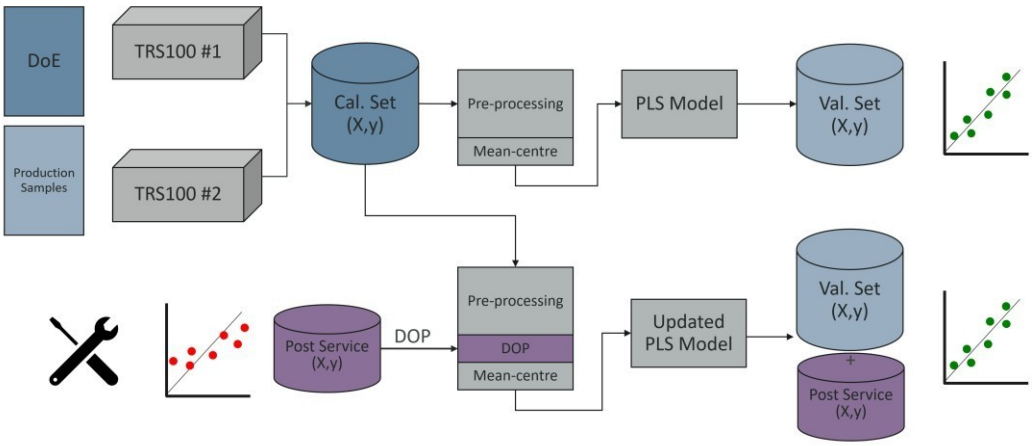


Figure 1 – Schematic showing how DOP was used to update an existing calibration set

References

[1] M. Zeaiter, J.M. Roger, V. Bellon-Maurel, *Chemo. Intelli Labs* **80** (2006) 227-235.
[2] J-M. Roger,J. C. Boulet, J. Chemo, **32** (2018), e3045.

CLASSIFICATION OF BITTER AND SWEET ALMONDS USING NIR MINIATURIZED INSTRUMENTS

Jordi Riu[1], Hawbeer Jamal Ahmed[1], Ricard Boqué[1], Barbara Giussani[2]

[1]*Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry, 43007 Tarragona (Catalonia), Spain*
[2]*Università degli Studi dell'Insubria, Science and High Technology Department, 22100, Como, Italy*
jordi.riu@urv.cat

Near-infrared (NIR) spectroscopy is a well-established and mature technique that has become the preferred tool in many fields of applications. NIR equipment, as most of the analytical instrumentation, has evolved from large, benchtop-based instruments through on-line and in-line instruments for industry, and finally to miniaturized portable lab-on-a-chip devices. Miniaturized NIR instrumentation is gaining importance in the last years despite the lack of a suitable and structured optimization in most of the analytical strategies used to obtain reliable results [1]. In order to have a clearer vision of the features and limitations of miniaturized NIR instruments, understanding the different sources of variability [2] may help to obtain a better performance of the results obtained with these instruments and to understand the capabilities and limitations of miniaturized NIR instruments.

In this study we have used two different NIR miniaturized instruments (SCiO, Consumer Physics and NeoSpectra Micro Development Kit, Si-Ware), with quite different characteristics and modes of operation, for the measurement of sweet and bitter almonds, since NIR spectroscopy has already shown its usefulness in measuring this type of samples [3]. The covariance and correlation matrices were used as methods to understand and evaluate the sources of variability that can influence the multivariate models built from these data and to infer the optimal spectroscopic range and the best pretreatments to be applied to obtain optimal results. This information has been used in the construction of classification models for the distinction of sweet and bitter almonds.

References

[1] B. Giussani, G. Gorla, J. Riu, *Crit. Rev. Anal. Chem.* (2022) DOI: 10.1080/10408347.2022.2047607
[2] G. Gorla, A. Taiana, R. Boqué, P. Bani, O. Gachiuta, B. Giussani, *Anal. Chim. Acta*, 1211 (2022) 339900
[3] M. Vega-Castellote, D. Pérez-Marín, I. Torres, J.M. Moreno-Rojas, M.T. Sánchez, *J. Food. Eng.*, 294 (2021) 110406

A solution based on sample weighting to the leverage problem in multivariate curve resolution – alternating least squares

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

[1] Université de Lille, LASIRE CNRS, Lille, France.
[2] Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Modena, Italy.
m.ahmad@live.nl

Multivariate curve resolution (MCR) encompasses several different algorithms that can tackle spectral mixture problems. One of the easiest routines to implement is the MCR – alternating least squares (MCR-ALS) algorithm which starts with an initial estimate for either the concentration profiles or spectra of the constituents, and alternates between estimating the other, with a least squares minimisation function. However, in situations where some constituents are present in a disproportionally large number of samples, relative to the others, one or more of the constituents can become "minor". This can result in the MCR-ALS solution drifting towards the "major"-constituents as the model is leveraged by the disproportionally large number of samples. This leverage issue, in MCR-ALS, has recently been investigated by Vitale et al. [1]. An example of this is visible in figure 1 (left), which shows the data cloud in the normalized row-space yielded by its principal component analysis (PCA) decomposition. In the plot, the three constituents and the sample coverage are clear. Although there are three linearly independent constituents, one is completely overshadowed by the other two. The MCR-ALS solution, obtained using only a non-negativity constraint and spectra of pure samples as initial estimates, is not accurate with respect to the true solutions (blue star) for the minor component. The estimated spectrum of the minor constituent (Comp. C) shows some irregularities, as well as non-zero values for certain spectral channels, even though the dataset is selective for the "minor"-constituent. As a solution to such an issue, we would like to propose using an alternating weighted LS (A-wLS) routine, instead of standard ALS. This was initially proposed by Wentzell et al. [2], but the algorithm used weights to incorporate measurement error information and impute missing data. We would like to propose a weighting scheme, where the samples are weighted based on their importance/relevance for the resolution process. An idea of how important/relevant samples are in this context can be obtained as described by Ghaffari et al. [3]. After applying a weighting scheme (figure 1, right), based on their importance, the solution obtained is significantly closer to the true solution, with the estimated spectra showing less irregularities with respect to the true spectra. We conclude with providing a general framework for a more robust MCR-ALS algorithm, exploiting the principles of weighted least squares.

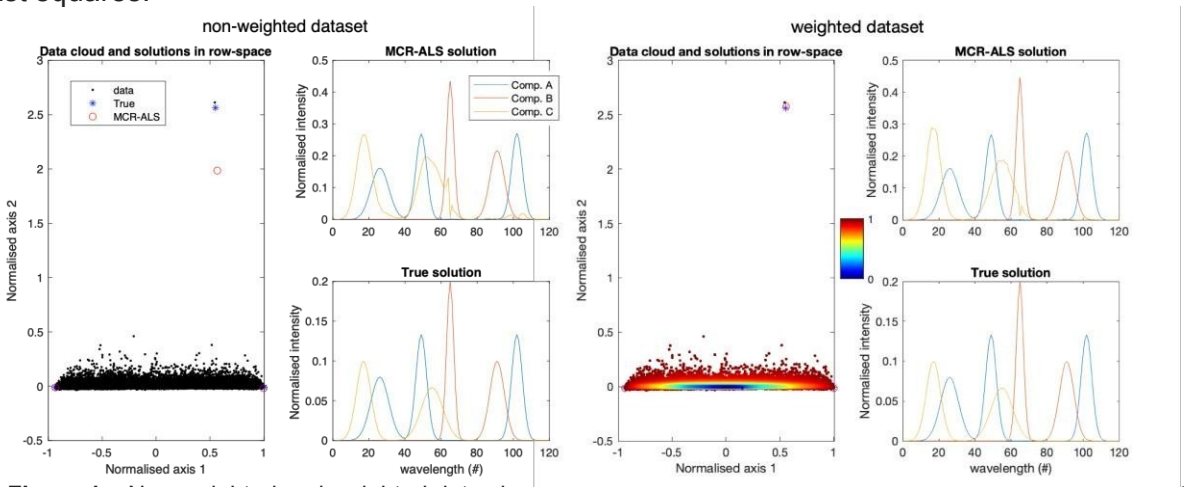


Figure 1 – Non-weighted and weighted data clouds, MCR-ALS and true solutions (colour-bar→ sample weights)

References

[1] Vitale R, Ruckebusch C., ChemRxiv. 2022, 10.26434/chemrxiv-2022-34g1j-v2
[2] Wentzell, P.D., Karakach, T.K., Roy, S. et al. BMC Bio. Inf. 7, 343 (2006), 10.1186/1471-2105-7-343
[3] Ghaffari M., Omidikia N., and Ruckebusch C., Anal. Chem. 2019 91 (17), 10.1021/acs.analchem.9b02890

DIFFERENT CHEMOMETRIC STRATEGIES TO CONTROL PTFE IN Ni-P/PTFE ELECTROLESS COATING BATHS BY UV-VIS

G. Albizu[1], G. Etxeberria[1], M. Ostra[1], M. Vidal[1]
[1]*University of the Basque Country (UPV/EHU), Department of Applied Chemistry, San Sebastian, Spain*
gorka.albizu@ehu.eus

In the coating industry, it is common to produce a composite coating incorporating particles into the metal matrix in order to improve some of the properties of the covering. One of the most widely used particles in Ni-P electroless is the polytetrafluoroethylene (PTFE) due to, among other properties, the high corrosion resistance and lower friction of the resulting Ni-P/PTFE coatings [1]. It is very important to control the PTFE concentration in the bath in order to readjust to the optimal value because as it is deposited, the concentration of the particles decreases. Moreover, it is also important to control that the concentration does not exceed some critical value, otherwise the particles can agglomerated and decrease the level of incorporation [2].

Nowadays, the weighing difference is the conventional methodology for PTFE determination in electroless bath. This method has several steps that increase the measurement error and is time consuming. Therefore, in this work a method based on UV-Vis spectrophotometric is used so that the measurement time and the amount of sample necessary to carry out the determination of PTFE are reduced. The PTFE is a polymer that has a characteristic peak at 250 nm, as can be seen in the Figure 1a, but since it is dispersed in solutions, it also absorbs in the visible zone.

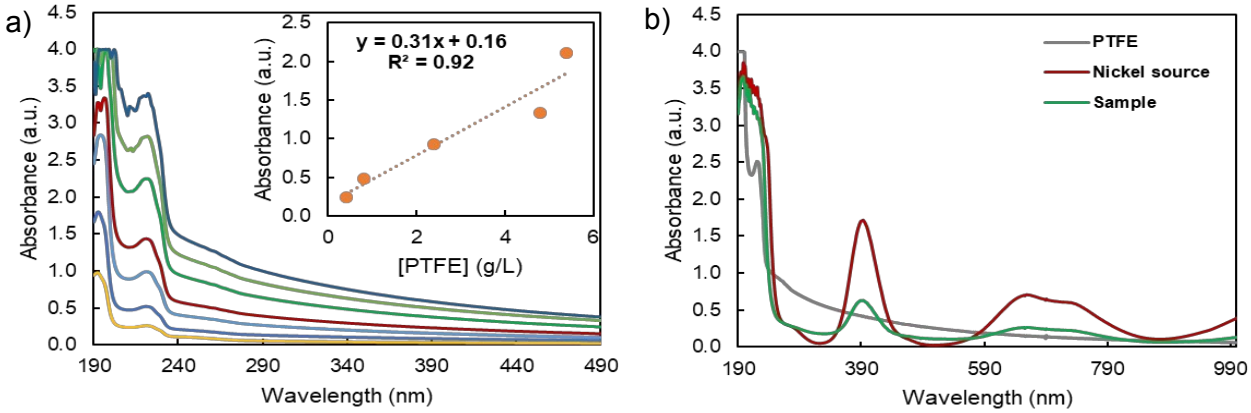


Figure 1 – a) Spectra of different amounts of PTFE and a calibration line using the absorbance at 225 nm. b) UV-Vis spectra of PTFE, Ni source and a real sample.

Figure 1a shows that there is a correlation between the PTFE concentration and the absorbance at 225 nm, but as it can be seen in the Figure 1b, the UV zone of the spectrum is saturated when a real sample is measured, due to the different organic compounds that are in this type of baths. To solve the problem different chemometrics algorithms have been applied. First, PLS models have been built in order to determine simultaneously the PTFE and the nickel concentration. Although the models had acceptable errors, as the bath was getting older the conditions changed, the prediction of the samples was not good enough. MCR was also applied to obtain the PTFE and Ni pure spectra from the real samples. With this algorithm, two different strategies were followed, an external calibration and the standard addition method. With the standard addition data, the classic standard addition method has been used after MCR, but also this data have been processing by using the net analyte signal (NAS) and the H-point standard addition method. In this work, the results obtained with the different methods and models are shown and compared.

References

[1] I. R. Mafi, C. Dehghanian, *Appl. Surf. Sci*. **257** (2011) 8653-8658.
[2] J. Sudagar et al, *J. Alloys Compd.* **571** (2013) 183-204.

PURE COMPONENT RECOVERY FOR RANK-DEFICIENT PROBLEMS

T. Andersons[1] ,M. Sawall[1], K. Neymeyr[1,2]
[1]*Universität Rostock, Rostock, Germany*
[2]*Leibniz-Institut für Katalyse, Rostock, Germany*
tomass.andersons@uni-rostock.de

Spectral data matrices can have an inherent rank-deficiency, that is, the rank of the product matrix $D=CS^T$ is smaller than the number of components in the factors. This can be caused by linear dependencies within the concentration profiles of pure components, for example, as is seen in the Michaelis-Menten kinetics. Typically, rank-deficiency occurs in the concentration profiles and the pure component spectra have full rank. Rank-deficient problems obfuscate the true chemical structure in multivariate curve resolution problems. This calls for an extension to the usual approaches to find feasible profiles for both factors.

We propose a twofold analysis of a rank deficient problem:

I. The Area of Feasible Solutions (AFS) can be found for the rank deficient factor. This results in bands of concentration profiles. [2]
II. If the proper concentration profiles are either given, calculated, measured or guessed, then the pure component spectra can be determined, but they are not unique, see Figure 1. The spectral band boundaries can be computed with linear programming problems for every feasible choice of concentration profiles. [1]

Together these steps provide the feasible solutions for both factors of a rank-deficient multivariate curve resolution problem.

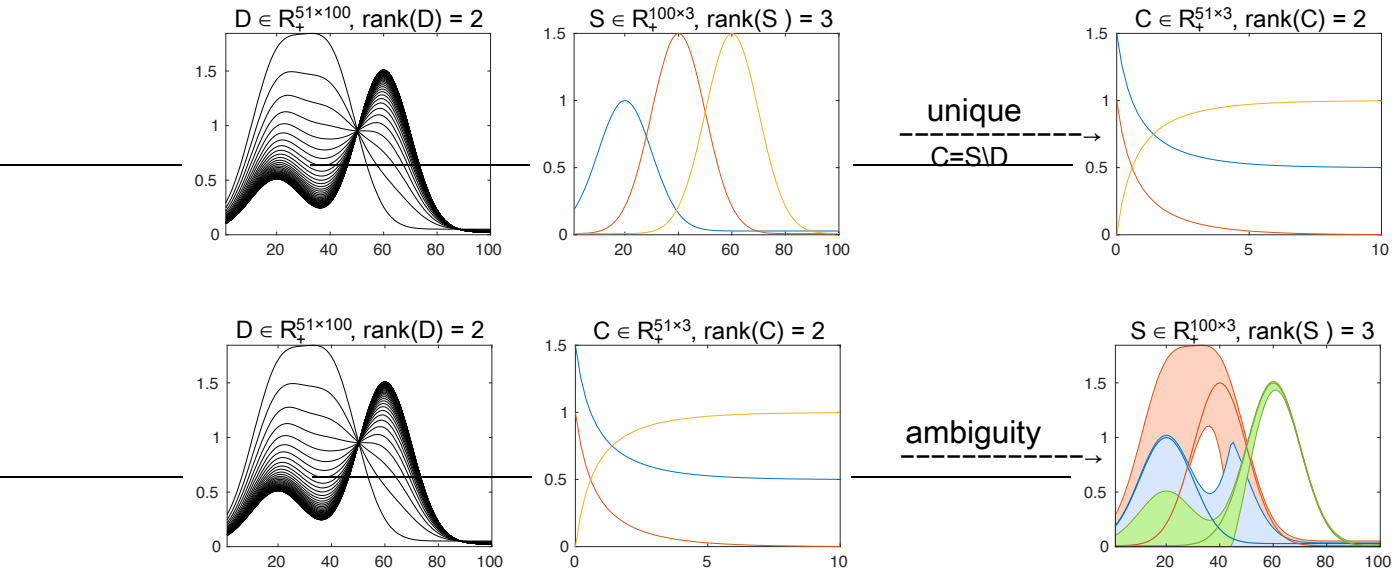


Figure 1 – The inherent uniqueness and ambiguity of a rank-deficient MCR problem.

References

[1] M. Sawall, T. Andersons, H. Abdollahi, S. Khodadadi Karimvand, B. Hemmateenejad, K. Neymeyr, *Chemom. Intell. Lab. Syst.* **226** (2022) 104577.
[2] M. Sawall, K. Neymeyr, *J. Chemom.* **35(3)** (2020) e3316.

Application of a segmented analysis by MCR-ALS on ^{1}H-NMR spectroscopy for the identification of adulterations in brown sugars

C. Fuentes[1,2], M. Rojas[1,2], R. Castillo[1,2], M. Öztop[3], A. Goksu[4]

[1]*Biospectroscopy and Chemometrics Laboratory, Biotechnology Center, Universidad de Concepción, Concepción, Chile*

[2]*Instrumental Analysis Deparment, Faculty of Pharmacy, Universidad de Concepción, 4070386 Concepción, Chile.*

[3]*Food Engineering Deparment, Middle East Technical University, Ankara, Turkey*

[4]*Kayseri Sugar, Kayseri, Turkey*

crisfuentes@udec.cl

Different types of brown sugar available in the market (Beet, Cane and Coconut) could be susceptible to food fraud given their similar organoleptic character and differences in their prices [1]. Certain minor organic compounds in addition to sucrose usually remain in the final product, which can be identified and used as markers for the differentiation of this sugars. The identification of a wide range of compounds in a complex matrix can be analyzed by nuclear magnetic resonance (NMR) spectroscopy; however, the information contained in an NMR spectrum can be a challenge given the complexity of its spectral interpretation. The objective of this study is to improve the differentiation of types of brown sugars and their mixtures through the integration and undirected resolution of the resonances in á^{1}H-NMR data set by Multivariate Resolution Curve-Alternating Least Squares (MCR-ALS) as an independent preprocessing method, which consists of dividing the data set into spectral windows containing between one and four resonances, where each independent window is resolved by MCR-ALS [2]. The exploration of data by PCA was evaluated comparing three analysis strategies. Using the original raw spectral dataset and binning data, it was not possible to differentiate the samples by sugar type, but using a reconstructed matrix with relative concentration values (MatrixC) obtained by MCR-ALS, it was possible to separate by type of sugar (brown or mixed) and even the coconut sample from its adulterations. In this way, the use of MCR-ALS as an independent preprocessing method applied to á^{1}H-NMR spectral data set demonstrates its potential by improving the selectivity of this technique and reducing the effect of overlapping signals, which would allow increasing classification rates of supervised classification methods such as k-nearest neighbor (KNN), smooth independent modeling of class analogies (SIMCA), or partial least squares discriminant analysis (PLS-DA) compared to preprocessing commonly applied to NMR data, which will be performed, evaluated and validated with a larger set of samples.

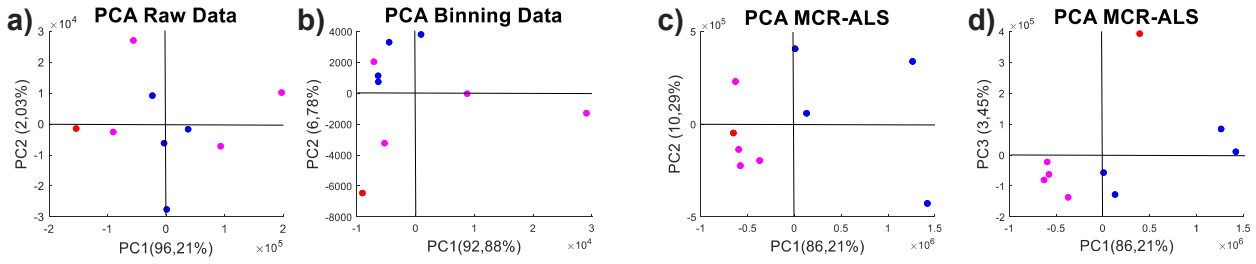


Figure 1 – **(a)** Scores of PC1 vs PC2 using Raw Data. **(b)** Scores of PC1 vs PC2 using Binning Data. **(c,d)** Scores of PC1 vs PC2 and PC1 vs PC3 using segmented data by MCR-ALS. Samples of Brown Cane Sugar are blue dots, Brown Coconut Sugar are red dot, and Mixtures are pink dots.

Acknowledgments:
The authors are grateful European Union's Horizon 2020 Research and Innovation Programmed-MSCA RISE under grant agreement # 101008228.

References

[1] R. Bachmann, A. Hornds, N. Paasch, R. Shcrieck, M. Weidner, I. Fransson & J.P. Schror. *Food Control*, **135** (2022).

[2] Y. Pérez, M. Casado, D. Raldúa, E. Prats, B. Piña, R. Tauler, I. Alfonso & F. Puig-Castellví. *Analytical and Bioanalytical Chemistry,* **412** (2020) 5696-5706.

Unmixing exponential signals by *Kernelizing*

Adrián Gómez-Sanchez[1,2], Olivier Devos[2], Raffaele Vitale[2], Anna de Juan[1], Cyril Ruckebusch[2]
[1]*Chemometrics group, Universitat de Barcelona, Barcelona, Spain*
[2]*LASIRE, Université de Lille, Lille, France*
aderegomez@gmail.com

Time-resolved fluorescence spectroscopy (TRFS) is a technique able to measure the emission decay of a fluorophore in the picosecond to nanosecond timescale [1]. However, unmixing fluorescence multiexponential decay signals is often very hard due to the high collinearity of the decays associated with the different fluorophores [2]. To solve this problem, slicing methodologies, such as PowerSlicing, have been proposed [3]. Slicing methodologies tensorize tables of exponential decay signals by putting together slices of local time ranges of the decay curves shifted by a certain lag, generating trilinear data and, thus, providing unique solutions via trilinear decomposition. Slicing approaches show very good performance but require that the number of sampling points along the time axis is large (typically a few hundreds) for the analysis of the fluorescence decay signals, for obtaining a large number of slices with rich time decay information. In this work, we propose a new methodology called *Kernelizing* that allows tensorizing tables of multiexponential decays for which only a few tens of sampling points have been measured.

The methodology is based on the properties of exponential decay curves that remain exponential (same exponential decay rate, only the pre-exponential factor changes) when convolved with another signal. Kernelizing works convoluting a set of distinct kernels with measured exponential decay signals. If multi-exponential decays have been measured, kernelizing will affect the individual contributions of individual signals. Thus, if a monoexponential decay is convolved by a kernel, the result is the same monoexponential decay with different intensity. If a multiexponential decay is convolved by a kernel, the monoexponential decays contributing to it will change intensity, generating a new multiexponential decay with the lifetimes of these monoexponential contributions remaining the same. Every different kernel, then, will provide, by convolution, a different convolved multiexponential decay. This particular behavior allows generating many new *samples* originating from the same original multiexponential decay curve. The Kernelizing approach works convoluting diverse kernels to an exponential decay. If Kernelizing is applied to a data table formed by a set of samples defined by decay curves, the result is, therefore, a data cube, which can be analyzed by PARAFAC, providing unique solutions.

To show the performance of the proposed approach, several simulated examples of Time Correlated Single Photon Counting (TCSPC) measurements with many and few tens of time channels were analyzed by PowerSlicing and Kernelizing to compare the results obtained. Kernelizing was also applied to TCSPC measurements of mixtures of ATTO fluorophores in solution to prove the good performance of the algorithm in real conditions. Finally, a FLIM image of a convallaria majari [4] was also analyzed to show how the lifetimes of the fluorophores present in a very complex sample can be extracted when only 12 sampling points are available.

References

[1] J. R. Lakowicz. Principles of Fluorescence Spectroscopy, 3rd ed. 2006, Springer, USA.
[2] A. A. Istratov, O.F. Vyvenko, Rev. Sci. Instrum. **70**(2) (1999) 1233-1257.
[3] S. B. Engelsen, R. Bro. J. Magn. Reson. **163**(1) (2003) 192-197.
[4] G. O. Williams, E. Williams, et al. Nat. Commun, 12(1) (**2021**) 1-9.

MULTI-LAYER MODELING OF TIME SERIES OF NMR SPECTRA

J. Hellwig[1,2] , D. Meinhardt, H. Schröder, E. Steimers[3], A. Friebel[3], T. Beweries[3], M. Sawall[1], E. von Harbou[3], K. Neymeyr[1,2]

[1]*Universität Rostock, Rostock, Germany*

[2]*Leibniz-Institut für Katalyse, Rostock, Germany*

[3]*Technische Universität Kaiserslautern, Kaiserslautern, Germany*

jan.hellwig@uni-rostock. de

Hard modeling of NMR spectra by Gauss-Lorentz peak models is an effective way for dimensionality reduction. In this manner high-dimensional measured data are reduced to low-dimensional information as peak centers, amplitudes or peak widths. For time series of spectra these parameters can be assumed to be smooth functions in time. We suggest to model these time-dependent parameter functions by cubic spline functions, which makes a stable quantitative analysis of NMR series possible even for crossing, highly overlapping peaks.

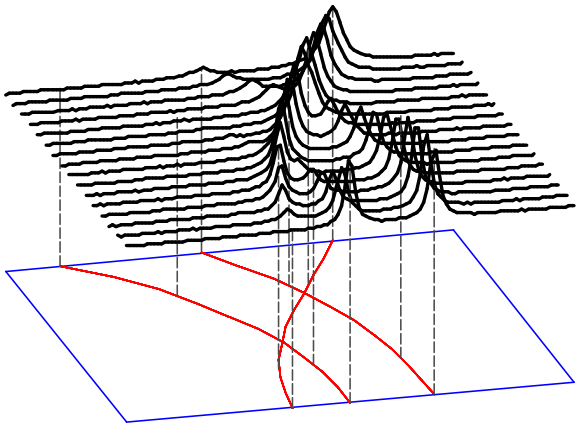


Figure 1 – Peak center values modeled at a few select spectra and the underlying cubic spline functions.

References

[1] D. Meinhardt, H. Schröder, J. Hellwig, E. Steimers, A. Friebel, T. Beweries, M.Sawall, E. von Harbou, K. Neymeyr, *J. Magn. Reson.* **339** (2022) 107212.

EXPLORING THE DYNAMIC EQUILIBRIA OF NON-CANONICAL DNA STRUCTURES BY MULTIVARIATE CURVE RESOLUTION AND 2D CORRELATION SPECTROSCOPY

N. Iaccarino[1], J. Amato[1], F. D'Aria[1], A. Randazzo[1], C. Giancola[1], A. Cesàro[2], S. Di Fonzo[2], B. Pagano[1]

[1]*Department of Pharmacy, University of Naples Federico II, Via D. Montesano, 49, 80131,Naples, Italy*
[2]*Elettra-Sincrotrone Trieste S. C. p. A., Science Park, 34149,Trieste, Italy*
nunzia.iaccarino@unina.it

Non-canonical DNA secondary structures include hairpins, cruciforms, parallel-stranded duplexes, triplexes, G-quadruplexes and i-motifs. Among all these structures, the i-motif (iM) structure has attracted a lot of attention in the last decade. It is formed by C-rich sequences and are composed of two intercalated duplexes, stabilized by hemi-protonated cytosine-cytosine$^+$ ($C{\cdot}C^+$) base pairs. Interestingly, putative iM-forming sequences have been identified in or near the promoter regions of more than 40% of human genes, including important oncogenes such as *BCL2* and *KRAS*, where they have been shown to be involved in the regulation of transcription, by means of mechanisms in which the iM structures appear to be in dynamic equilibrium with hairpin species (Figure 1).
Herein, the effects of pH, temperature, and presence of cell-mimicking molecular crowding conditions on conformational equilibria of the *BCL2* and *KRAS* i-motif-forming sequences were investigated by ultraviolet resonance Raman (UVRR) and circular dichroism (CD) spectroscopies. Multivariate curve resolution-alternating least square (MCR-ALS) of CD data was essential to model the presence and identity of the species involved. Analysis of UVRR spectra measured as a function of pH, performed also by the 2D correlation spectroscopy (2D-COS) technique, showed the role of several functional groups in the DNA conformational transitions, and provided structural and dynamic information. Then, the combined use of the two spectroscopic tools by means of heterospectral 2D-COS analysis made it possible to correlate the changes in the UVRR and CD spectra. The results of this study shed light on the factors that can influence at the molecular level the equilibrium between the different conformational species putatively involved in the oncogene expression [1].

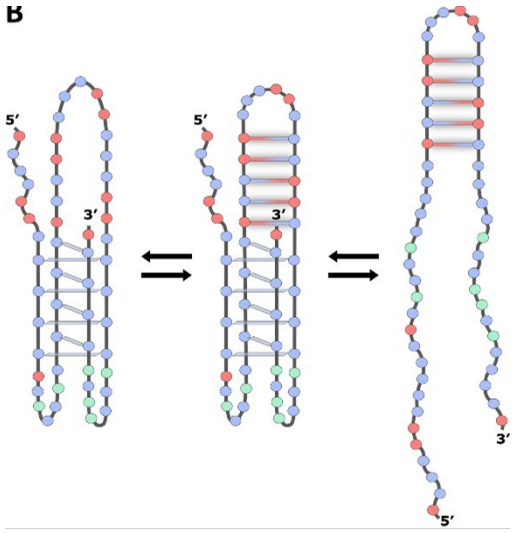


Figure 1 – Schematic illustration of the proposed folding pattern for the *KRAS* gene promoter iM in equilibrium with the hairpin species stabilized by Watson–Crick base pairs. The blue, red, orange, and green circles represent the cytosine, guanine, adenine, and thymine, respectively.

References

[1] J. Amato, N. Iaccarino, F. D'Aria, F. D'Amico, A. Randazzo, C. Giancola, A. Cesàro, S. Di Fonzo, B. Pagano *Phys. Chem. Chem. Phys.* **24** (2022) 7028-7044.

CAC 2022
XVIII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
29 AUGUST - 2 SEPTEMBER
ROMA

PSEUDO-UNIVARIATE CALIBRATION THROUGH MCR-ALS APPLIED TO ELECTROCHEMICAL DATA TO DETERMINE DIFFERENT AMINO ACIDS SIMULTANEOUSLY.

Luan P. Camargo[1], Mario H. M. Kilner[1], Dimas A. M. Zaia[1], Luiz H. Dall'Antonia[1], Paulo H. Março[2]

[1] *Graduate Program in Chemistry, Universidade estadual de Londrina (UEL) – CEP 86057-970, Londrina – Paraná – Brazil*

[2] *Post-Graduation Program of Food Technology, Federal University of Technology of the Paraná State (UTFPR) – CEP 87301-899, Campo Mourão – Paraná – Brazil*

paulohmarco@gmail.com

Molecules such as adenine (ADE), adenosine (ADO), and adenosine monophosphate (AMP) stand out due to their essential biological functions, which include cellular respiration, energy transduction processes, besides the constitution and formation of more complex molecules (for instance, DNA and RNA) [1,2]. Furthermore, studies have revealed its importance in prebiotic chemistry. Thus, to strengthen the knowledge about such molecules, it is necessary to develop a simple procedure for the simultaneous determination of ADE, ADO, and AMP. Electrochemical methods represent an important tool due to their high analytical frequency, good reproducibility, and relatively lower cost than traditional methods. In this sense, this work reported an efficient approach through pseudo-univariate calibration using Multivariate Curve Resolution with Alternating Least Squares (MCR-ALS) [3] to simultaneously determine the three species. The signals were provided by using an optimized differential pulse voltammetry method (DPV) with the boron-doped diamond electrode. The results indicated an excellent electrochemical response for the oxidation reaction of species individually (sweep from +0.80 to +1.50 V vs. Ag/AgCl), with concentrations varying from 5 to 300×10^{-7} mol L^{-1}. Although the results were promising and the electrochemical response reproducible, it was impossible to separate the oxidation peaks under the simultaneous form using the DPV. Nonetheless, MCR-ALS allowed for the individual signal recovery, besides a promising quantification of each amino acid. Figure 1 shows the results for (**A**) original normalized signals, (**B**) MCR-ALS normalized recovered signals, and fitting regarding concentrations (real vs. predicted by MCR-ALS) for (**C**) ADE, (**D**) ADO, and (**E**) AMP. This study encourages further exploration of the methodology in the study of amino acids.

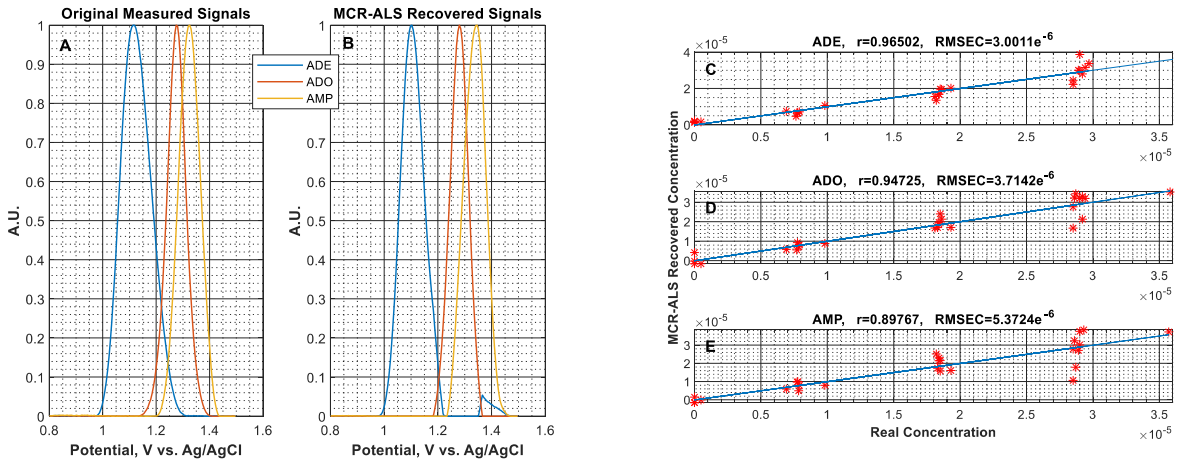


Figure 1 – (**A**) Original normalized signals, (**B**) MCR-ALS normalized recovered signals, and fitting regarding concentrations (real vs. predicted by MCR-ALS) for (**C**) ADE, (**D**) ADO, and (**E**) AMP.

References

[1] R. N. Goyal, A. Tyagi, *Nucleosides, Nucleotides & Nucleic Acids* **25** (2006) 1345-1362.
[2] M. Jauker, H. Griesser, C. Richert, *Angewandte Chemie* **54** (2015) 14564-14569.
[3] R. Tauler, B. Kowalski, S. Fleming, *Analytical Chemistry* (1993) 2040 – 2047.

On the Visualization of Bayesian Nonnegative Factor Analysis

N. Omidikia[1], J. Jansen[1], and R.Tauler[1,2]
1 Department of Analytical Chemistry, Institute for Molecules and Materials (IMM), Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
2 IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, Spain
Nemat.Omidikia@ru.nl

Factor analysis is still an evolving area of research dedicated to analyze data sets coming from various scientific disciplines [1]. Different algorithms have been proposed to componentize the measured data into physico-chemically meaningful profiles [2]. However, both accuracy and uncertainties of bilinear factor decomposition algorithms should be evaluated. Bayesian Non Negative Factor Analysis (BNFA) is a multivariate receptor modeling based on Markov chain Monte Carol (MCMC) method which is an attractive approach as it offers a great deal of flexibility in both modeling and estimation of parameter and specially in the estimation of the model uncertainty [3].

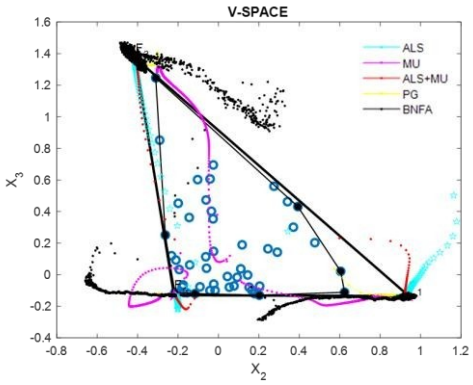


Fig. 1. Geometrical representation of the Area of Feasible Solutions (AFS) obtained in the investigation of an environmental data set and trajectrories followed by the different NMF optimization methods. The blue circles are the data rows and black triangle show the outer bondaies of data. BNFA drives the esttimations during the optimization iterations (MCMC samples,(black dots) toward the AFS and for the uncertainty calculations. Comparison of the onvergence tracks for various bilinear non-negative factor (NMF) decomposiution methods (ALS, NMF-MU, ALS+NMF-FU, NMF-PG and BNFA) is depicted.

In order to investigate the properties of the different NMF methods, a synthetic environmental three sources apportionment data set is used. Representation of the trajectories followed by the different NMF methods in the AFS plots provide a geometrical insight not only into the convergence properties of these different methods, but also on the microstructure of the investigated data set. In this work, BNFA convergence is investigated in detail to show how the uncertainty estimations are obtained in the successive MCMC iterations (see Fig. 1). On the other hand, identifiability conditions of the BNFA method can be investigated regarding the number and location of null elements in the factor profiles. Finally, the MCR-ALS, and non-negative matrix factorization (NMF) multiplicative update (MU) and projected gradient (PG) optimization methods are compared with the BNFA results obtained within the AFS for this three-component environmentasl source apportionment data system.

References

[1] Lee, D.D. and H.S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature, 1999. 401(6755): p. 788-791.
[2] Tauler R. Multivariate curve resolution applied to second order data, Chemomet. and Intell. Lab. Syst., 1995, 30, 133-146; 18.
[3] Sug Park E., Kyung Lee E., Suk Oh M., Bayesian multivariate receptor modeling software: BNFA and bayesMRM, Chemometrics and Intelligent Laboratory Systems 2021, 11, 104280.

Application of PARAFAC for curve resolution of fluorescence lifetime imaging data

N. Saburouh[1,2], O. Devos[1], P. Dedecker[2], C. Ruckebusch[1]
[1]*LASIRE, Univ. Lille, Villeneuve d'Ascq, France*
[2]*Chemistry Department, KU Leuven, Belgium*
nazanin.saburouhvahid@univ-lille.fr

Time correlated single photon counting (TCSPC) is a time resolved spectroscopy technique that has been used in solution studies to resolve mixtures of fluorophores with high spectral overlap [1]. The technique is amenable to fluorescence microscopy imaging (fluorescence lifetime imaging microscopy, FLIM) where the chemical contrast for each pixel is produced by the measured fluorescence decay. FLIM images are as a time resolved image [2]. However, unmixing of FLIM data is a difficult task due to the lack of selectivity, and multi-exponentiality of the decays. In this work, the application of a multilinear analysis (Parallel Factor Analysis, PARAFAC) [3] is investigated for curve resolution, i.e. identification and characterisation of the lifetime and spatial distribution of each fluorophore. We also consider the application of PARAFAC-Slicing[4].
To begin, a simple set of experiments was designed and ten solution mixtures of three dyes, (ALEXA-647, ATTO-655, and ATTO-665) were measured with TCSPC at 8 emission wavelengths. We analyzed the data using both PARAFAC and PARAFAC-Slicing, and show that exploiting the data obtained at any single emission wavelength would be sufficient here despite huge overlap of the profiles of individual contributions. Next, the autofluorescent daisy pollen was analyzed by FLIM [5]. The sample was excited by a pulsed laser at 485 nm and the fluorescent light emitted at 530 nm. Binning was used to preprocess the FLIM data in order to increase the signal-to-noise ratio. The two-way imaging data matrix was rearranged into a three-way array and PARAFAC slicing was applied. (Figure 1). The results were then compared with the ones obtained by a curve fitting approach.

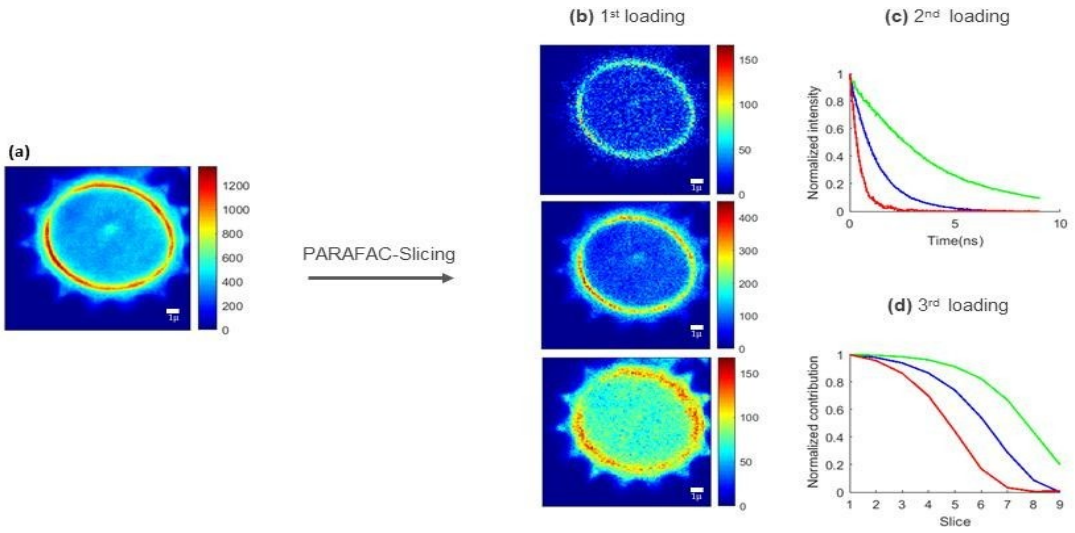


Figure 1 – A three-component PARAFAC-Slicing model of the FLIM data of a daisy pollen sample. (**a**) Mean image of the data, (**b**) distribution images of the components (2 layers of wall and 1 cytoplasm), (**c**) the corresponding decay time profiles, (**d**) the total number of slices and their associated contribution.

References

[1] W. Becker**,** *Advanced Time-Correlated Single Photon Counting Techniques*, **81** (2005)
[2] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, (2006) 740-752.
[3] R. Bro, *Chemometrics and Intelligent Laboratory Systems,* **38** (1997) 149-171.
[4] S. B. Engelsen and R. Bro, *J. Magnetic Resonance*, **163** (2003)192–197.
[5] C. Pohlker et al., Atmos. Meas. Tech., **6**, (2013)3369–3392.

UV ABSORPTION SPECTROPHOTOMETRY AND LC-DAD-MS COUPLED TO CHEMOMETRICS ANALYSIS OF THE DEGRADATION OF SULFAMETHOXAZOLE DRUG BY UV/CHLORINE ADVANCED OXIDATION PROCESSES

Aina Queral[1], Marc Marín[1], Sílvia Lacorte[1] and Romà Tauler[1]
[1]*Institute of Environmental Assessment and Water Research - Spanish National Research Council (IDAEA-CSIC), Barcelona, Spain*
aina.queral@idaea.csic.es

The presence of pharmaceuticals in surface waters and effluents due to human consumption, veterinary practice, or industrial activities is of concern. Sulfamethoxazole (SMX) is one of the most widely used antibiotics worldwide and has been detected at high concentrations in wastewater treatment plant effluents and river waters.

In this study, the SMX combined UV/Chlorine Advanced Oxidation Processes (AOP) are assessed and compared with the sunlight photodegradation [1] and the chlorine oxidation processes, each one individually and then also simultaneously. Photodegradation tests were performed using Suntest CPS equipment which simulated a sunlight exposure of the samples. Different experimental techniques, including UV-Visible spectrophotometry and liquid chromatography coupled to a diode array detector and positive and negative ionisation mass spectrometry (UPLC-DAD-MS), are proposed to evaluate the degradation reaction of SMX.

All the analytical data generated have been processed with the Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS) method [2,3] to monitor, resolve and identify the several transformation products generated during the studied degradation processes. A new data fusion analysis strategy is proposed to examine the three processes simultaneously (with only light, only chlorination, and simultaneous light+chlorination). Combined with the analysis of different analytical techniques individually, the fusion of all generated data improved the description of the degradation processes. Results obtained by the proposed procedure permitted the detection of all generated species and transformation products formed in each case and allowed the comparison of the effects produced by the two different oxidation processes.

References

[1] M. Marín-García, M. De Luca, G. Ragno, R. Tauler, *Talanta.* **279** (2022) 122953.
[2] A. de Juan, R. Tauler, Multivariate Curve Resolution-Alternating Least Squares for Spectroscopic Data, *Data Handl. Sci. Technol.* **30** (2016) 5–51.
[3] M. De Luca, G. Ioele, F. Grande, S. Platikanov, R. Tauler, G. Ragno, *J. Pharm. Biomed. Anal.* **186** (2020) 113332.

The potential of the ROIMCR methodology for swage water sample characterization in environmental proteomics

C. Pérez-López[1], D. Barceló[1,2], A. Ginebreda[1], M. Carrascal[3], J. Abian[3] and R. Tauler[1]

[1]*Institute of Environmental Assessment and Water Studies (IDAEA-CSIC), Department of Environmental Chemistry, Jordi Girona 18–26, 08034 Barcelona, Spain*

[2]*Catalan Institute for Water Research (ICRA-CERCA), Emili Grahit 101, Parc Científic i Tecnològic de la Universitat de Girona, Edifici H2O, 17003 Girona, Spain*

[3]*Biological and Environmental Proteomics, Institute of Biomedical Research of Barcelona, Spanish National Research Council (IIBB-CSIC/IDIBAPS), Rosellón 161, E-08036 Barcelona, Spain*

carlos.perezlopez@cid.csic.es

The Regions of Interest-Multivariate Curve Resolution (ROIMCR) methodology, recently proposed as a proteomic tool [1], combined with Partial Least Squares-Discriminant Analysis (PLS-DA) is shown for the analysis of environmental proteomics [2] samples.

Polymeric devices were placed 11 days in the influent water of the Gavà-Viladecans (Barcelona, Spain) wastewater treatment plant at 3 different times, between April and May 2020, during the quarantine period of the SARS-COV 2 pandemic. Slices from these polymeric devices were cut, purified, and digested with trypsin. Tryptic peptides (a triplicated at each sampling time) were analyzed by LC-HRMS/MS using an Orbitrap XL and the data obtained were exported to MATLAB computational environment and analyzed by the ROIMCR procedure [3]. A first step of filtering and compression of datasets was performed by the Regions of Interest (ROI) procedure, without losing mass accuracy. Multivariate Curve Resolution-Alternating Least Square (MCR-ALS) was applied to the ROI compressed data matrix resolving 181 'pure' components, explaining most of the experimental data variance (96.86%). Most of these components can be associated with peptide signals. The peak heights of the elution profiles of these components were then analyzed by Partial Least Squares-Discriminant Analysis (PLS-DA) [4], resulting in 41 MCR components (potential peptides) as being the main responsible of the observed differences among samples collected at different times.

A final identification step of peptides, and their protein inference, from the signals that belong to those 41 MCR components responsible for the main variability among samples was performed, and a number of them could be properly fragmented and identified. As a final result, a variety of proteins have been postulated to be present in the analyzed wastewater samples, such as those from immunoglobin domains, chaperonins, elongation factors, or ATP synthases belonging to different eukaryotic (e.g. human or mouse) and bacterial species. Further work is pursued at present using LC-HRMS/MS analysis of those target MS signals that were previously selected by the combination of the ROIMCR and PLSDA procedures. This data analysis will allow a more comprehensive identification of the peptide signals responsible of the variability observed among the different sampling times.

References

[1] C. Perez-Lopez et al. (2021) Non-target protein analysis of samples from wastewater treatment plants using the regions of interest-multivariate curve resolution (ROIMCR) chemometrics method. J. Environ. Chem. Eng. Volume 9, Issue 4, August, 105752.

[2] M. Carrascal et al. (2020) Discovery of large molecules as new biomarkers in wastewater using environmental proteomics and suitable polymer probes. Sci. Total Environ., 747 Article 141145.

[3] E. Gorrochategui et al. (2019) ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinforma., 20 (1), pp. 1-17.

[4] S. Wold et al. (2001) PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab Syst., 58 (2) (2001), pp. 109-130

Monitoring the State of Health (SOH) of green batteries (GreenBat)

Sandrucci E.[1*], Marini F.[1], Brutti S.[1]
[1]*Department of Chemistry, Sapienza University, Rome, Italy*
eugenio.sandrucci@uniroma1.it

The “GreenBat” project focuses on the study of state of health monitoring of secondary batteries with advanced “green” formulations. European recycle map on battery energy storage plans to replace positive electrodes consisting of cobalt (Co) and fluorinated (F) graphite-based negative electrodes constituents with environmentally sustainable alternatives, without compromises on performance and able to mitigate the environmental impact of production/recycling. Recently, "green formulations" of the already consolidated and widespread Li-ion technology (LIBs) have been validated; in particular, recent trends suggest the use of electrodes based on silicon (Si), Co-free manganites ($LiMnO_2$) aqueous soluble binders and fluorine-free electrolytes [1]. There is a consolidated and comprehensive *know-how* on battery SOH analysis through impedance spectroscopy (EIS) [2]. EIS is extremely flexible as it easily adapts to the innumerable possible formulations of the LIBs undergoing specific degradative mechanisms. SOH is defined as a quantitative descriptor of the performance of a battery in a given time of its use, compared to the data provided as *benchwork* by the manufacturer. In the GreenBat project, we want to get a description as detailed as possible of the chemical and morphological changes during the battery cycling, through spectroscopies (Raman, IR and fluorescence) and micro-spectroscopies operating in parallel with impedance analysis. These experimental data will be integrated with the electrochemical performance in a multiblock dataset, which will constitute the basis for chemometric processing. The purpose of this chemometric modelling is to integrate the SOH data estimated by electrochemical parameters, and correlate SOH evolution to the chemical and physical state of individual battery constituents [1,3]. Here we illustrate the analysis of SOH of a pouch cells formulation (100 mAh Customcell) constituted by LTO (Lithium-Titanate) as anode, LFP (Lithium-Iron-Phosphate) as cathode and the LP30 electrolyte (EC:DMC 1:1 and $LiPF_6$ 1m). After formation cycles (CC-CV), the cells have been submitted to galvanostatic charge/discharge cycles at C/2 at 30°C. After about 500 cycles, batteries show good performances both for coulombic efficiency and specific capacity (Figure 1a). Turning to the chemometric modelling, as a starting point we verified the ability of PLS regression applied to the voltage profiles during charge cycles to estimate the SOH of a benchmark dataset by Lin et al. [1] (Figure 1b).

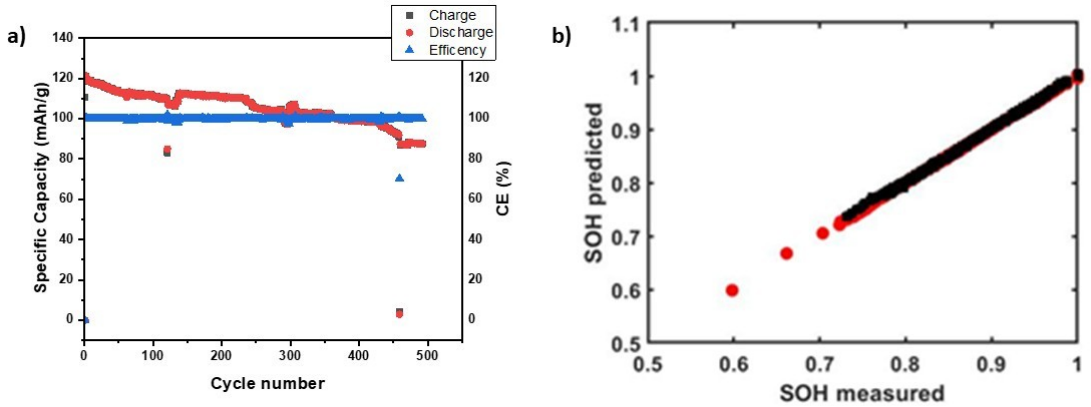


Figure 1: a) Specific capacity (mAh/g) and Coulombic Efficiency (%) during first 492 cycles; b) Results of PLS modelling of the benchmark data in [1]: the plot displays the comparison between predicted and measured SOH for the training (red circles, data from 5 cells) and the test (black squares, data from the remaining three cells) sets. For the test set, the value of R^2 is 0.9977 and RMSEP is 0.0034.

References

[1] M. Lin, D. Wu, J.Meng, J. Wu, H. Wu, A multi-feature-based multi-model fusion method for state of health estimation of lithium-ion batteries, *Journal of Power Sources*,2022, **518**.
[2] M. Chen, L. Zhang, F. Yu, L. Zhou, An Aging Experimental Study of Li-ion Batteries for Marine Energy Power Station Application, *Prognostics & System Health Management Conference—Qingdao*, **266100**, 2019.
[3] H. Rauf, M. Khalid, N. Arshad, Machine learning in state of health and remaining useful life estimation: Theoretical and technological development in battery degradation modelling, *Renewable and Sustainable Energy Reviews*, 2022, **156**.

SIMCA FRAMEWORK FOR MULTI-BLOCK CLASS MODELING

C. Scappaticci[1*], A. Biancolillo[2], F. Marini[1]
[1]*Department of Chemistry, Sapienza University, Rome, Italy*
[2]*Department of Physical and Chemical Sciences, University of L'Aquila, L'Aquila, Italy*
claudia.scappaticci@uniroma1.it

In this work, a new approach for model formulation and model building in the context of Soft Independent Modeling of Class Analogies (SIMCA), to overcome the limitations of the way it is normally implemented in the literature and used. This approach improves the performance of the SIMCA algorithm through the use of potential functions for non-parametrically estimating the probability density of the scores, at the same time taking into account the residuals of the model [1].
A further advantage of the proposed algorithm is that the new formulation of SIMCA can be easily applied to the case when dealing with more than a single block of data, thus allowing to apply, for the first time, a modelling classification technique also for so-called "multi-block" problems, i.e., where different matrices of data are used to describe and characterize the same sets of samples.
The proposed approach has been applied to three sets of spectroscopic data measured in the context of food authentication problems and involving the possibility of tracing the origin of different value-added products (PGI oranges from Sicily, a craft beer from Lazio [2] and the PGI Bell Pepper from Senise[3], respectively). In all cases, SIMCA models were built both considering a block of data at a time and through different strategies of "data fusion". The proposed modification introduced into the SIMCA algorithm has allowed to obtain models with increased sensitivity and specificity (so, also a better overall efficiency) with respect to those built following the conventionally used approach.
Moreover, as far as the multi-block analysis is concerned, the results evidenced on one hand that, if fusion occurs through a low-level approach (i.e., just concatenating the different data matrices) performances are comparable, if not even worse, to those of the best model built on a single block. On the other hand, when a mid-level fusion strategy was adopted, i.e., by concatenating principal components extracted from the individual data blocks, the integration of the information from the different matrices

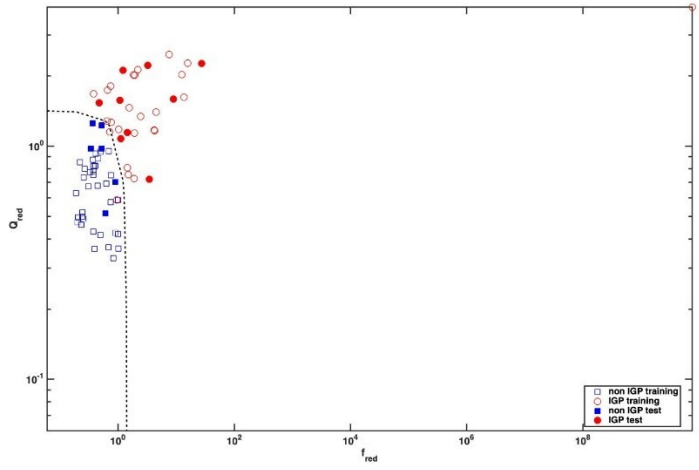


Figure 1 – Example of class modeling accomplished through the use of the modified SIMCA algorithm: Projection of the training and test samples onto the space of non-PGI samples for the orange multi-block data set.

References

[1] S. De Luca, R. Bucci, A. D. Magrì, F. Marini. In: R. Meyers (Ed.), Encyclopedia of Analytical Chemistry, Wiley, NY, 2018. https://doi.org/10.1002/9780470027318.a9578
[2] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, *Anal. Chim. Acta* **820** (2014) 23-31.
[3] A. Biancolillo, F. Di Donato, F. Merola, F. Marini, A.A. D'Archivio, *Appl. Sci.* **11** (2021) 1709.

ICP-OES ANALYSIS COUPLED WITH CHEMOMETRICS FOR THE CHARACTERIZATION AND THE DISCRIMINATION OF HIGH ADDED VALUE ITALIAN EMMER SAMPLES

F. Di Donato[1], G. Gornati[1], A. Biancolillo[1], A.A. D'Archivio[1]
[1]*Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio 67100, Coppito, L'Aquila, Italy*
alessandra.biancolillo@univaq.com

The elemental composition of 63 emmer samples produced in three different Italian areas, Monteleone di Spoleto (Umbria), Garfagnana (Tuscany) and National Park Gran Sasso-Laga (Abruzzo) were analyzed by Inductively Couples Plasma-Optical Emission Spectrometry (ICP-OES) combined with microwave-assisted digestion [1]. Recoveries were determined by the analysis of genuine and fortified samples spiked at different concentrations. The recoveries of minor elements: Ba, Cu, Fe, Mn and Zn, (1−50 µg/g concentration range) varied from 83% to 100%, whereas those for major elements (0.2− 5 mg/g): Ca, K, Mg and P, ranged from 85% to 98%. At first, Analysis of Variance (ANOVA) was applied to investigate the significance of the detected elements (Ba, Cu, Mn, Fe, Zn, Ca, K, Mg and P). Eventually, samples were divided into a training and a test set (of 33 and 30 objects, respectively) preprocessed (by log scaling and mean centering) and Partial Least Squares Discriminant Analysis (PLS-DA) was used to classify samples according to their geographical origin. PLS-DA highlighted that samples clearly group according to the geographical origin and the predictive model led to a correct classification rate of 100% for all the external samples. The investigation of the biplot (Figure 1) and the application of the Variable Importance in Prediction (VIP) analysis indicated that all the elements except Cu contributed to the characterization of the three different geographical classes. Briefly, the concentration of Fe is higher in samples from Spoleto while the emmer from Gran Sasso is richer in Zn. On the other hand, the class Garfagnana presents greater amounts of all the other elements.

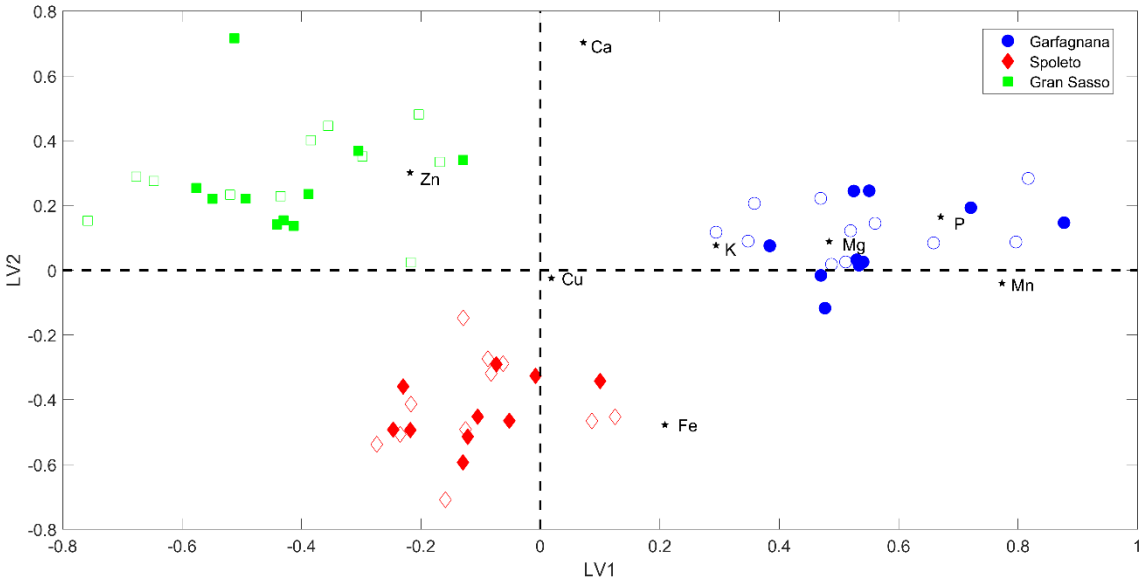


Figure 1 – Biplot. Legend: Class Garfagnana (Blue Circles); Class Spoleto: Red diamonds; Class Gran Sasso: Green Squares. Empty Symbols: Training samples; Filled Symbols: Test samples.

References

[1] F. Di Donato, G. Gornati, A. Biancolillo, A.A. D'Archivio, ICP-OES analysis coupled with chemometrics for the characterization and the discrimination of high added value Italian Emmer samples, *J. Food Compos. Anal.* **98** (2021), 103842.

GESTATIONAL DIABETES MELLITUS, PRETERM BIRTH AND MACROSOMIA: EARLY PREDICTION USING MULTIVARIATE ANALYSIS ON CLINICAL AND BIOCHEMICAL DATA

J. Araya[1-2], J. Appel[1], D. Mennickent[1-2], A. Rodríguez[2-3], E. Guzmán-Gutiérrez[1-2]
[1]*Universidad de Concepción, Concepción, Chile*
[2]*Machine Learning Applied in Biomedicine (MLAB), Chile*
[3]*Universidad del Bío-Bío, Chillán, Chile.*
jarayaq@udec.cl

In the last few years multivariate analysis has been started to penetrate biomedical sciences and just recently artifitial intelligence methods are being use to improve clinical diagnosis and disease prediction. In medical applications, clinical history records and blood tests such as biochemical profiles are routinely used as tools for decision making, however, most of the time only one or few analytes are used as predictors and no multivariate analysis is performed. In this scenario, the type of variables used are not too different from the data used in chemical, environmental or pharmaceutical sciences, except for the fact that the matrix is now the human body. Analytical chemists have been successfully using chemometrics for several decades to extract relevant information from chemical data, to find correlations or predict a sample property. Therefore, the robust chemometrical platform used in analytical chemistry for the analysis of this type of data could also be exploited in biomedical sciences.

The aim of this study was to predict different adverse pregnancy conditions (i.e. gestational diabetes mellitus [GDM], preterm birth [PB] and macrosomia) using obtetric and perinatal parameters commonly analyzed in pregnancy medical controls. Medical records of 71 pregnant women recruited from Concepcion (Chile), between 2015 to 2018 were analyzed retrospectively. Data include 22 obstetric and perinatal parameters from all the pregnancy, 8 of them from first trimester of gestation: maternal age (years), height (meters), weight (kg), BMI (kg/m2), basal glycaemia (mg/dL), hormone levels of: TSH (mIU/L), FT4 (ng/dL), TT4 (ug/mL) and TT3 (ng/mL).Principal Component Analysis (PCA) was used to explore unknown patterns among the data and Soft Independent Modelling of Class Analogy (SIMCA) was used for binary or multiclass assignment of the data. Finally, models were evaluated by their sensitivity (Se), specificity (Sp) and error rate (ER).
PCA analysis allowed to see differences and spontaneus separation tendencies for PB and macrosomia, however, for GDM samples two compact and completely separated clusters were formed vs healthy controls. SIMCA models allowed to predict PB and macrosomia with Se, Sp, and ER values of 1.0, 0.81 and 0.15 respectively for PB; and 0.88, 0.94 and 0.07, respectively for macrosomia. In both cases, although the evaluation of the models was favorable, class distance was low and most of the samples were located in the superposition zone. In contrast, SIMCA models for GDM showed compact and separated hyperboxes, allowing succesful prediction for this pathology.
The results showed that using obstetric and perinatal parameters from first trimester of gestation it is possible to generate a predictive model using chemometrics.The models generated used clinical information routinley collected during the first trimester of pregnancy without requiring additional clinical tests, image interpretation or trained staff. Although a bigger data set is required to do a properly validation, it seems multivariate analysis is a better alternative for prediction of negative outcomes in pregnancy.

Chemometrics-assisted microNIR spectroscopy for large-scale classification and authentication of high-quality Brazilian Canephora coffees

Michel Rocha Baqueta[1], Federico Marini[2], Enrique Anastácio Alves[3], Patrícia Valderrama[4], Juliana Azevedo Lima Pallone[1]

[1] *University of Campinas – UNICAMP, School of Food Engineering, Department of Food Science and Nutrition, Campinas, São Paulo, Brazil*

[2] *Department of Chemistry, University of Rome "La Sapienza", Rome, Italy*

[3] *Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Rondônia, Porto Velho, Rondônia, Brazil*

[4] *Universidade Tecnológica Federal do Paraná – UTFPR, Campo Mourão, Paraná, Brazil*

michelbaqueta@gmail.com

High-quality Brazilian Canephora coffees have become highly appreciated in Brazil and around the world, especially after two regions have received geographical indication (GI) [1]. As an innovative technology for the quality control of specialty foods, portable micro Near-Infrared (microNIR) spectroscopy has emerged to be cost-effective and extremely powerful to perform scientific measurements that would normally require advanced laboratory instruments [2]. In this study, mobile NIR spectroscopy was applied to identify and discriminate each Canephora producing origin in Brazil (Rondônia, Espírito Santo and Bahia), its botanical variety (Robusta or Conilon), to distinguish low-quality Canephora from specialty Canephora, and specialty Canephora from specialty Arabica. Principal Component Analysis (PCA) differentiated the samples according to coffee species and quality. Such changes were attributed to compounds (trigonelline, sugar, caffeine, etc.) that differ in Arabica and Canephora species and vary according to their quality. A 5-multi-class Partial Least Squares with Discriminant Analysis (PLS-DA) discriminated two Conilon classes, two Robusta classes, and one Arabica class. Binary PLS-DA discriminated GI Canephora coffees showing 100% sensitivity and specificity. Data-Driven Soft Independent Modeling of Class Analogy (DD-SIMCA) authenticated GI Canephora coffees with 100% sensitivity and specificity. These classification and authentication results were comparable and even superior to those of previous studies with coffee analysis [3–5]. While multi-class PLS-DA provides discrimination between origins, species, quality, and GI versus non-GI Canephora, DD-SIMCA provides GI versus non-GI Canephora authentication. The assignment of absorption bands related to chemical components of coffee supported the NIR analysis. A portable NIR can be a promising analytical technique when coupled to chemometric for quality control, certifying, and authenticating Canephora produced under GI specifications.

References

[1] C.A. de Souza, E.A. Alves, R.B. Rocha, M.C. Espindula, A.L. Teixeira. *Café Conilon*, 1st ed., (2021). 187–198.

[2] M.R. Baqueta, A. Coqueiro, P.H. Março, M. Mandrone, F. Poli, P. Valderrama, *Food Chem*. 355 (2021) 129618.

[3] M.R. Baqueta, A. Coqueiro, P.H. Março, P. Valderrama, *Talanta*. 222 (2021) 121526.

[4] R.C.E. Dias, P. Valderrama, P.H. Março, M.B. dos Santos Scholz, M. Edelmann, C. Yeretzian, *Food Chem*. 255 (2018) 132–138.

[5] J.V. Robert, J.S. de Gois, R.B. Rocha, A.S. Luna, *Food Chem*. 371 (2022) 131063.

CAC2022 will be an environmentally friendly event promoting social solidarity: thanks to the Food for Good project endorsement, we will collect the surplus food at the end of the Conference meals and deliver it to charitable organizations such as family homes, soup kitchens and refugee centers, in accordance with applicable hygiene regulations and in compliance with Italy's Good Samaritan law (Law 155/2003)

FOODFOR GOOD
from meetings to solidarity

SILVER SPONSORS

EIGENVECTOR RESEARCH

VIAVI

REGULAR SPONSORS

Kode
From Data to Knowledge

Metrohm

MathWorks

PICCIN

Journal of CHEMOMETRICS