

Book of Abstracts

Contents

Session 1: Chemometrics for process modelling/control/monitoring. Chair: Harald Martens

1. Frans van den Berg - Process chemometrics for dynamic systems
2. Geert H. van Kollenburg - Understanding chemical production processes through PLS path modelling
3. Noemí Marta Fuentes-García – PARAMO: Enhanced Data Pre-processing in Batch Multivariate Statistical Process Control
4. Ewa Szymańska – Embracing seasonal variation in milk composition in feed-forward control of cheese production process

Session 2: Spectroscopy. Chair: Alberto Ferrer

5. Ali Gahkani – t-SNE for Visualization of Spectroscopic Data
6. Shuxia Guo – Towards a Fast and Automatic Analysis of Fluorescence Lifetime Imaging Microscopy (FLIM) Data
7. Carl Emil Eskildsen – Correcting Inner Filter Effects in Fluorescence Measurements
8. Marta Bevilacqua – Front Face fluorescence and PARAFAC for fine interpretation of protein modification: how far can we go?
9. Andreas Baum – The potential of FTIR and PARAFAC–PCA analysis for quantification and subsequent comparison of enzyme activities originating from different origins
10. Nils Kristian Afseth – Hierarchical modeling in high-resolution spectroscopy – prediction of average molecular weights of protein hydrolysates using FTIR

Session 3: Chemometrics in the -omics area. Chair: Lennart Eriksson

11. Edoardo Saccenti - My first 15 years: learned lessons on critical steps in chemometrics applications to omics and systems biology
12. Jeroen Jansen - How to critically compare methods for resolving biomedical mixtures

Session 4: Deep learning, machine learning and chemometrics. Chair: Federico Marini

13. Ole Christian Lingjærde - Deep learning: past, present and future
14. Rickard Sjögren – Deep Learning – isn't it time for Chemometrics to embrace it?
15. Ulf Geir Indahl – The scikit-learn Data Science “pipeline approach” to Machine- (and Deep) Learning
16. Geert J. Postma – Deep learning for spectroscopic data analysis: an evaluation

Session 5: Chemometrics in action. Chair: Jens Petter Wold

17. Johan Trygg – Perspective on the application of multivariate technologies in biopharmaceutical manufacturing
18. Gerjen H. Tinnevelt – A novel unbiased method links variability of co-expression between multiple proteins on single cells to a clinical phenotype
19. Lars Munck – Natural Computing expressed in irreducible barley spectra reveal the functional composition in diagnostic fingerprints without compression
20. Giorgio Tomasi – Optim2DCOW: an algorithm for automated 2D Correlation Optimized Warping for GC \times GC – MS data

Session 6: PhD Projects. Chair: Ingrid Måge

21. Elise A. Kho – Characterization of *Haemonchus contortus* infections in sheep faeces by infrared spectroscopy
22. Raju Rimal – Simulation of multi-response linear model data and comparison of prediction methods
23. André van den Doel – Is river water out of control?
24. Silje S. Fuglerud – Aqueous glucose sensing by fiber-based near-infrared spectroscopy
25. George Stavropoulos – Data fusion strategies to improve prediction accuracy in Crohn's Disease
26. Anne Bech Risum – Multiway modelling of five-way protein fluorescence data; challenges and new approaches

Session 7: Path modelling, graphical modelling and causality. Chair: Jeroen Jansen

27. Rosaria Romano – University of Calabria, Italy: “Path modeling with multi-block regression method SO-PLS”

Session 8: Method Development. Chair: Age Smilde

28. José Camacho – Cross-Product Penalized Component Analysis: A new tool for Exploratory Data Analysis
29. Lennart Eriksson – Multiblock Orthogonal Component Analysis (MOCA) – A Novel Tool for Data Integration
30. Lars Erik Solberg – Consensus and distinct subspaces for blocks of distances
31. Kristian Hovde Liland – Fast “shortcut calculations” for cross validating Partial Least Squares prediction models
32. Raffaele Vitale – A novel procedure for the simultaneous optimisation of the complexity and significance level of SIMCA models in the presence of strong class overlap

33. Ryan Gosselin – A Novel Dynamic-PLS Algorithm for Meaningful and Robust Models
34. Erik Andries – Calibration Updating Using Unlabeled Secondary Samples

Session 9: Chemometrics in action. Chair: Barry Wise

35. Harald Martens – Big Data Cybernetics: Chemometrics and hybrid modelling for control theory
36. Federico Marini – A general SIMCA framework for single- and multi-block data
37. Chun Kiang Chua – Recent Development of Band-Target Entropy Minimization Algorithm for Hyphenated Techniques
38. Ingunn Berget – Sequential Clusterwise Rotations (SCR); a tool for clustering three-way data
39. Joan Borràs-Ferrís – Defining multivariate raw materials specifications via PLS model inversion
40. Anita Rácz – QSAR behind the curtains: best practices by multi-level comparisons
41. Jose M. González-Martinez – Energy Dispersive X-Ray Hyperspectral Imaging for Homogeneity Studies of Catalyst Extrudates

Posters

1. Lennart Eriksson – An OPLS®-based Multivariate Solver
2. Marian Kraus – Fast standoff investigation of chemical and biological samples using laser induced fluorescence signals, machine learning and an interactive interface
3. Andrei Barcaru – Chasing the interesting in the data with the Supervised Projection Pursuit
4. Ramin Nikzad-Langerodi – Domain Regularization in Partial Least Squares Regression: New Solutions for Old Problems
5. Dillen Augustijn – N-way Data Analysis of Protein Fluorescence in Formulation Screening
6. Kurt Varmuza – One-class classification for the recognition of relevant measurements – applied to mass spectra from cometary and meteoritic particles
7. Magnus Fransson – Applying Convolutional Neural Networks to Vibrational Spectroscopy Data
8. Rola Houhou – PCA – LDA in functional and discrete framework applied to Raman spectra

9. Alba González Cebrián – Dealing with outliers and missing data in PCA model building
10. José Camacho – Comparison of Sparse Principal Component Analysis for Data Interpretation
11. Robert van Vorstenbosch – The Detection of Colorectal Cancer using Exhaled Breath
12. Carl Emil Eskildsen – The cage of covariance: A consequence of regressing high dimensional response variables onto a lower dimensional subspace of explanatory variables
13. Tim Offermans – Improving process control of a dairy processing plant using a soft-sensor on parallel production data streams
14. Morten Arendt Rasmussen – One-Button Chemometrics
15. Agnese Brangule – Use of innovative FTIR spectroscopy sampling methods and chemometrics for authentication and differentiation of herbals
16. Johan Trygg – Data Fusion in metabolomics
17. Johan Trygg – Design of Experiments for data generation and data processing in ‘omics studies (genomics – metabolomics)
18. Johan Trygg – Multivariate patent analysis
19. Carlo G. Bertinetto – Effects of long distance walking analyzed by multidimensional flow cytometry analysis of neutrophils
20. Dávid Bajusz – Similarity metrics for binary data structures in cheminformatics, metabolomics and other fields
21. Veeramani Manokaran – Rapid identification of reaction systems using spectroscopic measurements and micro-reactors
22. Jacob Kræmer Hansen – Novel NIR analysis of Heterogeneous Powder
23. Mona Stefanakis – Infrared spectroscopy and multivariate data analysis for the label-free early stage diagnosis and demarcation of head and neck cancer in a mouse model
24. Roel Bouman – Process pls: A new path modeling algorithm for high dimensional and multicollinear data
25. Gavin Rhys Lloyd – Getting more from the PLS model: application to metabolomics
26. Gavin Rhys Lloyd – Statistics in R Using Class Templates (StRUCT)

27. Mercedes Bertotto – Detection of High Fructose Corn Syrup in Honey by Fourier Transform Infrared Spectroscopy and Chemometrics
28. Sumana Narayana – Mid-infrared spectroscopy and multivariate analysis to characterize *Lactobacillus acidophilus* fermentation processes
29. Ellen Færgestad Mosleth – Gene expression in petroleum workers exposed to sub-ppm benzene levels
30. Barry M. Wise - A Comparison of ANNs, SVMs, and XGBoost in Challenging Classification Problems
31. Mats Josefson - Experiments with complex numbered multivariate data analysis

Session 1

Chemometrics for process modelling/control/monitoring

Chair: Harald Martens

Process chemometrics for dynamic systems

Frans W.J. van den Berg

University of Copenhagen, Faculty of Science, Department of Food Science
Chemometrics and Analytical Technologies section
Rolighedsvej 26, DK-1958 Frederiksberg-C, Denmark
fb@food.ku.dk

Abstract:

Dynamics - both “good” or anticipated and “bad” or hidden - are at the core of any monitoring, control and optimization task in the processing industry. Whether yearly maintenance of equipment, regular optimization of production recipes or the machine related millisecond-scale periodicities present in a continues-tableting systems, proper interpretation and modeling of measurements and data requires an understanding of the engineering principles of the system, and thus the role of time. This insight can be utilized either explicitly via visualization or post-processing of the model parameters, or incorporated implicitly during model-building.

Chemometric methods have emerged as valuable tools in this area of research (and practice) as proven by the inclusion of factor models - such as PCA and PLS, including the powerful concept of principal or latent modes - in many engineering university curriculums and as established subject in process technology literature. Many modeling strategies from literature are however developed on clean, well-behaving, pure batch data or continues-production systems, and are at times lacking a practical fundament. Most operations in the processing industry e.g. run in a semi-continuous mode, with many hidden and unknown variations on top of a noisy signal.

In this paper I will present some of the unavoidable challenges based on examples from industry, and discuss some suggested solutions developed in our research group for working with real process data. I will cover both implicit and explicit use of the process characteristics, combining “soft model-builder” and “hard engineering” ideas. Via these examples I hope to convey my opinion that a valuable and realistic, working solution can be achieved by combining chemometric model building skills with process understanding.

Understanding chemical production processes through PLS path modelling

Geert H. van Kollenburg¹, Jan Gerretzen² & Jeroen J. Jansen¹

¹ Radboud University, the Netherlands. g.vankollenburg@science.ru.nl

² Nouryon; RD&I / Sustainability; Deventer, The Netherlands

An important aspect of the transition to industry 4.0 is that (chemical) production processes are closely monitored. The wealth of data obtained by measuring, among other things, pressures, temperatures and concentrations allows for a better understanding of the process itself. Incorporating substantive process knowledge into our statistical analyses may allow for a detailed understanding of the relationships between parts of a process. A natural choice to model relationships between various ‘blocks’ of a process is the PLS path model. Since in each block multiple variables may be measured, the PLS path modelling framework enables us to calculate the effects of each individual variable on the end-product, while retaining an interpretable process analytical model. Similarly, the PLS path model can aid in deciding on certain changes in a process to guarantee a required end-product quality.

The presented application entails the analysis of a continuous industrial process at Nouryon (previously AkzoNobel Specialty Chemicals). The data comprises approximately 17000 hourly process measurements of pressures, flowrates, concentrations and other process data. The production process itself requires a catalyst which degrades over time. A new batch of catalyst is installed when the previous one is depleted. The total yield from different batches of catalyst is highly variable and one of the main goals of the presented research is to understand this variation. Due to the nature of the production process, the variables can be easily grouped according to the separate parts of the process where they are measured. This multi-block approach is facilitated by the causal ordering of the distinct blocks in the production process. Due to having multiple batches, we can naively cross-validate model parameters across batches and relate differences in model estimates to catalyst lifetime.

PARAMO: Enhanced Data Pre-processing in Batch Multivariate Statistical Process Control

M. Fuentes-García¹, J. M. González-Martínez², G. Maciá-Fernández¹, J. Camacho¹

¹ Dpt. of Signal Theory, Telematics and Communications & CITIC-UGR, University of Granada - nmfuentes@ugr.es

² Shell Global Solutions International B.V., Shell Technology Centre Amsterdam

Since the pioneering works by Nomikos and MacGregor [1], the Batch Multivariate Statistical Process Control (BMSPC) methodology has been extensively revised and a sheer number of alternative monitoring approaches have been suggested. The different approaches vary in the batch data alignment, the pre-processing approach, the data arrangement and/or the type of model used, from two-way to three-way and from linear to non-linear [2]. One of the most accepted pre-processing schemes, referred to as Trajectory Centering and Scaling (TCS), is based on the normalization to zero mean and unit variance around the average trajectory [1]. However, the main drawback of TCS is the inherent increase of the level of uncertainty in the estimation of model parameters [2]. In this work, two main open questions are addressed: *i*) can the estimation of pre-processing parameters be improved, thereby reducing the parameter instability in the bilinear modeling of batch data? and *ii*) does the parameter stability have a statistical significant effect on fault detection?

We illustrate how to improve parameter estimation whilst maintaining the good properties of TCS. We propose a new pre-processing approach, PARAMO (PARAMeters from More Observations), which uses more observations than TCS to estimate the pre-processing parameters. We assess PARAMO and TCS by using the data generated from the *Saccharomyces Cerevisiae* cultivation process [3-4]. PARAMO outperforms the established methodology for pre-processing batch data in BMSPC. Using this proposal, both the parameter stability and the monitoring performance are improved. The results of this research work affect a large amount of the monitoring approaches proposed to date, and we advocate that the pre-processing procedure proposed here should be generally applied in BMSPC.

[1] P. Nomikos and J. F. MacGregor, “*Multivariate SPC Charts for Monitoring Batch Processes*”, *Technometrics*, vol. 37, no. 1, pp. 41–59, 1995.

[2] J. M. González-Martínez, J. Camacho, and A. Ferrer, “*Bilinear modelling of batch processes. Part III: parameter stability*”, *Journal of Chemometrics*, vol. 28, pp. 10–27, 2013.

[3] J. M. González, J. Camacho, and A. Ferrer, “*MVBatch: A matlab toolbox for batch process modeling and monitoring*”, *Chemometrics and Intelligent Laboratory Systems*, vol. 183, pp. 122–133, 2018.

[4] F. Lei, M. Rotboll, and S. B. Jorgensen, “*A biochemically structured model for Saccharomyces cerevisiae*”, *Journal of Biotechnology*, vol. 88, no. 3, pp. 205–221, 2001.

ACKNOWLEDGEMENT

This research work was partly conducted during a research stay at Shell Global Solutions International B.V. in Amsterdam (the Netherlands). The authors would also like to thank the Ministry of Economy and Competitiveness, and FEDER funding programs for partial financial support through the projects TIN2014-60346-R and TIN2017-83494-R.

Embracing seasonal variation in milk composition in feed-forward control of cheese production process

Ewa Szymańska, Frank Gielens, Emanuela Cavatorta, Ivo van der Zouwen, Tom van Hengstum

FrieslandCampina, Amersfoort, the Netherlands, ewa.szymanska@frieslandcampina.com

Seasonal variation in milk composition is a known but not yet fully characterized phenomenon which can strongly affect quality of dairy products. During cheese production different actions are being taken to respond and minimize seasonal changes in cheese and whey quality. However, often these actions are not taking milk composition changes into account and they are feedback control actions purely based on final product quality. In this project we have investigated the relationship between seasonal variation in milk, whey and cheese composition as well as process variables.

Firstly, we have collected and connected historical data over last 2 years including infrared spectra of milk, whey and cheese and process variables as temperatures, pressures, process step times etc. Secondly, in exploratory analysis, seasonal differences in each dataset have been investigated by univariate tests and principal component analysis (PCA) and linked together by generalized canonical analysis (GCA). Even though, seasonal trends are significant part of variation in each dataset, not always seasonal trends are well synchronized between datasets. Thirdly, different prediction models using sequential orthogonalization partial least squares (SO-PLS) have been built including milk, process variables and whey datasets. During model building process both cheese-making knowledge as well as variable selection with help of chemometric tools were used.

One of critical quality attributes of whey can be well predicted by infrared raw milk spectra and a small selection of process variables. Obtained model has been statistically validated and it can be deployed to give daily advise on process variables settings based on incoming milk spectra. Importantly, the seasonal variations in advised process settings are very well synchronized with seasonal variations in milk composition. Currently the model is being introduced at one of our cheese plants as an advisory tool for technologists. After successful validation by plant technologists it can be used as in an automated feed-forward control regime.

Session 2

Spectroscopy

Chair: Alberto Ferrer

t-SNE for Visualization of Spectroscopic Data

Ali Gahkani, Chris Piotrowski

Aunir (a division of AB Agri), Towcester, NN12 7LS, UK

Abstract:

The aim of dimensionality reduction (DR) is to preserve as much of the significant structure of the high dimensional data as possible into a low-dimensional map. PCA (Principal Component Analysis) is by far the most notable linear DR method. However PCA mainly focuses on retaining large pairwise distances, as opposed to small pairwise distances which are often more important in revealing local structures of a high dimensional data manifold. PCA embedding therefore has limitations with respect to revealing important within-cluster differences esp. when dealing with complex heterogeneous datasets. This is caused by domination of large (but irrelevant) sources of variability and or presence of outliers. t-SNE (t-Distributed Stochastic Neighbor Embedding) is an unsupervised nonlinear DR approach that has found widespread use in machine learning and big data applications in recent years. T-SNE has the advantage of allowing the implicit structure of all of the data to influence the way in which a subset of data is displayed [1]. t-SNE emphasizes on preserving the structure of important within cluster neighborhoods as well as between-cluster overlaps which can make it an indispensable tool for spectroscopic data. This method has proven to be very effective in optimization and validation of discriminant methods for NIR spectroscopy [2]. In this work, we will explore the concept and merits of t-SNE by using a large heterogeneous NIR dataset [3].

References:

- [1] Maaten LV, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9: 2579-605.
- [2] IDRC 2018 Shoot-out competition: <http://www.cnirs.org>
- [3] Miguel A., Gahkani A.S., An NIRS Prediction Engine for Discrimination of Animal Feed Ingredients, Chimie X VII, 2016, Namur, Belgium

Keywords:

PCA, t-SNE, nonlinear dimension reduction, clustering, classification, visualization, spectroscopy, NIR

Towards a Fast and Automatic Analysis of Fluorescence Lifetime Imaging Microscopy (FLIM) Data

Shuxia Guo, Tobias Meyer, Jürgen Popp, Thomas Bocklitz

Leibniz Institute of Photonic Technology Jena (IPHT Jena), Member of Leibniz Health Technologies, D-07745 Jena, Germany;

Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University of Jena, D-07743 Jena, Germany;

shuxia.guo@uni-jena.de

As an intrinsic property of a given fluorophore, the fluorescence lifetime (τ) is independent to the measurement parameters and insensitive to photobleaching. This makes fluorescence lifetime imaging microscopy (FLIM) a superior technique to conventional fluorescence intensity microscopy [1-3]. Nonetheless, the fluorescence lifetime is sensitive to changes in the local (micro) environment, such as pH, temperature, polarity and the presence of quenchers. Therefore, dynamic biological processes can be easily detected by FLIM with high spatial contrast. The great potential of FLIM has been already demonstrated in biological studies including intracellular protein-protein interactions [1-2] and the basic mechanisms of cancers [3]. The investigations of drug delivery and release in different pathologies, like diabetes, kidney and blood pathologies have benefited from FLIM as well [3]. All these applications require a deep understanding of the acquired FLIM signals and a well-designed data analysis protocol. The existent FLIM analysis approaches are roughly categorized as global fitting and graphical analysis [1]. Despite the wide application of global fitting, it is computationally expensive and its precision degrades dramatically for inhomogeneous samples like cells and tissues. Graphical analysis has gained more attention largely because it does not require any fitting procedures. While providing an excellent and intuitive visualization of FLIM data, graphical analysis falls short of quantitative analysis. In general, the FLIM data analysis remains an open issue, especially for the data with extremely low photon counts (i.e., high noise level).

In this contribution, we present a fast and automatic FLIM data analysis approach based on the mechanism of functional data analysis. The method was verified using a dataset measured on a time correlated single photon counting (TCSPC) system with by average one photon per temporal channel. In such highly noisy data, it is difficult to fit the decay curves correctly, i.e. to estimate the lifetime accurately. With the proposed approach, however, we could extract the lifetimes and amplitudes for each data voxel individually with a satisfactory accuracy. The extracted information can largely support subsequent data interpretation and medical diagnosis, if FLIM is utilized in a diagnostic manner.

Acknowledgements

The authors highly acknowledge the financial support of project Uro-MDD (FKZ 03ZZ0444J) funded by the alliance 3Dsensation.

References

- [1]. Borst, J.W. and Visser, A.J., *Measurement Science and Technology*, 2010. 21(10): p. 102002.
- [2]. Suhling, K., Hirvonen, L.M., Levitt, J.A., *et al.*, *Medical Photonics*, 2015. 27: p. 3-40.
- [3]. Berezin, M.Y. and Achilefu, S., *Chem. Rev.* 2010, 110: p. 2641-2684.

Correcting Inner Filter Effects in Fluorescence Measurements

Carl Emil Eskildsen, Katinka Dankel, Odin Øra, Tormod Næs, Jens Petter Wold

Nofima AS, NO-1433 Ås, Norway, carl.eskildsen@nofima.no

Fluorescence spectroscopy offers high-throughput multivariate measurements with high sensitivity and specificity. Fluorescence spectroscopy is used for measuring e.g. amino acids, vitamins and compounds like collagen, elastin, NADH and ATP. For these reasons, fluorescence spectroscopy is widely used within biological sciences in areas from medical diagnostics to process analysis measuring cellular reactions in biotech productions.

In order to obtain valid and robust calibration models (based on e.g. Partial Least Squares Regression), stable and linear relationships must exist between the multivariate measurements and the response. However, inner filter effects deteriorate fluorescence measurements and violate this linear relationship. This complicates calibration and limit the usefulness of fluorescence measurements obtained on complex biomaterials.

Inner filter effects appear as a two-step phenomenon. Primary inner filter effects appear as sample specific features absorbing part of the excitation radiation. This reduces the intensity of radiation left to excite the fluorophore. As a direct consequence, the emission intensity decreases (i.e. the magnitude of the emission spectrum decreases). Secondary inner filter effects appear as sample specific features absorbing part of the emission radiation. This will also reduce the emission intensity. Furthermore, secondary inner filter effects alter the shape of the emission peak (i.e. the direction of the emission spectrum changes). Altogether, inner filter effects disrupt the calibration between fluorescence measurements and the fluorophore of interest.

Theoretical models exist to correct inner filter effects in fluorescence measurements. However, these models do not work on highly absorbing samples. In this study, we develop a pragmatic approach to correct inner filter effects of complex and highly absorbing biomaterials. Front-face fluorescence measurements are obtained. The emission spectra contain information, deteriorated by the presence of inner filter effects, on the fluorophore of interest. Subsequent to fluorescence measurements, the samples are measured by reflection. Reflection measurements, in a wavelength region including the excitation wavelength, contain information on the primary inner filter effects. Likewise, reflection measurements, in a wavelength region similar to the emission wavelengths, contain information on the secondary inner filter effects.

To eliminate the secondary inner filter effects, the emission spectra are projected onto the null space of the space spanned by the reflection measurements, in a wavelength region similar to the emission wavelengths. Now, the projected emission spectra contain information, deteriorated solely by the primary inner filter effects, on the fluorophore of interest. The primary inner filter effects decrease the magnitude of the projected emission spectra. Reflection measurements, in the wavelength region of the excitation wavelength, is used to correct the magnitude of the projected emission spectra and thereby eliminate the primary inner filter effects. In that way, the corrected spectra contain information directly associated with the fluorophore of interest. Using this procedure, it will be demonstrated that predictions from corrected measurements have lower error than predictions from the original measurements.

Front Face fluorescence and PARAFAC for fine interpretation of protein modification: how far can we go?

Marta Bevilacqua, Therese Jansson, Åsmund Rinnan, Marianne Nissen Lund

University of Copenhagen, Dept. Food Science, Rolighedsvej 26, 1954, Frederiksberg C, Denmark (marta@food.ku.dk).

The use of front-face fluorescence spectroscopy, in combination with three-way chemometric analysis (like PARAFAC), has been largely reported in the literature for the analysis of samples without the need of any pre-treatment. In particular, in the field of food science, it has been used, for example, to classify foods undergone specific treatments or for traceability of Protected Designation of Origin (PDO) products. In the field of environmental chemistry, its use is largely and successfully applied for analysis of organic matter in water. Such good results reported in the literature, pave the way to the possibility of extending the use of this approach to reach a deeper understanding of the nature of food samples: to develop quantitative in situ methods for protein modifications in foods.

This innovative goal is based on the concept that the intrinsic fluorescence characteristic of the proteins (due to the three aromatic amino acids: phenylalanine, tyrosine and tryptophan) is sensitive to the local environment of the protein itself, and is therefore passible of changes in case the protein is modified in its structure or properties.

The use of front-face fluorescence (and three-way analysis of the signals by PARAFAC) seems an optimal choice for analyzing protein modifications as it can be used directly on food, thus being much easier and faster than the methods currently used to characterize such changes. This approach could answer the current lack of knowledge about how industrial food processing and storage affects the proteins present in the food and hereby its quality. To give light on this matter, indeed, it is important to detect and quantify the changes taking place to the proteins on a molecular level.

This requires the chemometrician to be able to accurately deconvolve many different fluorophores in a mixture and to be able to finely resolve very small differences and details in their spectra. Such challenging goal approaches the intrinsic limitation of PARAFAC. Mathematically, the PARAFAC method has unique solutions and clearly interpretable results only under very specific conditions: (almost) trilinear and low-rank data. These conditions are not inherently met in front-face fluorescence and this might hamper the fine interpretation of the peaks necessary to study protein modifications.

This work addresses these concerns. Front-face and right-angle geometries are used, in combination with PARAFAC, to study the behavior of the algorithm under different conditions of the samples, and accurately estimate the level of confidence that we can attribute to the PARAFAC loadings obtained in each case. This is done by means of ad-hoc designed samples containing known fluorophores together with different other absorbing and scattering molecules. Different strategies of spectral preprocessing are studied and compared by means of different applications both in models systems and in real food samples.

The potential of FTIR and PARAFAC–PCA analysis for quantification and subsequent comparison of enzyme activities originating from different origins

Andreas Baum, Valentina Perna, Heidi Ernst, Jane Agger, Anne Meyer

Assistant Professor, Section for Statistics and Data Analysis, Department of Applied Mathematics and Computer Science, Technical University of Denmark, andba@dtu.dk

Laccases are enzymes capable of catalyzing the oxidation of phenolic compounds using molecular oxygen as the electron acceptor. Depending on their origin the enzymes tend to react differently when exposed to a phenolic substrate. Up to date quantitative assays employ either substrate or product specific markers to monitor the enzyme activity, but do not allow for comparison of overall reaction patterns (univariate measure). In this study four different laccases: three fungal laccases from *Trametes versicolor*, *Trametes villosa* and *Ganoderma lucidum*, respectively, and one bacterial laccase from *Meiothermus ruber* were used.

We employed FTIR as a spectral fingerprinting technique to follow the enzymatic reactions in real-time. A FTIR spectrum can be understood as a snapshot representing the overall chemistry of a sample. When measured over the time-course of a reaction the spectral fingerprint changed as substrate(s) were consumed and product(s) were formed, resulting in a so-called spectral evolution profile. Several evolution profiles were measured using different laccase dosages leading to changes of the evolution intensity. Once obtained the evolution profiles were stacked to form a “data cube”, indicating enzyme dosage, spectral wave-numbers and reaction time along its modes.

The measurement routine was repeated for all four laccases and the resulting three-way tensors were decomposed using Parallel Factor Analysis (PARAFAC). As a result the PARAFAC scores were highly correlated with the enzyme dosage and, therefore, a suitable quantitative measure for the enzyme activity itself.

In a second step, the loadings from the spectral wavenumber mode from all PARAFAC models were used to perform Principal Component Analysis (PCA). This resulted in score plots which could be used to map the different enzyme origins and to interpret their differences with respect to the underlying reaction fingerprints.

The presented approach enables new possibilities within large-scale enzyme discovery. Nowadays thousands of genetically modified enzymes are evaluated as potential candidates for improved performance. Our results indicate that the methodology is capable of detecting significant deviations from a wild-type reference point-of-view and, hence, represents a valuable approach towards cracking the bottleneck of industrial enzyme activity assessment.

Hierarchical modeling in high-resolution spectroscopy - prediction of average molecular weights of protein hydrolysates using FTIR

Nils Kristian Afseth*, Sileshi Gizachew Wubshet, Kenneth Aase Kristoffersen, Ulrike Böcker, Diana Lindberg and Kristian Hovde Liland

* Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research, P.O. Box 210, 1431 Ås, Norway, nils.kristian.afseth@nofima.no.

High-resolution spectroscopic techniques like FTIR, Raman and NMR are providing new possibilities for detailed characterization, not only in research environments, but also for quantitative analysis in industrial settings. When aiming at obtaining generic calibration models across samples of different composition, the high chemical resolution might become a challenge, i.e. sample-dependent information can obscure the potential information from the feature of interest. This might in turn negatively affect calibration performance, and chemometric approaches to deal with this is needed. We encountered such challenge when we recently investigated the relationship between average molecular weights (M_w) and FTIR spectra in a total of 885 hydrolysates from enzymatic protein hydrolysis (EPH) of poultry and fish by-products. Monitoring EPH processes is an essential element of process control and the existing classical analytical methods are not easily applicable to industrial setups. It has recently been shown that FTIR is a potential technique to monitor protein sizes during enzymatic protein hydrolysis in industrial settings. From the FTIR spectra, it is apparent that protein hydrolysates originating from different raw materials and even different enzymes will display different FTIR fingerprints. Using PLSR, the coefficients of determination between FTIR spectra and M_w were reduced when samples of different raw material origins were combined in a single regression model. We therefore explored an alternative approach employing a two-level regression model, where the spectra were assigned to predefined groups consisting of known enzyme-raw material combinations in the first level using a supervised classification model (CPLS+LDA). On the second level, the spectra were subjected to a local regression model based on the specific enzyme-raw material classes. This type of hierarchical modeling proved very useful, providing better calibration results compared to the standard PLSR approach. In the presentation, the approach of FTIR-based hierarchical modeling will be discussed and compared to the standard PLSR approach. Further possibilities related to hierarchical modeling in high-resolution spectroscopy will also be discussed. The results suggest that hierarchical modeling of high-resolution spectroscopic measurements is an important approach that potentially will lead to increased use of these techniques in industrial settings for process monitoring and understanding.

Session 3

Chemometrics in the -omics area

Chair: Lennart Eriksson

My first 15 years: learned lessons on critical steps in chemometrics applications to omics and systems biology

Edoardo Saccenti

Wageningen University & Research, Wageningen, the Netherlands
edoardo.saccenti@wur.nl

The pioneering experimental work of Mamer and Horning (Horning and Horning 1971; Mamer and Crawhall 1971) and the first application by Pauling (1971) laid the bases for metabolomic profiling of samples. Contextually with experimental advancements, researchers realized that the potential of omics data could be exploited by deploying multivariate and pattern recognition methods. The use of components methods, such as principal component analysis and factor analysis was established early (Meuzelaar and Kistemaker 1973; Windig et al. 1980). Then, the analysis of omics data became rapidly intertwined in an almost symbiotic fashion with chemometrics. However, despite the introduction of advanced statistical tools, the pipeline leading from the formulation of a research hypothesis to the generation and verification of new results and knowledge is still disseminated by traps and bottlenecks. Paradoxically, the biggest limitations and problems reside in those steps that should have been the first to be addressed and evaluated because they are critical for our data understanding. In this presentation, based on my experience in the analysis of metabolomics in systems biology setting, I will present some results and idea on the FAIRification of chemometric analysis, data preprocessing, measurement error models and correlation estimation and on sparse principal component methods.

How to critically compare methods for resolving biomedical mixtures

Jeroen J. Jansen & Gerjen Tinnevelt

Radboud University, Department of Analytical Chemistry&Chemometrics Nijmegen, The Netherlands; jj.jansen@science.ru.nl

Translating analytical data into chemical ingredient lists is one of the key competences in chemometrics, specifically in the ever-increasing coverage of contemporary omics technologies. Most models of such mixtures are unsupervised, which makes their validation challenging—especially since this affinity with validation seems a very chemometric characteristic. As a result, also comparing between models is a challenge; most characteristics of a cluster or ordination model do not directly relate to its quality in recovering the biomedical information within. However, such critical comparison is of the utmost importance to define best practices for the end-user and to truly understand the quantitative representation of biological processes.

For Multicolour Flow Cytometry, several high-impact publications have introduced novel specific analysis methods to distinguish different types of white blood cells. Studies of equally high-impact have tried to critically compare these methods for recovery of different cell types, based on comparison of the algorithm performance to manually resolving cells into different populations. However, the criteria used for this comparison are very limited, compared to the biomedical information the end-user would need to interpret. Therefore, we have defined a series of criteria that cover many more aspects of the reconstruction. We use this list to compare methods based on a dataset about which much biomedical information is available as golden standard.

We also show the power of supervised analyses in resolving more cell types than with unsupervised methods alone. However, the criterion used for supervision needs to be evaluated critically. The results of sparse analyses for this need to be interpreted critically: broad-spectrum discoveries and optimal simplification may not go hand-in-hand.

Session 4

Deep learning, machine learning and chemometrics

Chair: Federico Marini

Deep learning: past, present and future

Ole Christian Lingjærde

Department of Informatics, University of Oslo

ole@ifi.uio.no

Deep learning has revolutionized computer vision and has shown great promise in a range of applications, including in such diverse areas as machine translation, self driving cars and computer games. We are likely to see many novel applications of deep learning in the years to come, including in chemometrics and in other fields where analysis of high-dimensional data with complex structure is central. In this talk, I will go back to the early roots of deep learning, starting with McCulloch and Pitts' seminal work in 1943. I will briefly revisit some major events in the history of neural network theory and the methodological developments that led to the transition from neural networks to deep neural networks. Why do deep networks work so much better than traditional neural networks? At the end of the talk I will give a few examples of deep learning applications in my own field of research and speculate a bit about what the future will bring.

Deep Learning – isn't it time for Chemometrics to embrace it?

Rickard Sjögren^{1,2}

1. Department of Chemistry, Umeå University, Umeå, Sweden

2. Advanced Data Analytics, Corporate Research, Sartorius AG.

rickard.sjoegren@sartorius.com

In recent years, Deep Learning has received a great deal of attention and is the main driver behind the ongoing hype surrounding Artificial Intelligence (AI). It has consistently demonstrated dramatic performance improvements in computer vision, speech recognition, and natural language processing.

In this presentation we present how deep learning can unlock new applications and expand the toolbox for Chemometrics by using *learned* features rather than *engineered* features. By using deep neural networks for feature learning rather than traditional feature engineering, a new paradigm in data science has emerged. We will also argue against the notion of Deep Learning models as black-boxes by recent advances in model interpretability, visualization and explainable AI. We will also highlight areas where Chemometrics may contribute to safer use of and less bias in Deep Learning models and applications.

The *scikit-learn* Data Science “pipeline approach” to Machine- (and Deep) Learning

Ulf G. Indahl, Kristian H. Liland and Oliver Tomic

NMBU/Faculty of Science and Technology, ulf.indahl@nmbu.no

Modern machine learning tools provide numerous powerful approaches to multivariate data analysis. The [scikit-learn](#) (*Machine Learning in Python*)-library is among the most attractive open source packages available for any purpose of application (commercial or non-commercial - as long as its copyright notices and the license's disclaimers of warranty are maintained).

We will give a brief “hands on” presentation of how to use various modelling methods implemented in [scikit-learn](#), as well as other Python tools, for modern data analysis purposes. Our demonstrations will include the pipeline-like and “easy-to-compare” modelling possibilities from the “Smörgåsbord” of modern data analysis methodology (ranging from PLS and Ridge Regression to Support Vector Machines, Decision Trees and Deep Learning).

The application programming interfaces (APIs) of the scikit-learn modelling objects are based on a relatively simple common API skeleton limiting the number of methods an object is required to include. It is therefore relatively simple to implement other methods (that are not available in scikit-learn) for a joint pipeline benchmarking.

The applications in our demonstration will cover model selection and -predictions for several classification and regression problems associated with both “tall” and “wide” datasets.

Deep learning for spectroscopic data analysis: an evaluation.

G.J. Postma¹, J. Acquarelli², T. van Laarhoven², E. Marchiori², L.M.C. Buydens¹, J.J. Jansen¹

1: Radboud University Nijmegen, Institute for Molecules and Materials

2: Radboud University Nijmegen, Institute for Computing and Information Science

g.j.postma@science.ru.nl

Over the last years, deep learning has gained popularity and relevance because of its effectiveness in classifying images and other multimedia data. Deep learning often uses Convolutional Neural Networks (CNN), composed of a large number of convolutional layers. These layers are the true novelty of CNNs; they facilitate the use of CNN for spectroscopic data sets which frequently are of limited size ('fat' data). Moreover, convolution with kernels of a fixed size allows exploitation of the spectral feature locality of spectroscopic data. One important drawback about ANNs from a chemometrics perspective is the limited interpretability of the predictive features in the data. We therefore have developed an embedded feature selection method to provide interpretation of the trained network. We also show that the purely data-driven analysis by CNNs may benefit from analytical chemical domain knowledge, through chemometric data preprocessing.

We have investigated the applicability of CNNs for a wide range of chemometrics case studies, such as the classification of beers, tablets and wines based on Raman or FTIR spectra, the prediction of brain tumours and Alzheimer disease based on Magnetic Resonance Spectroscopic Imaging data, and the classification of hyperspectral images. In case of the latter dependencies on the size of the training sets and methods to enhance the size of the training set were also investigated.

We compared performances of the CNNs with those of the more classical chemometric methods like PLS-DA. In case of hyper spectral images also with a true deep convolutional neural network. We show how CNNs often outperform such more conventional methods, such that embedding CNNs in the knowledge-supported and interpretation-driven approach of chemometrics may be very beneficial.

References:

1. Jacopo Acquarelli, Twan van Laarhoven, Jan Gerretzen, Thanh Tran, Lutgarde Buydens and Elena Marchiori, Convolutional neural networks for vibrational spectroscopic data analysis, *Analytica Chimica Acta* 954 (2017) 22-31
2. Jacopo Acquarelli, E. Marchiori, L. Buydens, T.N. Tran, T. van Laarhoven, Spectral-Spatial Classification of Hyperspectral Images: Three Tricks and a New Learning Setting, *Remote Sensing*, 10 (7) (2018) 1156
3. Jacopo Acquarelli, Twan van Laarhoven, Geert J. Postma, Jeroen J. Jansen, Arend Heerschap, Anne Rijpma, Sjaak van Asten, Elena Marchiori, Lutgarde M.C. Buydens, Convolutional neural networks for tumor severity and Alzheimer Disease prediction with Magnetic Resonance Spectroscopic Imaging data, in preparation
4. Jacopo Acquarelli, Twan van Laarhoven, Jeroen J. Jansen, Lutgarde M.C. Buydens, Elena Marchiori, Deep learning for spectroscopic data analysis: a critical assessment, in preparation.

Session 5

Chemometrics in action

Chair: Jens Petter Wold

Perspective on the application of multivariate technologies in biopharmaceutical manufacturing

Chris McCready², Olivier Cloarec², Johan Trygg^{1,2}

1. Department of Chemistry, Umeå University, Umeå, Sweden

2. Advanced Data Analytics, Corporate Research, Sartorius AG.

rickard.sjoegren@sartorius.com

Multivariate methods have become the technology of choice for monitoring and prediction of biopharmaceutical process performance. Many challenges presented themselves requiring the development of technological advancements to enable application in cell culture and chromatography. For example, hierarchical methods were introduced to handle multiple blocks of data resulting from multi-step process trains or combining of process and multivariate data sources such as spectral, tools for use of PAT instrumentation were developed, handling the alignment of events through the use of phases, time warping for dealing with variable batch lengths, strategies for combining data with varying sample rates, the development of advanced methods such as OPLS to focus on most important variability and more recently multivariate methods were integrated into a model predictive control framework to provide control of final batch quality.

In this talk a historical perspective on the challenges and solutions developed over the last 15 years of application of MVA in cell culture, chromatography and batch manufacturing systems in general is provided including case studies and a perspective of future and emerging technologies

A novel unbiased method links variability of co-expression between multiple proteins on single cells to a clinical phenotype

Gerjen H. Tinnevelt^{1,2}, Selma van Staveren^{2,3}, Kristiaan Wouters⁴, Bart Hilvering³, Rita Folcarelli¹, Leo Koenderman³, Lutgarde M.C. Buydens¹, Jeroen J. Jansen¹

¹Radboud University, Institute for Molecules and Materials, (Analytical Chemistry/Chemometrics), P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

²TI-COAST, Science Park 904, 1098 XH Amsterdam, The Netherlands

³Department of Respiratory Medicine and laboratory of translational immunology (LTI), University Medical Center Utrecht, Heidelberglaan 100, 3584CX, Utrecht, The Netherlands

⁴Dept. of Internal Medicine, Laboratory of Metabolism and Vascular Medicine, P.O. Box 616 (UNS50/14), 6200 MD Maastricht, The Netherlands

Email: gtinnevelt@science.ru.nl

Abstract: The immune system has evolved to combat a wide variety of infections, regulate healing processes, and displays many functions in homeostasis. This leads to an enormous complex cell mixture with a high cellular and functional diversity, which can be studied using flow cytometry. A typical multicolour flow cytometry sample may contain a large number of cells (>10,000), of which specific protein expressions are measured at the single-cell level. Individuals that exhibit an immune response, may have both changes in protein expressions on individual cells and changes in ratios of similar cells. Analysis of flow cytometry first requires to describe the cellular distribution in each individual and subsequently study systematic changes between groups of individuals. Our new data analysis method Discriminant Analysis of Multi-Aspect flow Cytometry (DAMACY) uses Principal Component Analysis to describe the cellular distribution and subsequently uses Orthogonal Partial Least Squares – Discriminant Analysis to show systematic changes. Thus, DAMACY merges and quantitatively integrates all the relevant characteristics on protein co-expression, the specific cells on which these are expressed, the distribution of these cells within all samples and the systematic change in this distribution upon changes in homeostasis of the host such as immune responses. The resulting model is comprehensively statistically validated to optimize the model information content and robustness.

We also show how DAMACY may be used to quantitatively integrate different multicolour flow cytometry tubes. The multiple tubes are needed because the number of proteins per measurement is technologically limited. This is unfortunate, as many high-impact studies reveal that interrogation of more proteins simultaneously leads to a more comprehensive view on the immune system. We show how data fusion of all tubes may find all the relevant cells in an activated immune states in obese versus lean people. The resulting model may not find the best biomarker, but will show how very different cells may function together in the individual development of type 2 diabetes.

Acknowledgements: This research received funding from the Netherlands Organization for Scientific Research (NWO) in the framework of the Technology Area COAST of the Fund New Chemical Innovations.

Natural Computing expressed in irreducible barley spectra reveal the functional composition in diagnostic fingerprints without compression.

Lars Munck

Chemometrics and Spectroscopy Group. Institute of Food Science Copenhagen University, Denmark lmu@ food.ku.dk

Scope: Natural computing inspired by Nature was observed in a designed near isogenic barley mutant/environmental model by perturbation. The result was visually observed as uncompressed irreducible chemical fingerprints carrying maximal information for functional global quality by Near Infrared Spectroscopy (NIRS). They were validated by proteome, amino acid, metabolite, carbohydrate and coarse chemical fingerprints. The molecular functions of the mutants were defined in literature. Combined together all these fingerprints in each individe represent *one deterministic globally coherent* fingerprint that defines each barley seed genotype. The aim is by relying on coherence in Natural Computing to reveal the uncompressed irreducible information on global functional quality by a diagnostic differential fingerprint to a certified marker. **Material and data:** Documentation on the barley mutant model was published in 15 papers since 2001.

Results: The aim is by selection by NIRS-MSC 2250-2400nm in a high lysine *lys3.a* mutant barley recombinant material to breed for improved high starch low fibre chemical composition. Compositional classification is compared between a PCA score plot on NIRS and diagnostic differential intact NIRS fingerprints by Natural Computing. The low starch, high fibre *lys3.a* mutants and its high lysine recombinants are situated to the right in the PCA with the high starch low fibre controls to the left. In between there are four starch *lys3.a* recombinants with improved starch from 47% to 52-54 %. Eight spectral positions covering the variation in the PCA were selected for *differential chemical spectral visualization of composition* with a normal high starch low fibre barley spectrum as a control. The degree of *linearization* of the spectrum from a tested line gave an accurate estimation of the desired optimal composition by a linearization index. There is a breeding response in starch from 47 to 54 % compared to 101 to 24 in the uncompressed spectral index. *Surprisingly it is now possible accurately to classify chemical composition in barley by just two chemical intact spectra one to test and one as a certified marker.* The deterministic precision in coherent chemical Natural computing was confirmed in an experiment over six years in two environments with two genetic identical barley lines with the NIRS reproducibility of stunning abs. $1-2 \times 10^{-4}$ at 1710-1810nm. **Implications:** The results from single seed Near Infrared Transmission Spectral sorters (Bomill A/S) confirm the functional informative capacity of Natural computing in sorting a wheat batch for baking and feed quality confirmed by chemical and pilot analyses. *Natural Computing precluding hidden compression by observation and experimental interpretation of intact visual chemical fingerprint-traits is now open for discovery and inclusion in the chemometric toolbox.*

Literature: L. Munck (2019) Introducing the era of NIRS-integrated functional fingerprinting in Proc.18th NIRS conf. <https://doi.org/10.1255/nir.2017.xx>. L. Munck, Å. Rinnan (2019) The Scientific Basis of a NIRS fingerprint as a Functional Trait for Coherent Chemical Information. Journal NIRS, second referee amended version submitted 15.2.2019.

Optim2DCOW: an algorithm for automated 2D Correlation Optimized Warping for GC \times GC – MS data

Giorgio Tomasi, Guilherme L. Alexandrino

Dept. of Plant and Environmental Sciences

University of Copenhagen

Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark

gito@plen.ku.dk / guialexandrino@plen.ku.dk

The Correlation Optimised Warping algorithm [1] was published 20 years ago and has since become a standard in chemometrics for the alignment of chromatographic data. The original algorithm has seen several improvements, e.g., to deal with high-dimensional mass – spectrometry data (CODA-COW [2]), or to align data from comprehensive two-dimensional gas chromatography – GC \times GC using the Total Ion Chromatogram (2D-COW [3]).

As the original COW, one important drawback of 2D-COW is that it requires the optimization of multiple meta-parameters; namely, the segment length ℓ and the so-called slack parameter s for both first (1D) and second (2D) chromatographic dimension to correct for retention time shifts in both columns. Skov *et al* [4] successfully addressed this problem in 1D through the optimCOW algorithm, in which a grid search for the two meta-parameters is followed by a simplex optimization and which uses a computationally efficient COW implementation.

In this presentation, we will introduce the optim2DCOW algorithm, an extension of optimCOW for 2D chromatography which can also handle multiple channels directly, together with some substantial improvements on the computational efficiency of the method that also apply to the 1D case. We will show how Optim2DCOW can successfully align GC \times GC data acquired with single-channel as well as multiple channel detection and how aligning multiple samples simultaneously can have significant computational benefits also for the basic 2DCOW.

We will present the results of optim2DCOW for two environmental studies based on GC \times GC – (high-resolution) MS chromatograms of i) diesel spills in marine environments, and ii) extracts from polluted areas of the Copenhagen municipality. In both cases, the 4-way data sets (sample \times retention time $^1D \times$ retention time $^2D \times m/z$) presented retention time shifts in both 1D and 2D that required correction prior to the data analysis. In particular, we will focus on the fast optimization strategy implemented here, as well as on its advantages when using chemometrics for environmental applications.

[1] Nielsen N.P.V. *et al.* J. Chromatogr. A, 805 (1998), 17 – 35

[2] Christin C. *et al.*, Anal. Chem., 80 (2008), 7012 – 7021.

[3] Zhang D. *et al.*, Anal. Chem., 80, (2008), 2664 – 2671.

[4] Skov T. *et al.* J. Chemometrics, 20 (2006), 484 – 497.

Session 6

PhD Projects

Chair: Ingrid Måge

Characterization of *Haemonchus contortus* infections in sheep faeces by infrared spectroscopy

Elise A. Kho¹, Jill N. Fernandes¹, Andrew C. Kotze², Maggy T. Lord³, Glen P. Fox¹, Anne M. Beasley⁴, Stephen S. Moore¹ and Peter J. James¹

1. The Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, Queensland 4067, Australia.

2. CSIRO Agriculture and Food, St. Lucia, Queensland 4067, Australia.

3. The School of Public Health, The University of Queensland, Herston, Queensland 4006, Australia.

4. The School of Agriculture & Food Sciences, The University of Queensland, St. Lucia, Queensland 4067, Australia.

e-mail: elise.kho@uq.edu.au

Gastrointestinal worm infections cause debilitating disease in sheep and goats worldwide. Heavy infestations of the Barber's pole worm, *Haemonchus contortus*, can cause severe wasting, morbidity and mortality in animals if not promptly treated [1]. The current detection methods for this blood-sucking parasite involve faecal worm egg counts and diagnosis of anaemia, both of which are time consuming and require expertise [2]. We investigated the use of infrared spectroscopy as a quick and easy alternative for detecting *H. contortus* eggs and the presence of blood in sheep faeces as an indirect measure of *H. contortus* infection.

We assessed haemoglobin (Hb) at various concentrations in sheep faeces using portable near-infrared (NIR) spectrometers (QualitySpec Trek and Felix F-750). Calibration models built within the 400 – 600 nm region by PLS-DA resulted in high accuracies ($R^2_{\text{prediction}} > 0.79$ and SEP $> 2.84 \mu\text{g Hb/mg faeces}$). Pre-processing with Savitzky-Golay smoothing further improved the prediction accuracies ($R^2_{\text{Pred}} > 0.81$ and SEP $> 2.75 \mu\text{g Hb/mg faeces}$). For the measurement of *H. contortus* worm eggs, we used a LabSpec 4 semi-portable NIR spectrometer in reflectance mode. We found that *H. contortus* eggs can be characterized in water within the 1880–2100 nm region. Similar chemical properties of dried *H. contortus* eggs were identified by mid-IR spectroscopy. However, when *H. contortus* eggs were combined with the complex matrix of sheep faeces, the development of a robust calibration model for eggs in sheep faeces proved challenging ($R^2_{\text{cal}} < 0.47$). This is the first infrared characterization of *H. contortus* eggs, which provides a baseline for future studies. Furthermore, our success in detecting Hb in sheep faeces indicates the potential of NIR spectroscopy as a rapid on-farm diagnostic method for blood in sheep faeces, which could have a range of veterinary applications.

[1] Besier, R., et al., Chapter Six-Diagnosis, Treatment and Management of *Haemonchus contortus* in Small Ruminants. Advances in parasitology, 2016. 93: p. 181-238.

[2] Preston, S.J.M., et al., Current status for gastrointestinal nematode diagnosis in small ruminants: where are we and where are we going? Journal of immunology research, 2014. 2014.

Simulation of multi-response linear model data and comparison of prediction methods

Raju Rimal, Trygve Almøy and Solve Sæbø

Norwegian University of Life Sciences (NMBU), Faculty of Chemistry and Bioinformatics
raju.rimal@nmbu.no

A linear model is a widely used relationship structure which we encounter. A versatile tool to simulate linear model data controlling various aspects of it can be useful not only for comparing methods, algorithms and models but also accessing and understanding their properties. Here we will present an R-package `simrel` that can control properties such as collinearity between the variables, information content in predictors that is relevant for responses and position of the predictor components that contain this information. With few parameters, the package can simulate data with a wide range of properties. Various multivariate prediction methods are developed over time to address the problem related to such properties.

In the second part of the talk, we will compare relatively established methods such as Principal Component Regression, Partial Least Squares Regression together with newly developed envelope methods using the multi-response data simulated with varying properties. These methods deal with the relevant and irrelevant regression structure in a different way. This part will present these differences and make some comparisons on these methods.

Is river water out of control?

**André van den Doel, Matteo Mastropierro, Geert van Kollenburg,
Gerard Stroomberg, Jeroen Jansen**

Institute of Molecules and Materials, Department of Analytical Chemistry, Radboud University, The Netherlands; TI-COAST, The Netherlands

Clean surface water is of vital importance to many aspects of life. Therefore, regulatory bodies closely monitor water quality. In the Netherlands, Rijkswaterstaat operates several monitoring stations, where multiple water samples per day are analysed. In current practice, warnings are sent out when any parameter exceeds its regulatory threshold. However, pollution events are rarely restricted to just one parameter. Industrial waste water or ship fuel that is (illegally) discharged often contains many different compounds for example. A multivariate approach should therefore allow for a more robust and specific detection of such pollution events.

We have applied Multivariate Statistical Process Control (MSPC) principles to the monitoring of water in Dutch rivers. We have used concentrations of organic chemical compounds that have been measured with purge-and-trap GC-MS. Based on these concentrations we have defined the normal range of river water quality. In terms of MSPC, water samples which fall outside of these 'normal operating conditions' are 'out of control'. Cluster analysis on these deviant samples was used to identify causes of the pollution (fault identification). With this approach we were able to identify several distinct types of pollution related to specific industries or events, giving us a new tool to find the perpetrator.

This research is part of the 'Outfitting the Factory of the Future with Online analysis' (OFF/On) project, which aims to provide innovative chemometric and statistical methods for monitoring chemical processes. Applications range from describing feedstock variability (surface water is also a source for the production of drinking water) to the multi-block modelling of a complete chemical production plant to ensure consistent end-product quality through a quality-by-design approach.

This research received funding from the Netherlands Organisation for Scientific Research (NWO) in the framework of the Programmatic Technology Area PTA-COAST3 of the Fund New Chemical Innovations. This publication reflects only the author's view and NWO is not liable for any use that may be made of the information contained herein.

Aqueous glucose sensing by fiber-based near-infrared spectroscopy

Silje S. Fuglerud^{1,2,*}, Karolina B. Milenko¹, Reinold Ellingsen¹, Astrid Aksnes¹, Harald Martens³, Dag R. Hjelme¹

¹ Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim NO-7491, Norway

² Department of Endocrinology, St. Olavs University Hospital, Trondheim NO-7030, Norway

³ Department of Cybernetics, Norwegian University of Science and Technology, Trondheim NO-7491, Norway

* silje.fuglerud@ntnu.no

An industry of glucose sensors and insulin delivery exists in pursuit of more effective treatments for diabetes mellitus (type 1). The goal is to avoid harmful effects of either too high (hyperglycemia) or too low (hypoglycemia) blood sugar levels, of which the first causes many long-term complications, including damages to the nervous system and sight, poorer general health and earlier death. The current glucose sensing gold standard is continuous subcutaneous measurements, where an electrochemical sensor is placed in the interstitial fluid. These sensors have good selectivity but suffer from stability problems, resulting in limited lifetime and the need for frequent calibration and a complete exchange every 1-2 weeks. As an alternative, optical glucose sensing could provide a long-term option for continuous glucose measurements.

This project is part of the Artificial Pancreas Trondheim (APT) group. APTs long term goal is to develop a robust closed-loop system where both glucose measurement and insulin infusion is administered in the peritoneum. Investigations of the dynamics in the peritoneum has indicated that the glucose concentration in peritoneal fluid is less delayed compared to blood glucose in interstitial fluid. Thus, the project seeks to measure aqueous glucose towards the goal of peritoneal glucose sensing. The focus of this PhD project is on the sensor design and subsequent data analysis of continuous glucose measurements to be further used in the control algorithms.

The current experimental setup is based on a 512-channel NIRQuest (OceanOptics) spectrometer spanning wavelengths 0.9-2.5 μm , targeting mainly the first overtone and combination band vibrations of glucose. Lensed optical fibers have been deployed for a flexible probe design without cuvettes and to increase the SNR. The current analysis protocol was developed on a larger dataset of glucose and interferents measured with a commercial spectrometer (Metrohm XDS, Foss Instruments) at Nofima. Extended multiplicative scatter correction (EMSC) is applied as pre-processing and builds a model with partial least squares regression (PLSR) from calibration samples. However, strong water absorption, weak signal and high detector noise is a challenge. By exploring the data analysis further, we seek to reduce the effect of some of these challenges.

Data fusion strategies to improve prediction accuracy in Crohn's Disease

George Stavropoulos¹, Frederik-Jan van Schooten¹, Jane Hill², Agnieszka Smolinska¹

¹*Department of Pharmacology and Toxicology, NUTRIM School of Nutrition and Translational Research, Maastricht University, Maastricht, The Netherlands*

²*Thayer School of Engineering, Dartmouth College, Hanover, USA*

g.stavropoulos@maastrichtuniversity.nl

Crohn's Disease (CD) is a type of Inflammatory Bowel Disease. CD can affect any part of the gastrointestinal tract, from the mouth to the anus. That is why CD may lead to debilitating or even life-threatening complications, and it usually requires heavy medication. The cause of CD remains unknown and exacerbated symptoms include severe abdominal pain, bloody diarrhoea, as well as high fever. Inflammatory stage of the disease is highly related to changes in human metabolome, volatilome, and gut microbiome. Affected and inflamed organs in the CD patients produce, and therefore, release Volatile Organic Compounds (VOCs) in blood and exhaled breath, as well as specific metabolites in plasma. At the same time, gut microbiome has an enormous influence on disease progression and development; thus, microbiome dysbiosis is always observed. Previously, analyses of VOCs in breath and gut microbiome, individually, have successfully differentiated CD patients in the active stage of the disease from remission CD patients with prediction accuracies of 80% and 82%, respectively.

In the present study, 130 breath, faecal, and blood samples were collected from CD patients while visiting an outpatient clinic, and they were classified into active (n = 65) and remission (n = 65) cases by using a combination of biochemical biomarkers and the Harvey Bradshaw index. VOCs in breath were measured by Gas Chromatography *time-of-flight* Mass Spectrometry (GC-*tof*-MS), while VOCs in blood headspace were measured by GC/GC-*tof*-MS. Metabolites in blood were analysed by Nuclear Magnetic Resonance, whereas Operational Taxonomic Units (OTUs) in faeces were assessed by 454 pyrosequencing (16S rRNA).

The present study aims to check whether the prediction accuracy for disease activity in CD patients can be improved by joint analysis of four different data platforms: VOCs in breath, VOCs in blood headspace, metabolites in blood, and OTUs in gut microbiome.

For that purpose, different fusion strategies were examined, compared, and evaluated: mid-level, high-level, as well as Multiple Kernel Learning (MKL), a specific case of fusion approach. Low-level fusion attempts were not made due to the enormous increase in data dimensionality. Random Forest (RF) and Gradient Boosting Trees (GBT) were applied to find the discriminatory features to be concatenated in the mid-level fusion case and to get predictions to be fused in the high-level fusion case. RF and GBT were also implemented in the MKL fusion attempt. All fusion strategies were evaluated based on their sensitivity and specificity to detect disease activity in CD patients. The fusion strategies, as mentioned above, demonstrated a comparable or improved prediction accuracy, and at the same time, correlations among all these data platform variables were found.

Multiway modelling of five-way protein fluorescence data; challenges and new approaches

Anne Bech Risum, Marta Bevilacqua, Marianne Nissen Lund, Åsmund Rinnan

University of Copenhagen, Dept. of Food Science, Rolighedsvej 26, 1954, Frederiksberg C, Denmark (anne.risum@food.ku.dk)

Current state of the art methods for characterization of protein modifications are laborious and slow, and risk introducing artifacts. The overall scope of my PhD is to develop a label-free fluorescence and chemometrics based methodology to characterize and quantify protein modifications directly in complex food samples.

Proteins are intrinsically fluorescent due to the presence of aromatic amino acids, of which tryptophan (Trp) is the dominant contributor. However, standard fluorescence studies based solely on excitation-emission matrices (EEM) of protein systems are too unspecific to connect signal changes to chemical modifications on a detailed level (i.e. Maillard reactions, oxidations, ligand binding sites etc.). We will solve this selectivity problem by increasing the analytical dimensionality. Fluorescence EEM measurements are inherently a multidimensional technique. Here, we extend this further by including a time decay mode, an anisotropy mode and a quencher mode, each increasing the selective information regarding the local environments around different Trp residues in the proteins.

To analyze the resulting complex five-way data array, we use multiway data analysis. Our base model is parallel factor analysis (PARAFAC), which has been widely applied for trilinear fluorescence EEM data. In the time decay mode, we will further increase the dimensionality by implementing the slicing technique, thus optimizing the resolution of multiple decays.

We will model the data in the intact higher order structure and thereby take advantage of the selectivity inherent herein. However, new challenges emerge when using PARAFAC on this type of higher order data. The multi-linearity condition may break due to interactions between the quencher and decay modes and in case of resonance energy transfer between residues. Furthermore, PARAFAC requires an equal number of components in all modes, which may not be a reasonable assumption for this data, as a single Trp fluorophore is expected to display multiple decay components. The focus of my talk will be on these modelling challenges and how to solve them, including a discussion of other possible multiway techniques such as PARALIND (PARALLEL Profiles with LINEAR Dependencies) and Tucker3 decomposition.

To illustrate these issues, data from fluorescence measurements on two different proteins (β -lactoglobulin and bovine serum albumin) will be shown. Both contain two Trp residues, and are considered relatively simple systems. As an example of a more complex sample matrix, a mixture of the two proteins will also be included. Chemical changes to the proteins will be induced by e.g. changing the pH, and the results will be compared to LC-MS results on the same samples.

Session 7

Path modelling, graphical modelling and causality

Chair: Jeroen Jansen

Path modeling with multi-block regression method SO-PLS

Rosaria Romano

Department of Economics and Statistics, University of Naples Federico II, Italy
rosaroma@unina.it

Abstract:

In many application fields the variables used to measure a phenomenon can be grouped into homogeneous blocks that measure partial aspects of the phenomenon. For example, in sensory analysis the overall quality of products may depend on the taste variables, the odor variables, etc. In consumer analysis, consumer preferences may depend on physical-chemical and sensory variables. In some contexts, there may exist a structure of relations between the different blocks that gives rise to a chain of influences. Within each relation, the blocks of predictor variables are called *input blocks*, while the block of dependent variables is called the *output block*. If the input blocks do not depend on any other block then they are defined *exogenous blocks*, while those that depend on other input blocks in the same relation are called *intermediate blocks*. If there is a chain of relationship between the blocks, we are then dealing with what is often called a *mediation model*, and must interpret both *indirect* and *direct effects* among blocks.

Within the scope of multiblock data analysis with a directional path among the blocks, we will present a new approach to path modeling named SO-PLS path modeling (SO-PLS-PM).

The approach splits the estimation up into separate sequential orthogonalized PLS regressions (SO-PLS) for each output block. The new method is flexible and graphically oriented and allows for handling multidimensional blocks and diagnosing missing paths. New definitions of total, direct, indirect and additional effects in terms of explained variances will be proposed, along with new methods for graphical representation.

In this presentation, some interesting properties of the method will be shown both on simulated and real data. The real data concerns consumer, sensory and process modelling data. Results will also be compared to those of alternative path modeling methods.

Session 8

Method Development

Chair: Age Smilde

Cross-Product Penalized Component Analysis: A new tool for Exploratory Data Analysis

J. Camacho⁽¹⁾, E. Acar⁽²⁾, M. Rasmussen^(3,4), R. Bro⁽⁴⁾

(1) CITIC, University of Granada (josecamacho@ugr.es)

(2) Simula Research Laboratory

(3) Dept. Food Science, University of Copenhagen

(4) Copenhagen Studies on Asthma in Childhood, University of Copenhagen

Matrix factorization methods are extensively employed to understand complex data. Principal component analysis is a key tool for that purpose. However, PCA is often difficult to interpret for two interrelated reasons, that have been addressed with a number of techniques in the literature:

- PCA is often difficult to interpret because the resulting PCs are linear combinations of all the variables. It is desirable to find factorizations that correspond to a limited number of original variables, so that they are easier to interpret. This can be achieved by means of rotation or sparse methods like sparse principal component analysis (SPCA).
- PCA does not distinguish between unique variance in each variable and shared variance, that is, the variance that is common among a group of variables. The result is that single factors may combine unrelated variability, which is a serious limitation for interpretation. A factorization method that focuses on shared variance rather than on any type of variance, like factor analysis (FA), solves this problem. Another approach is to constrain loadings to agree with the structure of the covariance matrix, like in the group-wise principal component analysis (GPCA).

In this contribution, we introduce a sparse matrix factorization approach (called cross-product penalized component analysis) based on the optimization of a loss function that allows a trade-off between variance maximization and structural preservation. The approach is based on previous developments, notably (i) the SPCA framework based on the lasso (least absolute shrinkage and selection operator) (ii) extensions of SPCA to constrain both modes of the factorization, like co-clustering or the penalized matrix decomposition (PMD), and (iii) GPCA. The result is a flexible modelling approach that can be used for data exploration in a large variety of problems, and we demonstrate its use with applications from different disciplines.

ACKNOWLEDGEMENT

This research work was partly funded by the Ministry of Economy and Competitiveness, and FEDER funding programs for partial financial support through the project TIN2017-83494-R.

Multiblock Orthogonal Component Analysis (MOCA) – A Novel Tool for Data Integration

Lennart Eriksson, Stefan Rännar, Rickard Sjögren & Johan Trygg.

Sartorius Stedim Data Analytics AB, Umeå, Sweden. lennart.eriksson@sartorius.com.

Multiblock Orthogonal Component Analysis (MOCA) is a new data analytics tool, which is used for fast and transparent analysis of multiple blocks of data registered for the same set of observations. In the case of two blocks, MOCA is similar in scope to O2PLS, but MOCA generalizes to situations involving any number of matrices without giving preference to any particular block of data.

MOCA extracts two sets of components; joint and unique components. More specifically, joint components express systematic structure found in multiple data blocks being analyzed. The joint components may either be:

- Globally joint, meaning that the same structures are found in all data blocks, or
- Locally joint, where the same structure is found in a subset of the data blocks.

Unique components express systematic structure that is only found in a single data block.

The objective of this contribution is to introduce MOCA to a wider audience and discuss its model structure and relationship to PCA and O2PLS. MOCA is available as a new tool in SIMCA® 16 multivariate data analysis solution. The benefits of MOCA are exemplified using three examples, the first involving modeling of a multi-phase chemical process, the second representing sensory analysis of a dairy product, and the third dealing with sequence-property descriptions of G-protein coupled receptors (GPCR).

Consensus and distinct subspaces for blocks of distances

Lars Erik Solberg(*), Tobias Dahl(x), Age Smilde(+) and Tormod Næs(*)

(*) Nofima, Norway. (+) University of Amsterdam, The Netherlands. (x) SINTEF, Norway.
lars.erik.solberg@nofima.no.

In multiple domains, the analysis of results concerns pairwise distances between samples for a set of methods or assessors. For instance, in psychology and sensory science, “projective mapping” is a technique where several assessors arrange objects on a table according to how similar they are judged to be. Another example is comparing methods for quantifying similarity between images of faces.

When analyzing this type of data, one is usually interested in relations among the samples. However, analyzing similarities and differences among assessors or methods may be just as relevant: what do assessors and methods agree about, and where do they differ?

Multidimensional scaling is a well-established tool, or set of tools, for representing a distance matrix using a set of points in some finite dimensional space. There are solutions within this tradition for handling sets of distance matrices: INDSCAL uses a model that accounts for such sets, while other approaches could analyze individual matrices and then perform a secondary analysis on the sets of points.

In this paper, we will compare different methods for modelling sets of distances matrices and emphasize how well they represent common and distinct parts. For this purpose, we will borrow concepts and methods from related multi-block analyses of regular matrices. An important aim of the present paper is to link the two areas, i.e. analysis of individual distance data and concepts of common and distinct components borrowed from multi-block analysis of matrix data.

We will use simulations as a tool to identify and understand properties of the different approaches because simulations allow full control of the actual common and distinct parts among methods or assessors. This will show how noise may affect approaches, what happens when assumptions about sizes of common and distinct parts do not hold, and how relative sizes of the common part versus distinct parts are handled by these approaches. We also aim to illustrate differences between approaches using examples from the literature.

The most important implication of the research is to be able to identify and interpret what the assessors agree and disagree upon and to quantify the relative variability of the two aspects.

Fast “shortcut calculations” for cross validating Partial Least Squares prediction models

Kristian Hovde Liland, Ulf Geir Indahl

Faculty of Science and Technology, Norwegian University of Life Sciences
kristian.liland@nmbu.no

Abstract

An alternative formulation of PLS for wide predictor matrices is proposed where only \mathbf{X} -scores and \mathbf{y} -loadings are calculated in the component estimation loop. The sample inner product matrix \mathbf{XX}' is calculated just once, prior to the component estimations, and requires no updating (all deflations are restricted to the response \mathbf{y}). Loading weights and \mathbf{X} -loadings can be obtained at low cost after the component estimation loop is completed.

Our approach is indeed more than “yet another theoretical exercise on PLS”, as it turns out to provide the possibility of conducting highly efficient cross-validation (of any type) by repeated use of the \mathbf{XX}' inner products through simple indexing and computationally “cheap” re-centring. A simple mathematical argument shows that explicit calculations of *loading weights* and \mathbf{X} -loadings are superfluous in the cross-validation calculations, as the predictions $\hat{\mathbf{y}}_{out} (= \mathbf{X}_{out}\hat{\boldsymbol{\beta}}_{in})$ can be found without explicit calculation of the regression coefficients $\hat{\boldsymbol{\beta}}_{in}$. The computational savings of our approach become even larger when the \mathbf{XX}' -inner products are utilized at both levels of the “double cross-validation” strategy (and across responses when multi-response problems are approached one response at the time).

We demonstrate the efficiency and stability of the proposed PLS formulation with simulated data in both single- and multi-response cases as well as for conducting “double cross-validation”. Finally, we demonstrate the potential of our approach for doing multi-class classification with a real data set (Prokaryote classifications of K-mers from 16S rRNA sequences).

A novel procedure for the simultaneous optimisation of the complexity and significance level of SIMCA models in the presence of strong class overlap

Raffaele Vitale, Federico Marini, Cyril Ruckebusch

Molecular Imaging and Photonics Unit, Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001 Leuven, Belgium
DyNaChem, Laboratoire de Spectrochimie Infrarouge et Raman – UMR 8516, Université de Lille, Bâtiment C5, 59655 Villeneuve d'Ascq, France
rvitale86@gmail.com

Nowadays, a large number of problems in fields like foodstuff origin authentication, quality control or process monitoring is addressed by Class Modelling (CM) statistical methods. Techniques such as UNEQual class modelling (UNEQ) or Soft Independent Modelling of Class Analogy (SIMCA) have been extensively used in the last decades for similar purposes. Contrarily to the more popular Discriminant Analysis (DA), the basic principle of CM is that classification rules are derived using only samples/objects belonging to a single target category. Faults in the definition of non-target categories, which could bias the classification performance, can thus be avoided.

Nevertheless, it is also well-known that if the classes under study present a high degree of overlap, CM approaches might suffer from severe limitations. In similar cases, properly adjusting the significance level of the resulting models can represent a potential solution to guarantee a better compromise between True Positive and True Negative rate. In this work, a new data-driven methodology that exploits the concept of Receiver Operating Characteristic (ROC) curve is proposed to address such a task. Its only requirement is that measurements for samples belonging to non-target classes are also available. Although this is actually not strictly needed in the CM context, it can be highly beneficial in all situations in which a significant overlap exists between categories. This presentation explores the potential of this procedure as a possible way of tuning SIMCA model parameters in circumstances like this. More specifically, an algorithm is here proposed that allows both complexity (number of principal components) and significance level of a SIMCA model to be simultaneously tuned through the construction of cross-validated ROC curves. It will be compared to a more standard procedure for tuning SIMCA model parameters which was described in and that is based on fixing such a significance level *a priori*. The performance of the two methodologies will be assessed in terms of classification sensitivity, specificity and efficiency in external validation in both simulated and real case-studies.

Two interesting points will arise from the analysis of the different handled datasets:

- in cases of clear and definite separation among classes, the two aforementioned methodologies enable a similar and equally satisfactory classification of unknown test samples;
- in the presence of strong overlap amongst classes, the implemented approach leads to better classification efficiency in external validation compared to the more standard procedure based on a fixed significance level.

Therefore, it can be said that adequately tuning the significance level guarantees a classification that is more robust towards the dispersion of the target category (i.e., a better compromise between classification sensitivity and specificity may be achieved).

Calibration Updating Using Unlabeled Secondary Samples

Erik Andries*, John H. Kalivas†

*Department of Mathematics, Science and engineering, Central New Mexico Community College, Albuquerque, New Mexico, USA; eandries@cnm.edu

†Department of Chemistry, Idaho State University, Pocatello, Idaho, USA

Calibration updating (transfer and/or maintenance) occurs at many levels: (1) adjusting the instruments; (2) adjusting the spectra using various spectral pre-treatments and transformations; (3) adjusting the calibration model (e.g., the regression vector); or (4) adjusting the final predicted results (e.g., bias and slope adjustments). Many practitioners reasonably insist upon a minimum set of established criteria (e.g., transfer or standardization samples) as a precondition for calibration transfer. However, as spectroscopic applications get more mobile, the migration from professional platforms (benchtop or hand-held) to home-based/DIY platforms (mobile phones, wearable tech, small/inexpensive NIR or Raman devices for consumer use) presents challenges for calibration updating. The primary challenge is that basic updating criteria often cannot be met for a variety of practical reasons: expense, added laboratory time, design of experiment, use-case scenario, etc.

We examine the scenario when one simply has a pool of labeled primary samples (primary samples with reference measurements) and a separate distinct pool of unlabeled secondary samples (secondary samples without reference measurements). This impoverished but all-too-common scenario rules out the vast majority of “go-to” calibration updating procedures from levels (1) through (4). We discuss level (3) algorithms and present encouraging results for this calibration updating scenario.

Session 9

Chemometrics in action

Chair: Barry Wise

Big Data Cybernetics: Chemometrics and hybrid modelling for control theory

Harald Martens^{1,2}

¹ Idletechs AS, Havnegata 7, 7562 Trondheim Norway

² Department of Cybernetics, Norwegian University of Science and Technology/NTNU, Trondheim NO-7491, Norway

* harald.martens@idletechs.com

How to deal with tomorrow's overwhelming, everlasting, high-dimensional torrent of technical measurements, in industry, medicine, economy, society? We need to prepare for a future where the wonderful new opportunities of Big Data in science and technology can be utilized, without people becoming alienated.

The chemometrics culture is very good at interpreting multichannel input data. But most of us are not so strong in differential equations, finite element models, graph theory and other aspects of industrial mathematics and systems theory.

The cybernetic control theory culture is very good at modelling dynamic systems with feedback, and also handles other industrial modelling techniques. But traditionally, it has had a preference for low-dimensional input data.

We see the need for better cooperation between these cultures.

The lecture will outline a Data Science culture with a distinct chemometrics/cybernetics focus on multivariate dynamic modelling in three main ontological domains of life: time, space and properties (e.g. wavelengths).

It will also summarize a new, versatile system for utilizing real-time streams of high-dimensional sensor data. This "Big Data on a Laptop" system provides interpretable hybrid modelling, combining prior mechanistic modelling, purely data-driven modelling and human interpretation in terms of tacit background knowledge. Industrial applications: Thermal and hyperspectral video, InSAR satellite data, spectroscopy, vibration analysis etc..

The main goal of this type of explainable AI is not to replace people by automata. It is to make good technology people even better, by giving them more natural, save and complete overview and insight, and fewer but better alarms.

A general SIMCA framework for single- and multi-block data

Alessandra Biancolillo, Federico Marini

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

federico.marini@uniroma1.it

Modeling classification techniques, sometimes also called one-class classifiers, have several advantages over discriminant ones, especially when dealing with asymmetric problems, where there is only one category of interest [1]. Indeed, in class modeling, attention is focused on a single category at the time, whose class space is built only on the basis of the data collected on samples from that particular group. Classification is then carried out as an outlier detection problem: if a sample is found to be an outlier with respect to the class model (usually, according to a distance to the model criterion), is predicted as not belonging to the category under exam. Among the methods available in the literature for class modeling, soft independent modeling of class analogies (SIMCA) [2] is by far the most commonly used. In SIMCA, the distance to the model is calculated by combining residuals with a distance in the scores space, which is usually Mahalanobis-like. When dealing with irregularly dispersed or, in general, moderately to highly heterogeneous classes, this may result in a shape of the model class space not corresponding to the actual one, so that high sensitivity can be achieved only at the price of low specificity and vice versa. In such situations, the use of a recently developed ROC-based approach to fine tune the classification thresholds [3] can help in finding the best model efficiency, but further improvements may be expected by redefining the way the class space itself is calculated. In the present communication, the possibility of defining the scores distribution non-parametrically by means of a gaussian mixture model (potential functions) is presented. Such approach allows a more-tailored definition of the class space even in the case of severe deviations of the distribution of the class scores from normality. Due to this property, this approach can easily be extended to the multi-block case in a framework which could be defined of mid-level data fusion.

The potential of the proposed approach will be illustrated by different examples involving food authentication both for the single- and the multi-block implementation.

1. C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, S. Wold, Four levels of pattern recognition. *Anal. Chim. Acta*, 103 (1978) 429-443.
2. S. Wold, M. Sjöström, SIMCA: a method for analyzing chemical data in terms of similarity and analogy. In: B. Kowalski (Ed.), *Chemometrics: theory and application*. ACS Symposium Series, vol. 52. American Chemical Society, Washington, DC, 1977, pp. 243-282
3. R. Vitale, F. Marini, C. Ruckebusch, SIMCA Modeling for Overlapping Classes: Fixed or Optimized Decision Threshold?, *Anal. Chem.* 90 (2018) 10738-10747.

Recent Development of Band-Target Entropy Minimization Algorithm for Hyphenated Techniques

Hua Jun Zhang, Chun Kiang Chua and Yunbo Lv

ChemoPower Technology Pte. Ltd., Singapore.

Email: george@chemopower.com

The implementation of Shannon entropy minimization in multivariate curve resolution approach for chemical data analysis has led to the development of a band-target entropy minimization (BTEM) method. BTEM has been applied to various chemical data such as ultraviolet, infrared and nuclear magnetic resonance spectroscopies, as well as mass spectrometry. This talk will explore the recent development, challenges, and applications of BTEM for hyphenated techniques data (i.e. gas/liquid chromatography-mass spectrometry (GC-MS & LC-MS) and liquid chromatography-photodiode array (LC-PDA)). In particular, I will discuss on the recent development of a new BTEM variant known as raw band-target entropy minimization (rBTEM). The rBTEM method enables the resolution of trace and co-eluting components without the need of any parameter settings or prior information about the system. The advantages of the rBTEM method over empirical modeling approach will be highlighted with several case studies.

Reference:

1. Chew, W., E. Widjaja, and M. Garland. Band-target entropy minimization (BTEM): An advanced method for recovering unknown pure component spectra. application to the FTIR spectra of unstable organometallic mixtures. *Organometallics* **2002**, 2, 1982-1990.
2. Zhang, H. J., M. Garland, Y. Z. Zeng, and P. Wu. Weighted two-band target entropy minimization for the reconstruction of pure component mass spectra: Simulation studies and the application to real systems. *J. Am. Soc. Mass. Spectrom.* **2003**, 14, 1295-1305.
3. Chua, C. K., B. Lu, Y. Lv, X. Y. Gu, A. D. Thng, and H. J. Zhang. An optimized band-target entropy minimization for mass spectral reconstruction of severely co-eluting and trace-level components. *Anal. Bioanal. Chem.* **2018**, 410, 6549-6560.

Sequential Clusterwise Rotations (SCR); a tool for clustering three-way data

Ingunn Berget, Quoc-Cuong Nguyen, Ingrid Måge, Paula Varela and Tormod Næs

Nofima AS, ingunn.berget@nofima.no

Three-way data occur frequently within chemometrics and sensometrics. In some cases, it is of interest to segment slices of the data sets for improved interpretation and precision and understanding of individual differences. Recently, Sequential Clusterwise Rotations (SCR) was developed for three-way data originated from the descriptive sensory method projective mapping (PM). In projective mapping, consumers are asked to organise a set of samples in a bi-dimensional map - a screen or sheet of paper- according to their similarities and differences. This results in a data cube, where each slice comprises two variables, the x- and y-coordinates for each product in each consumers' individual map. SCR is based on procrustes rotations to make maps within one cluster as similar as possible, combining this clustering criterion with the fuzzy C means methodology. Clusters are identified sequentially in such a way that the "best" cluster is identified first and shaved off before the next cluster is identified. This sequential procedure is repeated until the desired number of clusters is identified, or there are too few objects left in the data for meaningful cluster analysis. The methodology depends on parameters that define the threshold of what is a good cluster and the noise cluster in each step of the sequential procedure. Since SCR was developed for PM data which has a very special structure (N samples x M consumers x 2 variables), there is a need to test this methodology for other examples with possibly different structures. In this work, SCR will be demonstrated on data from PM experiments, other consumer data with multiple responses (liking, intake, eating rate), and microbiota data. For consumer studies the aim is to find groups of consumers with similar data slices, whereas for the microbiota, the purpose is to cluster OTUs (bacterial identification at genus level) with similar profiles for a set of individuals and conditions. The strategy of utilizing a sequential strategy for segmentation is not limited to SCR but could also be implemented for other clustering criteria both for two-way, three-way and multi set data. Directions for further research will be discussed.

Defining multivariate raw materials specifications via PLS model inversion

Joan Borràs-Ferrís¹, Daniel Palací-López¹, Carl Duchesne², Alberto Ferrer¹

¹ Multivariate Statistical Engineering Group (GIEM), Dept. of Applied Statistics Operations Research and Quality, Universitat Politècnica de València, València, Spain.

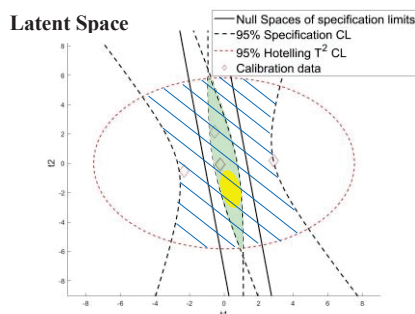
² Dept. of Chemical Engineering, Université Laval, Québec, Canada.

e-mail: aferrer@eio.upv.es


Developing meaningful multivariate specifications regions for raw materials properties is crucial to ensure that the desired final product quality can be achieved. This requires being able to infer causality through a model that explains how changes in raw materials properties (X) affect product quality (Y). Due to the complexity of many processes, though, either using first principles models or design of experiments may be unfeasible. This is why using Partial Least Squares Regression (PLS-Regression) has been proposed, since it allows using happenstance data to build causal models for both the X and Y spaces, while relating them [1]. Thus, by means of PLS model inversion, one may find a window in the latent space which guarantees the quality properties to be within specifications, i.e., the raw materials design space (DS). This point is in line with the Quality-by-Design initiative, which promotes the determination of the DS by science-based methodologies.


However, data-driven modeling is affected by uncertainty, which are backpropagated when a PLS model is inverted. Hence, a good estimation of the true DS relies on experimentation. Several papers have been proposed to segment the knowledge space (historical operating conditions) in such a way as to identify a subspace of it which is expected to contain the DS (referred to as the experimental space (ES)) [2]. Note that this delimitation of the ES as a subregion of the KS, and not the whole KS itself, greatly reduces the experimental effort required for a good estimation of the true DS.

In our proposal, we do not focus on the definition of the ES, but on the estimation of the DS for the raw material properties, where assurance of quality is expected with a high probability. Since uncertainty is not ignored, this space will presumably be slightly smaller than the true DS. Then, given the raw materials operating space (RMOS), i.e. the projection of their properties onto the latent space, it would be possible to check to what extent the RMOS overlaps with the estimated DS. Besides, this methodology allows assessing the adequacy of different suppliers by comparing the estimated DS with the RMOS for the materials provided by each one of them, the definition of multivariate specifications on the raw materials properties and to estimate a multivariate capability index.



 **Experimental Space:** high probability of containing the true DS.

 **Estimated DS:** high probability to provide assurance of quality, i.e. low probability of falling outside of the true DS (our proposal).

 **RMOS:** where we expect to operate given the properties of a supplier's raw materials.

[1] C. Duchesne and J. F. MacGregor, "Establishing Multivariate Specification Regions for Incoming Materials," *J. Qual. Technol.*, 36, 1, 78–94, Jan. 2004.

[2] P. Facco, F. Dal Pasto, N. Meneghetti, F. Bezzo, and M. Barolo, "Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development," *Ind. Eng. Chem. Res.*, 54, 18, 5128–5138, 2015.

QSAR behind the curtains: best practices by multi-level comparisons

Anita RÁCZ, DÁVID BAJUSZ, KÁROLY HÉBERGER

Research Centre for Natural Sciences Hungarian Academy of Sciences,

Magyar tudósok krt. 2, 1117 Budapest, Hungary

racz.anita@ttk.mta.hu

Quantitative structure–activity (toxicity, property, *etc.*) relationships have a paradoxical status in the field of computational chemistry. These modeling approaches can be dated back to the sixties, which means that they had sufficient time to be matured. However, the number of novel methods and alternatives for the sub-tasks associated with QSAR modeling are endless. A typical QSAR workflow encompasses various sub-tasks from the generation of molecular descriptors to the calculation of performance parameters.

Naturally, debates among the community have arisen due to the various and unambiguous statements about the used protocols; such as in the issue of internal *vs.* external validation, or in model selection based on the different performance parameters¹. Our aim was to find the most consistent alternatives in the various sub-tasks and collect the best practices for QSAR modeling, to make the entire process more robust and reliable. This entails providing the best (or most consistent) combinations of the common tools for different scenarios.

Multi-level comparisons based on commonly used alternatives were conducted, all along the long route of model building, involving many case studies for the activity or toxicity prediction of diverse ligands (molecules). Different intercorrelation limits, variable selection and modeling methods, validation tools and twenty performance parameters were tested and compared with the use of the sum of ranking differences (SRD) method and ANOVA. The SRD method is a novel and robust technique for comparison tasks² and, as we have demonstrated, for multi criteria decision making³ as well. Proper validation of the procedure was also implemented in the workflow.

Our results highlighted that even in the preparatory phase of QSAR modeling, proper molecular descriptor selection can play a significant role. Intercorrelation limits between 0.95 and 0.9999 are highly recommended for descriptor pre-selection. In the following steps, the modeling methods have a large influence on model building, and the best validation tools are usually method-dependent. On the other hand, the contiguous blocks cross-validation variant was clearly an outdated choice from the various opportunities. Machine learning methods such as SVM are less robust than the “old-timers”, like PLS or PCR, but they can easily outperform them. Finally, in the model selection phase, internal and external performance parameters could provide complementary information; external validation parameters were the most dissimilar from the consensus³.

1 P. Gramatica and A. Sangion, *J. Chem. Inf. Model.*, 2016, **56**, 1127–1131.

2 K. Héberger, *TrAC Trends Anal. Chem.*, 2010, **29**, 101–109.

3 A. RÁCZ, D. BAJUSZ and K. HÉBERGER, *SAR QSAR Environ. Res.*, 2015, **26**, 683–700.

Energy Dispersive X-Ray Hyperspectral Imaging for Homogeneity Studies of Catalyst Extrudates

J.M. González-Martínez¹, J.M. Prats-Montalbán², R. Haswell¹, C. Guédon¹, L. Ortiz-Soto³, A. Ferrer²

¹ Shell Global Solutions International B.V, Shell Technology Centre Amsterdam, PO Box 38000, 1030 BN, Amsterdam, the Netherlands

² Multivariate Statistical Engineering Group (GIEM), Department of Applied Statistics, Operational Research and Quality, Universitat Politècnica de València, Camino de Vera s/n Edificio 7A, 46022, Valencia, Spain

³ Shell International Exploration & Production, Shell Technology Center Houston N-2032-B, TX 77082 Houston, US.

The development of new catalyst requires a deep understanding of the active metal distribution at the level of the individual crystals. The combination of Scanning Transmission Electron Microscopy (STEM) and Energy Dispersive X-ray spectroscopy (EDX) is proposed to characterize the surface area of catalysts. The measurements are done by scanning the electron beam over the sample and at the same time measuring the X-ray spectrum – so-called spectral imaging [1]. The spectral dimension of the hyperspectral images consists of X-ray counts acquired at different energy channels. Chemical data is corrupted with noise, which comes from the time-dependent arrival of discrete particles on the sensor. This noise typically follows a Poisson distribution, which represents the probability of occurrence or events (X-ray counts) during a given period of time.

The application of Multivariate Image Analysis (MIA) techniques and Multivariate Curve Resolution (MCR) models becomes essential for the analysis of EDX hyperspectral images. This contribution proposes a modeling framework that permits segregating hyperspectral X-Ray images into simpler images (so-called Distribution Maps, DM's), which can be directly related to each of the chemical compounds present in the mixture [2]. From these DM's, chemical-textural score images (SI's) are further obtained. Finally, from all these DM's and SI's, different types of features, such as quantitative, morphological or textural can be extracted and combined into new data structures [3]. This new source of information is used to build multivariate statistical models for process understanding and prediction purposes at a MIA image-based level. This approach allows us to study similarities and differences between and within types of catalysts -i.e. different samples of the same catalyst across batches, and across locations, and the potential effects on quality properties of interest. To illustrate the chemometric framework for homogeneity studies, STEM-EDX images of real industrial catalyst will be used.

[1] R. Haswell, D. W. McComb and W. Smith, Preparation of site-specific cross-sections of heterogeneous catalysts prepared by focused ion beam milling, *Journal of Microscopy* (2003) Vol. 211, Pt 2, pp. 161-166

[2] J.M. Prats-Montalbán, A. de Juan, A. Ferrer, Multivariate image analysis: a review with application, *Chemometrics and Intelligent Laboratory Systems*, 107: 1-23, 2011.

[3] C Duchesne, JJ Liu, JF MacGregor, Multivariate image analysis in the process industries: A review, *Chemometrics and Intelligent Laboratory Systems* 117, 116-128, 2012.

Posters

An OPLS[®]-based Multivariate Solver

Lennart Eriksson, Stefan Rännar, Joakim Sundström, Johan Trygg and Jerker Öhman

Sartorius Stedim Data Analytics AB, Umeå, Sweden. lennart.eriksson@sartorius.com.

Two decades ago Jaeckle and MacGregor presented a multivariate approach for finding a window of operating conditions within which it would be possible to manufacture a product with a desired set of quality characteristics [1]. Their approach was based on using principal component regression (PCR) linking process input variables (X) to process output variables (Y).

In this contribution we present an alternative way to achieve a similar solver functionality, but which is based on the orthogonal partial least squares (OPLS) method [2]. This solver works for linear single-Y and multi-Y OPLS models. It uses the predictive components of the OPLS model in question to calculate which X-values give the desired Y-values. This means score values for suggested future observations will always be zero in the orthogonal components of the OPLS model.

The implementation in SIMCA[®] 16 Multivariate Data Analysis Solution, lets the user set target values for uncorrelated Y-variables. This ensures we get an exact solution from the solver. The user may distinguish between X-variables with constraints and without constraints. There must, however, be at least as many X-variables without constraints as there are with constraints. Additionally, in order to avoid unrealistic extrapolations far outside the validity domain of the OPLS model, it is possible to constrain the solver to observe restrictions in score space and residuals space.

The applicability of the new solver will be illustrated using datasets drawn from continuous manufacturing and quantitative structure-property modeling.

[1] Jaeckle and MacGregor, AIChE Journal, 1998.

[2] Trygg and Wold, Journal of Chemometrics, 2002.

Fast standoff investigation of chemical and biological samples using laser induced fluorescence signals, machine learning and an interactive interface

Marian Kraus, Lea Fellner, Florian Gebert, Carsten Pargmann, Arne Walter, Frank Duschek

German Aerospace Center (DLR), Lampoldshausen, Germany. marian.kraus@dlr.de

Release of hazardous substances may cause severe consequences to humans and infrastructure. A fast detection system classifying these substances can be used to initiate countermeasures quickly and reduce damage to general public significantly. Nowadays, the investigation of such materials is time and money consuming but essential for damage limitation. Current procedures take long and require methods that involve measurements at close range and/or sampling for subsequent laboratory analyses. Both, sampling time and distance, can be improved using standoff laser induced fluorescence (LIF) spectroscopy which enables detection in seconds reaching distances over 100 m. By now the specificity of the technique is not sufficient to identify samples but it can be helpful for risk assessment and to guide first responders to salient regions for subsequent in situ measurements.

This contribution presents an interactive graphical user interface as well as the practical workflow from generating training data and classification models to forecasting new records concurrently after the measurement. The foregoing modeling process is based on datasets generated previously with well-defined samples. Each sample from a considerable set of different chemical, botanical and bacterial substances can be distinguished using the LIF signals excited with short laser pulses of two UV wavelengths within a few seconds. Simultaneously, the fluorescence lifetime is recorded to provide additional information for a further enhanced discriminability. Estimating the sample species for new measurements consumes just a few seconds - including data acquisition, preprocessing, model application and visualization to the operator. As an example, the workflow is presented together with performance results for a test classification of 20 different substances.

Chasing the interesting in the data with the Supervised Projection Pursuit

Andrei Barcaru

Department of Laboratory Medicine, University Medical Center Groningen, PO Box 30001, 9700 RB Groningen, The Netherlands. Email: a.barcaru@umcg.nl

An important step in multivariate analysis is the dimensionality reduction, which allows for an easier visualization of the class structures in the data and sometimes a better classification. The techniques like PCA, PLS-DA and LDA dominated this side of data analysis for many decades. Yet the data does not always reveal properly the structures when these techniques are applied. To this end, a supervised projection pursuit (SuPP) is proposed, based on Jensen-Shannon divergence. The combination of this metric with powerful Monte Carlo based optimization algorithm, yields a versatile dimensionality reduction technique capable of working with highly dimensional data and missing observations. The method was successfully applied on 3 different data sets: classical Iris (I) and Wine (II) datasets, and gene expression adenocarcinoma (III) data set. For validation purposes, the data sets were split into training and validation sets, iterating through different fractions of this subsets and different random shuffles of the data. The classification accuracies obtained with SuPP-SVM, SuPP-NB, PCA-SVM, PCA-NB, PLS-DA and LDA were compared. Where SVM denotes Support Vector Machine and NB denotes Naïve Bayes respectively. In some cases, like in the case of Iris dataset, SuPP is capable to separate the classes in a lower dimensional latent space better than PLS-DA and PCA. Combined with Naïve Bayes classifier, SuPP becomes a powerful preprocessing tool for classification. The robustness of the algorithm to the shape of the distribution of the data and to the missing values makes SuPP a potentially useful tool for biological data analysis.

Domain Regularization in Partial Least Squares Regression: New Solutions for Old Problems

Ramin Nikzad-Langerodi

Research Center for Non-Destructive Testing (RECENDT GmbH) Linz Austria

Ramin.nikzad-langerodi@recendt.at

Domain regularization constitutes an emergent technique for tackling various challenges associated with multivariate calibration. It will be demonstrated that model adaptation between similar domains (e.g. instruments), calibration from incomplete data (i.e. semi-supervised calibration) and reduction of clutter (i.e. uninformative variation) can all be cast as constrained Partial Least Squares (PLS) problem that employs some form of domain regularization. Along these lines, it will be shown that several “old problems” in multivariate calibration can be addressed using a single algorithmic framework. The underlying theory will be discussed from a practical, algorithmic and learning theoretical perspective. Finally, different applications of domain regularization in PLS regression will be discussed on simulated and real-world data sets in order to underpin the versatility of the technique.

N-way Data Analysis of Protein Fluorescence in Formulation Screening

Dillen Augustijn, Alina Kulakova, Pernille Harris, Åsmund Rinnan

Department of Food Science, Faculty of Life Sciences, University of Copenhagen,
Rolighedsvej 30, DK-1958 Frederiksberg, Denmark. Dillen@food.ku.dk

INTRODUCTION. Fluorescence spectroscopy is a powerful tool in the evaluation of thermal stability of proteins. Protein unfolding introduces shifts in amino acid residue (Trp, Tyr, Phe) emission peaks, and thus these shifts become indicative of physical changes. Isothermal Chemical Denaturation (ICD) investigates protein stability by monitoring a denaturant induced unfolding. Various factors are introduced that may influence protein peak shape, such as: *Temperature, pH, buffer, excipient, ionic strength, and denaturant type and concentration.*¹

MATERIALS. A range of proteins have been investigated by ICD, with formulation at pH 5 to 9, salt at 0, 70 140 mM NaCl, various excipients, buffers and denaturants (GuHCl, Urea), to yield a large collection of high-quality fluorescence data. Protein fluorescence spectra are recorded at 300-500 nm (excitation at 285 nm).

DATA DECOMPOSITION. PARAFAC2² allows the modelling of sample variety in denaturant response curve under uniform protein fluorescence. The mode dictating denaturant based mixing of the fluorescence signals can thus be determined per sample, as shown in Figure 2 for some of the samples.

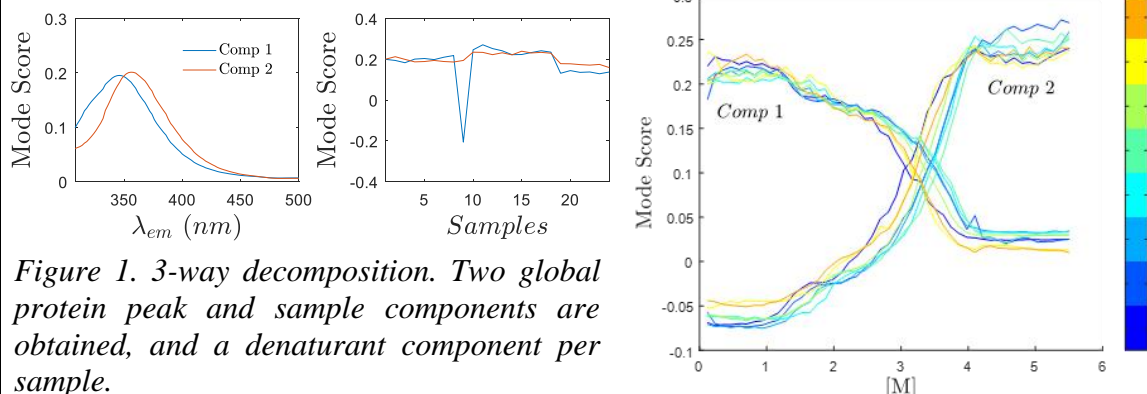


Figure 1. 3-way decomposition. Two global protein peak and sample components are obtained, and a denaturant component per sample.

RESULTS & CONCLUSION. We show how PARAFAC2 can be employed in the processing of ICD data. We further explore methods to estimate outliers that are commonplace for this type of measurements based on model component trends and errors.

ACKNOWLEDGMENTS. This study was funded as part of the EU Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreement No 675074.

REFERENCES

1. Garidel, P.; Hegyi, M.; Bassarab, S.; Weichel, M. J. B. J. H. N. T., A rapid, sensitive and economical assessment of monoclonal antibody conformational stability by intrinsic tryptophan fluorescence spectroscopy. **2008**, *3* (9-10), 1201-1211.
2. Kiers, H. A.; Ten Berge, J. M.; Bro, R. J. J. o. C. A. J. o. t. C. S., PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. **1999**, *13* (3-4), 275-294.

One-class classification for the recognition of relevant measurements - applied to mass spectra from cometary and meteoritic particles

Varmuza K.¹, Filzmoser P.¹, Ortner I.¹, Hilchenbach M.², Kissel J.², Merouane S.², Paquette J.², Stenzel O.², Engrand C.³, Cottin H.⁴, Fray N.⁴, Isnard R.⁴, Briois C.⁵, Thirkell L.⁵, Baklouti D.⁶, Bardyn A.⁷, Siljeström S.⁸, Schulz R.⁹, Silen J.¹⁰, Brandstätter F.¹¹, Ferrière L.¹¹, Koeberl C.^{11,12}

¹ TU Wien - Vienna University of Technology (Austria), Institute of Statistics and Mathematical Methods in Economics (Computational Statistics);
kurt.varmuza@tuwien.ac.at

Motivation. The mass spectrometer COSIMA on board of the ESA mission Rosetta to comet Churyumov-Gerasimenko (67P) collected particles (20 - 1000 μm diameter) at distances 10 - 1500 km from the comet and measured TOF-SIMS spectra at the particle surfaces. Because of the special conditions for these remote experiments, it is not trivial to assign the spectra either to particles or to the background (target). An objective classification of the spectra's origin (measuring spot 35 μm x 50 μm with position uncertainties up to 70 μm) has been developed by applying multivariate one-class classification strategies.

Method. The single class (target, background) for one-class classification is described by a set of multivariate objects (spectral data) measured on the target (gold). Two methods for modelling the target class are applied: robust PCA, and KNN. Criteria are defined for characterizing the dissimilarity (δ) between a query object and the target class: for robust PCA the orthogonal and the score distances from the median; for KNN the median of the distances to the k nearest neighbors. The cutoff values of δ for assigning a query object to the target class or not (the latter indicates a potentially relevant object) are derived from the distributions of δ for the target objects, based on median, 0.8-quantile and an adjustable parameter (controlling the efficiency of classification). Because of the nature of the data, concepts for compositional data and robust methods have been preferred.

Application. The data used consist of 275 spectra measured on three cometary particles, and 701 spectra measured by a laboratory twin instrument of COSIMA on particles from three meteorites (carbonaceous chondrites, often considered having similar composition as comet material). A set of nine variables is derived from the measured ion counts at masses 12-15 ($\text{CH}_{0.3}^+$), 24 (Mg^+), 27 (Al^+), 39 (K^+), 40 (Ca^+), and 56 (Fe^+) characterizing minerals and presumed organics. Results show distinctive differences between the cometary and the meteoritic samples with considerably more carbon containing material in the comet particles.

Affiliations of coauthors. ²Max Planck Inst. for Solar System Res., Göttingen (Germany); ³CSNSM, CNRS/Univ. Paris Sud, Univ. Paris Saclay, Orsay (France); ⁴Lab. Interuniversitaire des Systèmes Atmosphériques, Univ. Paris Est, Créteil (France); ⁵Lab. de Physique et Chimie de l'Environnement et de l'Espace, Univ. d'Orléans (France); ⁶IAS, CNRS/Univ. Paris Sud, Univ. Paris Saclay, Orsay (France); ⁷Carnegie Institution of Washington, DC (USA); ⁸Bioscience and Materials / Chemistry and Materials, Res. Inst. of Sweden, Stockholm (Sweden); ⁹European Space Agency, Noordwijk (The Netherlands); ¹⁰Finnish Meteorological Inst., Helsinki (Finland); ¹¹Natural History Museum, Vienna (Austria); ¹²Dept. of Lithospheric Res., Univ. of Vienna (Austria).

Acknowledgement. Austrian Science Fund (FWF), project P26871-N20.

Applying Convolutional Neural Networks to Vibrational Spectroscopy Data

Magnus Fransson, Esben Bjerrum, Mats Josefson

AstraZeneca Gothenburg (email:magnus.fransson@astrazeneca.com)

Introduction: The use of multivariate analysis in combination with spectroscopic data like near-infrared (NIR) and Raman has become a standard for non-destructive calibration of properties such as assay. Partial least squares (PLS) in its different variations has been the go-to method for this type of task for decades because it works well in most cases. Choosing an optimal spectral pre-treatment method can make a considerable difference to the accuracy and precision of the PLS predictions. However, finding this hypothetical optimal pre-treatment is often down to personal experience and preferences. Image analysis and classification models have seen a large increase in accuracy with the introduction of deep learning methods such as convolutional neural networks (CNN). A spectrum can be thought of as a one-dimensional image giving possibilities to apply the same neural network architecture for modelling of spectroscopic data.

Scope: Investigate the possibility of using CNNs to improve experience-based pre-treatment and calibration models using standard chemometric tools, in this case PLS and OPLS.

Results: In this study CNNs were applied to transmission Raman tablet data and in-line NIR powder data from a continuous direct compression manufacturing process of an immediate release tablet. The response was content of the active pharmaceutical ingredient. In both cases a CNN could be trained that gave about 5-15% lower prediction error compared to the experience-based combination of PLS and spectral pre-treatment.

Conclusion: The CNN works in a way that can be thought of as self-learning the optimal spectral pre-treatment. The CNN architecture needed 6-8 layers compared to the networks used for image data that sometimes needs over a hundred, this is likely due to that a spectrum in general carries less information than an image and the modelled property in this case has a more direct linear relationship to the information contained in the spectrum. The CNN, in contrast to a classical neural network, offers the possibility to investigate the trained network and draw conclusions about which wavelengths are used for the calibration in the same way as the loadings are used in a PLS model. In this study, the activated regions in the CNNs for the spectral data were very similar to the loading vector for the corresponding PLS model.

PCA – LDA in functional and discrete framework applied to Raman spectra

Rola Houhou^{1,2}, Jürgen Popp and Thomas Bocklitz

1. Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University Jena, Germany
2. Leibniz Institute of Photonic Technology Jena (IPHT Jena), Member of Leibniz Health Technology, Albert-Einstein-Straße 9, 07745 Jena, Institute of Photonic Technology Jena, Germany
rola.houhou@uni-jena.de

Raman spectroscopy allows the extraction of a molecular fingerprint of the sample, which is characterised by the superposition of its chemical components. During the detection process, the Raman spectra are recorded at discrete data points although the Raman effect is a continuous stochastic process. Accordingly, these Raman spectra should be converted into functions, which might contain more information. This is the core of the so-called functional data analysis that evaluates data in the form of functions [1]. Such a transformation requires *a priori* definition of a set of basis functions, from which a linear combination is constructed to estimate the discrete data. The characteristics of the sharp Raman peaks triggered the choice of the cubic B-spline basis functions; a piecewise polynomial approximation defined on selected knots. Thereafter, functional data analysis equivalents of known discrete analysis methods are applied and compared to their discrete counter-part. In our case, we utilized functional and discrete principal component analysis and linear discriminant analysis.

We demonstrated the application of functional data analysis for a bio-medical application of Raman spectroscopy. This case study was focusing on the classification of five cell types: leukocytes, erythrocytes, myeloid leukaemia cells (OCI-AML3) and two breast carcinoma derived cell lines (MCF-7, BT-20) based on their Raman spectra [2]. Initially, a cubic B-spline basis functions were approximated by means of the discrete Raman spectra of these cells. This functional data is of high or infinite dimension, which makes the use of dimension reduction tools very important. Therefore, a principal component analysis in the functional framework was applied to the continuous spectral data. Subsequently, a classification method via linear discriminant analysis was computed on selected functional principal components to discriminate between different cell types. At the end, a conventional principal component analysis followed by a linear discriminant analysis was performed. Both types of classification models showed similar performance identified by a similar mean sensitivity of 75.12% and 74.32% for the functional and discrete analysis, respectively. This performance can be further optimized by finding an appropriate basis functions for our spectral data.

Acknowledgements

The Deutsche Forschungsgemeinschaft (DFG) CRC 1076 “AquaDiva” is highly acknowledged for their financial support.

References

- [1] J.-L. Wang, Review of functional data analysis, *Annual Review of Statistics and its Applications*, 257-295, 2016
- [2] C. Beleites, Sample Size Planning for Classification Models, *Analytica chimica acta*, 25-33, 2013

Dealing with outliers and missing data in PCA model building

Alba González Cebrián, Abel Folch Fortuny, Francisco Arteaga, Alberto Ferrer.

Multivariate Statistics Engineering Group, Department of Applied Statistics and O.R. and Quality, Universitat Politècnica de València (UPV), València, Spain

In chemometrics, the model building task often involves data sets with missing values and potential outliers. The adaptation of Principal Component Analysis (PCA) model building to deal with missing data has been addressed in previous works [1]. Among different approaches, Trimmed Scores Regression (TSR) [2] performs extraordinarily well in terms of prediction quality and computation time, regardless the data structure and the percentage of missing data [1,3]. However, given the least squares nature of this procedure, the presence of potential outliers in the data set can have a critical influence on the missing data imputation. Consequently, further analysis based on the imputed data matrix will be affected as well by the presence of these outliers.

In this work we address the adaptation of TSR for PCA model building with missing data to deal with the presence of potential outliers. The proposed algorithm implements an iterative scheme for outliers' detection system over observations used to build the model.

Finally, in order to compare with other state-of-art methods which can deal with missing data and outliers in PCA model building contexts [4,5], a comparative study based on the Mean Squared Prediction Error (MSPE) is performed using real and simulated datasets, different missing data percentages and several structures of outlying values.

- [1] Folch-Fortuny, A., Arteaga, F., & Ferrer, A. (2015). PCA model building with missing data: New proposals and a comparative study. *Chemometrics and Intelligent Laboratory Systems*, 146, 77–88.
<http://dx.doi.org/10.1016/j.chemolab.2015.05.006>
- [2] Arteaga, F., & Ferrer, A. (2002). Dealing with missing data in MSPC: Several methods, different interpretations, some examples. *Journal of Chemometrics*, 16(8–10), 408–418.
<https://doi.org/10.1002/cem.750>
- [3] Folch-Fortuny, A., Arteaga, F., & Ferrer, A. (2016). Missing Data Imputation Toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems*, 154, 93–100.
<https://doi.org/10.1016/j.chemolab.2016.03.019>
- [4] Stanimirova, I., Daszykowski, M., & Walczak, B. (2007). Dealing with missing values and outliers in principal component analysis. *Talanta*, 72(1), 172–178.
<https://doi.org/10.1016/j.talanta.2006.10.011>
- [5] Hubert, M., Rousseeuw, P. J., Van den Bossche, W., & Bossche, W. Van den. (2018). MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, 1–18.
<https://doi.org/10.1080/00401706.2018.1562989>

Comparison of Sparse Principal Component Analysis for Data Interpretation

J. Camacho⁽¹⁾, A.K. Smilde, E. Saccenti, J.A. Westerhuis

(1) CITIC, University of Granada (josecamacho@ugr.es)

In modern data analysis there is an increased interest in sparse (component) methods for reasons of simplicity and interpretability. A popular of such methods is Sparse Principal Component Analysis (SPCA) [1,2]. The properties of SPCA have already been discussed elsewhere. However, there are some aspects that have not been treated in detail thus far and that are relevant for the interpretation of the models. One of them is in what space the scores and loadings of the successive components live. Whereas in (ordinary) PCA this is simple: they all live in the column and row space of the original matrix, this is not the case for sparse components. This means that sparse models cannot be entirely explained by the data they are calibrated from, and therefore there is the risk of creating artifacts that may confound data interpretation. Another relevant aspect is how scores, residuals and captured variance should be computed for sparse components, especially in those domains in which the interpretation of the scores is necessary, like in chemometrics. Moreover, the percentage of captured variance is often used as a criterion for model quality, and for the comparison of model variants. Therefore, its accurate computation is mandatory relevant issue.

In this contribution we review several extended SPCA algorithms [1-5] and compare them through simulation. We show that some have flaws in the computation of scores and captured variance, and how to correct for this. We also show that both scores and loadings can be correlated in SPCA, and the approaches, benefits and drawbacks of correcting for such correlation. Finally, we assess to which extent sparsity can create variance artifacts, harmful for interpretation.

- [1] Jolliffe I.T., Trendafilov N.T., Uddin M. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*. 2003.
- [2] Zou Hui, Hastie Trevor, Tibshirani Robert. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*. 2006;15:265-286.
- [3] Witten Daniela M., Tibshirani Robert, Hastie Trevor. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 2009;10:515-534.
- [4] Camacho José, Rodríguez-Gómez Rafael A., Saccenti Edoardo. Group-wise Principal Component Analysis for Exploratory Data Analysis. *Journal of Computational and Graphical Statistics*. 2017;26:501-512.
- [5] Sjstrand Karl, Clemmensen Line, Larsen Rasmus, Einarsson Gudmundur, Ersbll Bjarne. SpaSM: A MATLAB Toolbox for Sparse Statistical Modeling. *Journal of Statistical Software, Articles*. 2018;84:1-37.

ACKNOWLEDGEMENT

This research work was partly funded by the Ministry of Economy and Competitiveness, and FEDER funding programs for partial financial support through the project TIN2017-83494-R.

The Detection of Colorectal Cancer using Exhaled Breath

Robert van Vorstenbosch, Hao Ran Cheng, Zlatan Mujagic, Daisy Jonkers, Frederik-Jan van Schooten, Agnieszka Smolinska

Department of Pharmacology and Toxicology, NUTRIM School of Nutrition and Translational Research, Maastricht University, Maastricht, The Netherlands
R.vanVorstenbosch@maastrichtuniversity.nl

Colorectal cancer (CRC) is one of the most prevalent diseases within the EU. The survival rate of CRC patients is dependent on the stage at which the disease is diagnosed. To facilitate early diagnosis, national screening programs are organized. In the Netherlands this consists of a non-invasive iFOBT test and, if tested positive, a follow-up colonoscopy. Unfortunately, the iFOBT test suffers from low sensitivity and a high false positive rate. Therefore, a novel non-invasive diagnostics tool is urgently needed.

The analysis of Volatile Organic Compounds (VOCs) in exhaled breath might be a potential alternative. VOCs relate to host metabolism, gut micro bacteria, oxidative stress, and inflammation, which have previously been shown to correlate to pre-cancerous stages and colorectal cancer. Thus, the aim of this study is to demonstrate the feasibility of exhaled breath analysis to diagnose CRC and to compare its accuracy with the iFOBT test.

In this study, exhaled breath samples (and colonoscopy outcomes) were collected from patients tested positive for the iFOBT test (n=410) by inflating a 3L Tedlar bags and transferring the contents within 1h to stainless thermal desorption tubes where volatile metabolites were trapped. Later, they were analyzed using Gas Chromatography -time of flight- Mass Spectrometry (GC-tof-MS). Before the actual statistical analysis, the data was pre-processed using wavelets and p-splines to diminish effects of noise and baseline. After aligning the chromatograms, they were normalized by probabilistic quotient normalization.

The statistical analysis consisted of several steps. First, data was visualized using PCA, robust-PCA and unsupervised Random Forest. In the consequent step, the data was randomly divided into training (70%) and test set (30%). In order to find the CRC-specific volatile metabolites in exhaled breath, feature extraction was implemented in combination with linear and non-linear classification models including PLS-DA, tree based techniques and SVM. The prediction accuracy of each classification technique was accessed by means of precision-recall curve. This study shows that a specific set of volatile metabolites in breath enables the differentiation among CRC, pre-stages of cancer and healthy patients. Thus, it may prove to be of crucial importance in patients' diagnosis and treatment processes. Exhaled breath analysis can therefore be a potential new non-invasive diagnostic and monitoring tool.

THE CAGE OF COVARIANCE: A CONSEQUENCE OF REGRESSING HIGH DIMENSIONAL RESPONSE VARIABLES ONTO A LOWER DIMENSIONAL SUBSPACE OF EXPLANATORY VARIABLES

Carl Emil Eskildsen

Nofima AS, NO-1433 Ås, Norway, carl.eskildsen@nofima.no

When regressing a response onto explanatory variables, the estimated response will be in a subspace of the explanatory variables. This is important to consider when regressing multiple response variables onto the same explanatory variables. If the response variables are partly correlated, the estimates may end up in the same subspace of the explanatory variables. As a result, the estimated responses will be linearly dependent, even though the true response variables do not have a causal relationship. Consequently, the relationship between the estimated response variables in future data sets will be locked and defined by the calibration data. Hence, future estimated response variables are forever caught in a *Cage of Covariance* with each other.

To avoid linear dependencies among the estimated response variables, both the *true* response variables and the estimated response variables must have full rank. If linear dependencies exist among the estimated response variables (i.e. they are estimated from the same subspace of the explanatory variables), then the rank decreases.

This study estimates multiple fatty acids from vibrational spectroscopic measurements obtained from three samples sets 1) Salmon muscle tissue, 2) Porcine adipose tissue and 3) bovine milk. Individual partial least squares regression models are constructed to predict the fatty acids. All fatty acids are well predicted when evaluated by a mean squared error between the *true* values (obtained by gas chromatography) and the predicted values.

Nevertheless, the rank of the *true* fatty acids is remarkably higher than the rank of the predicted fatty acids. This is observed in all 3 data sets. This indicates that fatty acids estimated from vibrational spectroscopic measurements are trapped in a *Cage of Covariance*.

The *Cage of Covariance* highly compromises validity and robustness of the calibration models. Viewing these fatty acid predictions as independent, while in fact they are not, may lead to serious misinterpretation of the studied system.

Improving process control of a dairy processing plant using a soft-sensor on parallel production data streams

Tim Offermans¹, Ewa Szymanska², Jeroen Jansen¹

¹Radboud University, Nijmegen (The Netherlands)

²FrieslandCampina, Amersfoort (The Netherlands)

t.offermans@science.ru.nl

The goal of the presented study is to develop a soft-sensor for monitoring the byproduct concentration of a milk powder production plant in real-time. The concentration of this byproduct is the critical parameter for the product, and the control of the plant should be directed at keeping this concentration as low as possible. As this concentration can however not be measured directly in/on-line, there is a desire for a prediction model that can reliably estimate this concentration during production time. Such a prediction model will allow the plant operators to more accurately control the process and thereby optimize the product quality and the energy consumption of the plant.

The available production data are on-line near-infrared (NIR) spectroscopic measurements of the incoming milk, in-line measurements of process variables and at-line reference values for the byproduct concentration. Sequential and Orthogonalized Partial Least Squares regression (SO-PLS) was used to establish the predictive information from the NIR spectra and the process variables for the byproduct concentration. Different strategies for aligning the data sources have been investigated, of which median-filtering has been proven to be most fruitful for predictive modelling. The prediction model with the smallest prediction error used a selection of only the process variables. It is currently investigated whether transformation of the process variables and a stricter filtering approach can improve the prediction. The unprocessed NIR spectra were found to have no significant contribution to the prediction. Efforts are now being made to investigate whether the spectra can improve the prediction after transformation and/or preprocessing. Future work will also concentrate on improving the prediction by means of sparse modelling and extension to non-linear regression models, and on implementing the soft-sensor in an automated closed-loop control strategy.

Morten Arendt Rasmussen^{1,2*}, Åsmund Rinnan¹, Anne Bech Risum¹ and Rasmus Bro¹

¹⁾ Department of Food Science, University of Copenhagen

²⁾ Copenhagen Studies on Asthma in childhood, University of Copenhagen

* mortenr@food.ku.dk

Machine Learning solutions are getting more and more complex, while being nicely enveloped into black box pipelines handling all parts of the data analytical pipeline. This seeds the hope for being able to scale the use of data analytics beyond trained personal, e.g. such that a good calibration model can be built by anyone that has access to data. However, much of the domain knowledge and fingerspitzgefühl that trained researchers actively use when building models can be hard to automate.

In this project we ask the question of how well does a fully automated *one-button* procedure perform on a classical NIR-calibration task in comparison with three levels of trained human beings. Two NIR datasets with increasing complexity, prediction of protein in barley ($n_{\text{train}} = 308$, $n_{\text{test}} = 80$) and prediction of the assay value in pharmaceutical tablets ($n_{\text{train}} = 195$, $n_{\text{test}} = 460$), are used as data material for this task.

Two levels of students; level1: 35 bachelor in food science students attending the 2nd year course on Exploratory Data Analysis just after being briefly introduced to PLS and level2: 15 master in food science students attending the course on Advanced Chemometrics at the end of the course. This in comparison with expert level chemometric scientists at phd level and with a *one-button* procedure.

The one-button procedure performs automatic outlier removal, chooses preprocessing and variable selection, and selects the optimal model based on repeated cross validation procedures.

All models are evaluated on left out test-sets.

The results will be presented at SSC in Norway and will eventually point towards which teaching efforts that the scientific research community should focus on in the future.

Use of innovative FTIR spectroscopy sampling methods and chemometrics for authentication and differentiation of herbals

Agnese Brangule, Pēteris Tretjakovs

Riga Stradiņš University, Department of Human Physiology and Biochemistry
agnese.brangule@rsu.lv

Introduction. Herbal medicine (HM) has been used worldwide for more than hundreds of years as one of the most traditional forms of health care [1]. The chemical composition of herbs may vary depending on the species, a location of growth, age, harvesting season, drying conditions, and other conditions. [2]. In the field of HMs, the FTIR fingerprint spectra have been used since early 1987 and are used less frequently than chromatography methods [3]. Until now, the introduction of FTIR methods was limited by the complexity of spectra and its interpretation. FTIR spectroscopy, in conjunction with multidimensional statistical analysis (chemometrics), offers a very wide scope for HM studies [4]. **This study demonstrates** the significant potential of using innovative cantilever-enhanced Fourier transform infrared photoacoustic spectroscopy (PAS) principles and diffuse reflective infrared spectroscopy (DRIFT) with a diamond sampling stick and aluminum sampling cup in combination with a multivariate data processing methods. **Plant materials and**

Methods. In the present work, we evaluated dried herbals and herbal extracts in ethanol. PAS and DRIFT (PerkinElmer Spectrum One) spectra were taken at $450\text{--}4000\text{ cm}^{-1}$, at a resolution of 4 cm^{-1} , and an average made from 10 scans. For PAS, the homogenized samples were placed in the PAS cell filled with helium gas (flow 0.5 l/min). For DRIFT homogenized samples were placed on the diamond sampling stick, but extracts in the aluminum sampling cup and evaporated.

Spectral pre-processing. The FTIR spectra were viewed, smoothed, and had baseline correction and normalization performed with *SpectraGryph 1.2*. **Statistical Analysis.** The Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were performed using *SIMCA 14* software. PCA was used to identify the dominant clusters in the data set. For the HCA, Ward's algorithm was used. **Conclusions.** Comparison between spectra recorded by PAS and DRIFT showed high sensitivity and good resolution. The results obtained provide information about the spectral behavior of homogenized herbal and herbal extracts can be useful for establishing identification and discrimination criteria. It has been demonstrated that PAS and DRIFT can be a useful experimental tool for the characterization and discrimination of herbals.

References: [1] P. Wang, et al., Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review, *J. Pharm. Anal.*, vol. 5, no. 5, pp. 277–284, 2015. [2] H.A. Gad, et al., Application of chemometrics in authentication of herbal medicines: a review, *Phytochem. Anal.* 24, 1–24, 2013. [3] D. Riolo, et al., Raman spectroscopy as a PAT for pharmaceutical blending: advantages and disadvantages, *J. Pharm. Biomed. Anal.*, 5;149:329–334, 2018. [4] A. Bansal, et al., Chemometrics: A new scenario in herbal drug standardization, *J. Pharm. Anal.*, vol. 4, no. 4, pp. 223–233, 2014.

Acknowledgments. The research received funding from the ERAF Post-doctoral Research Support Program project Nr. 1.1.1.2/16/I/001 Research application "Development of screening methods by innovative spectroscopy techniques and chemometrics in research of herbal medicine", Nr. 1.1.1.2/VIAA/2/18/273.

Data Fusion in metabolomics

**Tomas Skotare¹, Rickard Sjögren^{1,2}, Frida Torell¹, Izabella Surowiec^{1,2},
Johan Trygg^{1,2}**

¹. Department of Chemistry, Umeå University, Umeå, Sweden

². Advanced Data Analytics, Corporate Research, Sartorius AG.

johan.trygg@umu.se

Today, large amounts of experimental data are being generated by modern high-throughput ‘omics’ technologies. The possibility to characterize a single sample using several profiling techniques generates multiple sets of large complex data. Extracting and integrating useful information from these data sets is a nontrivial task. Here, descriptive data analysis by means of a multi-block and data fusion approach can be used to quantitatively describe the main features in a collection of datasets and effectively map how different types of variation in each dataset is connected with the others. It can also provide the unique variation each individual dataset holds that is not found in the other datasets. More specifically, multi-block modelling can decompose each dataset into a globally joint part containing variation shared with all other connected data sets, a locally joint part containing variation that is shared with some, but not all, other data sets and unique variation, present only in a single data set.

We will demonstrate several examples on data fusion technology to integrate large and complex multi-omics datasets to facilitate improved understanding of the studied systems.

Design of Experiments for data generation and data processing in ‘omics studies (genomics - metabolomics)

Daniel Svensson¹, Rickard Sjögren^{1,2}, Izabella Surowiec^{1,2}, Johan Trygg^{1,2}

¹. Department of Chemistry, Umeå University, Umeå, Sweden

². Advanced Data Analytics, Corporate Research, Sartorius AG.

johan.trygg@umu.se

Today, massive amounts of data experimental data is being generated by modern high-throughput ‘omics and sensor technologies with increasing availability and decreasing experimental costs. This overwhelming size and complexity of modern ‘omics’ and phenotypic data have driven systems analysis towards the adoption of multivariate analysis and machine learning methods incl deep learning.

However, the principle, garbage in – garbage out, highlights the importance of good quality data and to minimize bias in data. Here, Design of Experiments, DOE, provides an established and easy to use strategy to enable the generation and data processing into high-quality and representative output data in ‘omics studies.

Recently developed generalized subset designs (GSD) based on fractional factorial designs have demonstrated success to both generate representative and balanced subset selections from large cohorts and biobanks and also providing an efficient strategy for optimizing data processing parameters.

Multivariate patent analysis

Rickard Sjögren^{1,2}, Kjell Stridh¹, Johan Trygg^{1,2}

¹. Department of Chemistry, Umeå University, Umeå, Sweden

². Advanced Data Analytics, Corporate Research, Sartorius AG.

johan.trygg@umu.se

Patents are an important source of technological knowledge but the amount of existing patents is very vast and quickly growing. This makes development of tools and methodologies for quickly reveal patterns in patent collections important. In this paper we describe how chemometric latent variable methods and principles of multivariate data analysis can be used to study collections of chemical patents represented as term-document matrices, or bags-of-words.

Using principal component analysis (PCA) on a collection of 12338 patent abstracts from 25 companies in big pharma revealed sub-fields which the companies are active in. Using PCA on a smaller collection of patents retrieved by searching for a specific term proved useful to quickly understand how patent classifications relate to the search term. With OPLS for patent classification we then were able to separate patents on a more detailed level than using PCA. Lastly we performed multi-block modelling using OnPLS on bag-of-words representations of abstracts, claims and detailed descriptions showing that semantic variation relating to patent classification is consistent across multiple text blocks.

We conclude that chemometric latent variable methods can provide a useful tool to understand collections of chemical patents due to ease of model interpretation.

Effects of long distance walking analyzed by multidimensional flow cytometry analysis of neutrophils

Carlo G. Bertinetto,¹ Roy Spijkerman,² Lilian Hesselink,² Jeroen J. Jansen,¹ Leo Koenderman.²

¹Radboud University, Institute for Molecules and Materials, (Analytical Chemistry), Nijmegen, The Netherlands.

²Laboratory of Translational Immunology, University Medical Center Utrecht, Utrecht, The Netherlands.

c.bertinetto@science.ru.nl

The innate immune system plays a major role in health, as it is the means to prevent infections and keep homeostasis. The neutrophil is an very important effector cell in the innate immune response. It is known that intense physical exercise induces changes in the neutrophil count and activity, but how exactly this affects the immunoresponse of different types of subjects is not fully understood.

In the current study, we analyzed a set of 45 people (mean age 63.8 ± 6.9 , 35% females) who participated in the “4-Day Marches”, an annual four-day walking event in Nijmegen, The Netherlands. Blood samples were collected from the subjects at baseline and shortly after each day of walking (30, 40 or 50 km). They were measured on-site, using a fully automated load-and-go Multicolor Flow Cytometer (MFC) in a mobile lab. Measurements were done with and without adding a bacterial stimulus that mimics a response to an infection.

The resulting MFC data were gated on the neutrophils and analyzed mainly using two methods recently developed at Radboud University for such purposes. The first method, named Elimination of Cells Lying in Patterns Similar to Endogeneity (ECLIPSE), combines PCA with Kernel Density Estimates to distinguish and characterize the treatment-related cells from the control ones. The second method is named Discriminant Analysis of MultiAspect CYtometry (DAMACY), which applies OPLS-DA on the multidimensional histogram of the scores obtained from the overall PCA model.

These methods enabled to delineate the main response pattern for the data set under study, but also to observe differences specific to individual subjects. The main modes of response could be highlighted by clustering the ECLIPSE-PCA scores of each individual. Among the most notable results was the observation of an effect, for the last analyzed day, in opposite direction as compared to the previous ones (see Fig. 1), suggesting that the body counters the initial changes. We also observed an interesting similarity between the response from several days of walking and the samples activated with the bacterial stimulus.

In conclusion, this study indicates that four days of chronic exercise is associated with initial activation of the neutrophil compartment followed by partial normalization at day 3-4.

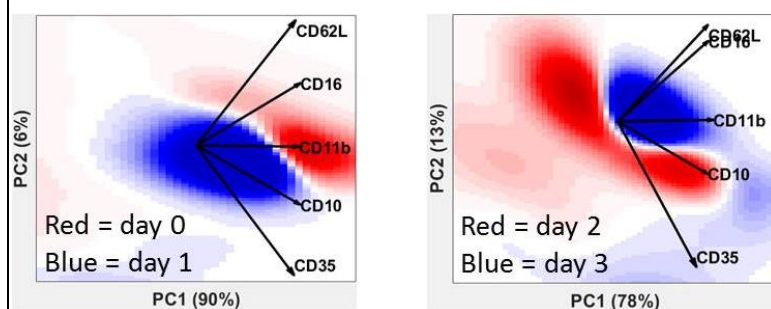


Fig. 1 OPLS weights of DAMACY models built on the control-response pairs of (left) baseline (day 0)-day 1 and (right) day 2-day 3, respectively. Loadings indicate different fluorescent markers.

Similarity metrics for binary data structures in cheminformatics, metabolomics and other fields

Dávid Bajusz, Anita Rácz, Károly Héberger

Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary, Bajusz.david@ttk.mta.hu

Binary data structures are ubiquitous in many, often unrelated scientific areas. A binary vector (or bitstring) representing an object is often called a fingerprint, and considered to be a unique and compact identifier of said object. Examples for binary fingerprints include molecular fingerprints, encoding molecular structure (with bit positions associated to the presence or absence of certain substructures),^[1] or metabolomics fingerprints, encoding sample composition (with bit positions associated to the presence or absence of certain components/metabolites). It is often desirable to calculate the similarity of such binary fingerprints, for various purposes including hierarchical clustering of samples, virtual screening for drug candidates, *etc.* To that end, a large number of similarity metrics were proposed by researchers of diverse fields over the past century, most of which have remained virtually unknown for the broader scientific community (they were, however, collected in a recent work^[2]).

Recently, we have conducted several large-scale comparative studies to explore an exhaustive set of binary similarity metrics for various applications, with the aid of a robust statistical comparison method, sum of ranking differences (SRD).^[3] For molecular fingerprints, we have found that from a small pool of the most commonly known similarity metrics, the well-known and generally favored Tanimoto coefficient is indeed a valid preference.^[4] In metabolomics, we have demonstrated that with a suitable similarity metric, binary metabolomic fingerprints can be used for the clustering of samples with minimal misclassification (as compared to the use of quantitative chemical composition data).^[5] For protein-ligand interaction fingerprints, we have highlighted six similarity measures as better or comparable alternatives for the popular Tanimoto similarity coefficient.^[6] We have published a compact, open-source Python package, implementing 51 similarity metrics for binary data formats: <https://github.com/davidbajusz/fpkit>

References

- [1] D. Bajusz, A. Rácz, K. Héberger, in *Compr. Med. Chem. III* (Eds.: S. Chackalamannil, D.P. Rotella, S.E. Ward), Elsevier, Oxford, **2017**, pp. 329–378.
- [2] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, *J. Chem. Inf. Model.* **2012**, 52, 2884–2901.
- [3] K. Héberger, *TrAC Trends Anal. Chem.* **2010**, 29, 101–109.
- [4] D. Bajusz, A. Rácz, K. Héberger, *J. Cheminform.* **2015**, 7, 20.
- [5] A. Rácz, F. Andrić, D. Bajusz, K. Héberger, *Metabolomics* **2018**, 14, 29.
- [6] A. Rácz, D. Bajusz, K. Héberger, *J. Cheminform.* **2018**, 10, 48.

Rapid identification of reaction systems using spectroscopic measurements and micro-reactors

Manokaran V

Ph.D. Scholar, Department of Chemical Engineering, Indian Institute of Technology Madras,

email – manokaran.deva@gmail.com

Advisors: Dr. Sridharakumar Narasimhan, Dr. Nirav Bhatt

Abstract

Micro-reactors are continuous flow reactors with better heat and mass transfer characteristics, can be used as alternate over the conventional reactors. In micro-reactors concentrations are measured as a function of residence time (τ) as compared to reaction time (t) in batch reactors. This work involves rapid identification of reaction systems using online spectral data from experiments carried out in continuous flow micro-reactors. Micro-reactors are assumed to follow plug flow reactor model where concentrations are measured indirectly using spectral measurements. Typically, prediction of concentrations from spectral data involves building a calibration model. These predicted concentrations will be later used for identification of reaction systems [1, 2]. Calibration model developed is reaction system dependent, and hence, a new calibration model is need to develop for each reaction system. Also, building a calibration model consumes considerable amount of reagents. So, calibration-free approach reduces experimental efforts and aids in faster identification of reaction systems [3]. We have demonstrated that the calibration-free method to analyse spectral data obtained from micro-reactors can be used to identify reaction kinetics in complex reaction systems via in-silico studies [4]. In this presentation, experimental validation of the proposed approach using a model reaction (Wittig reaction) will be presented.

If reaction rate model structure is unknown, several models are proposed for each reaction. The identification step will be combinatorial using spectral data, leading to higher computational time in simultaneous identification [5]. In contrast to this approach, the extent based incremental identification deals with decoupled rate equations and hence reduces the computational time. However, it is not straightforward to compute the extents from the experimental data due to rank-deficient spectral matrix [6]. In this work, it is proposed to compute reaction-variant form of spectral data, and then, to develop a blind source separation technique for recovering the extents of reaction from the data. These estimated extent can be used to perform reaction model discrimination task for each reaction separately. It is also proposed to apply wavelength (variable) selection methodologies to select the information rich wavelength ranges of spectral data. The parameter estimation results obtained using the entire spectral data and selected wavelength will be compared.

References

1. S. Mozharov et al. Journal of the American Chemical Society, 2011, 133, 3601-3608.
2. J. S. Moore and K. F. Jensen. Angewandte Chemie, 2014, 126, 480-483.
3. M. Amrhein et al. Chemometrics and Intelligent Laboratory Systems, 1996, 33, 17-33.
4. M. Veeramani et al. Computer Aided Chemical Engineering, 2018, 44, 931-936.
5. M. Brendel et al. Chemical Engineering Science, 2006, 61, 5404 – 5420.
6. N Bhatt et al. Industrial & Engineering Chemistry Research, 2011, 50, 12960–12974.

Novel NIR analysis of Heterogeneous Powder

Jacob Kræmer Hansen, and Åsmund Rinnan

University of Copenhagen, JKH@food.ku.dk

The scope of my project is to utilize NIR to analyze heterogeneous powders, without loss of information. To obtain this goal three issues must be handled: Representative sampling of the powders, representative scanning of the powder during the NIR measurement and selection of an appropriate frequency of spectral recordings. As representative sampling of powders has been widely described in literature, this will not be the main scope of the project, though the importance of this should not be underestimated.

There are many commercially available solutions of getting representative scanning of powders. Most of these rely on rotating sample vials or petri dishes to enable recording of spectra from a larger surface area of the sample. Despite of the rotation the actual amount of powder analyzed is low. Often, the rotation is around a vertical axis (in the horizontal plane); only measuring the powder at the bottom of the petri dish or at a specific point in the sample vial. Furthermore, most instrumental software only records an average of multiple scans as the sample container spins to record one single measurement for one sample. We propose to utilize a half-filled cylindrical sample vial rotating around a horizontal axis, while being translated in the same horizontal axis. During this rotation and translation, we can measure the entire length and surface of the vial. This allows for measurements over a greater surface area in combination with achieving a mixing of the powder during analysis. The resulting NIR spectra would therefore be representative of a greater volume of powder, compared to the commercially available setup.

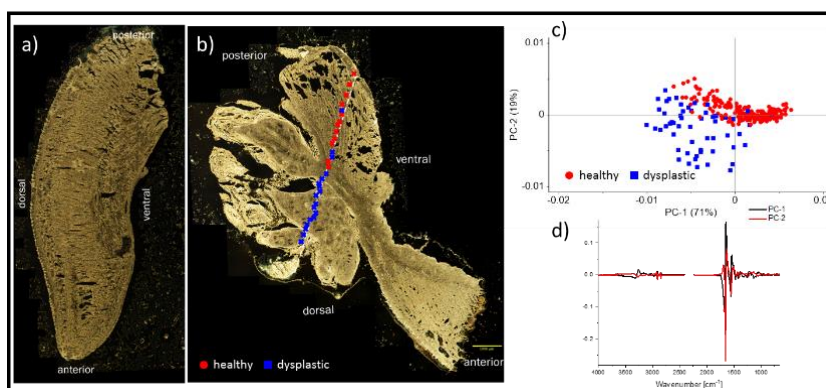
This way of measuring a single sample, will lead to a vast amount of data, especially for larger sample set. Therefore, the preferred method of analysis, is to average across all these scans for each sample. We propose to keep all the scans, and instead inspect the total variation present in the data, as we hypothesis that there is a large amount of information present in the raw, non-averaged data. By this, we are capable of investigating local differences in the chemical and/or physical properties of the sample. For each sample point we can now do pre-processing, storing the pre-processing parameters, and thus get an increased understanding of the changes according to physical changes in the powder (i.e. particle size), while the subsequent PCA will give us information about chemical differences (i.e. moisture level, polymorph type).

Infrared spectroscopy and multivariate data analysis for the label-free early stage diagnosis and demarcation of head and neck cancer in a mouse model

Mona Stefanakis, Edwin Ostertag, Karsten Rebner, Alfred J. Meixner, Marc Brecht

Process Analysis & Technology, Reutlingen University, Alteburgstr. 150 72762 Reutlingen, Germany, mona.stefanakis@reutlingen-university.de

Histopathological tissue characterization (Hematoxylin-Eosin), MIR-microspectroscopy (2400 cm^{-1} to 650 cm^{-1}) and multivariate data analysis (PCA-DA) is used to investigate cross sections of a mouse model for head and neck (HaN) cancer. An early detection and a clear identification of the tumor free margins is essential for therapy of HaN cancer. [1, 2] The mouse model consists of eleven mice, where seven are treated by a carcinogen and four are used as a control group. The carcinogen 4-nitroquinoline N-oxide is administered to the seven mice in the drinking water inducing a tumor genesis of HaN carcinoma in the tongue. The growth of different dysplastic stages in the seven treated mice and the development of four untreated mice are observed under defined conditions. The cross sections of these mouse tongues are cut in longitudinal and transversal cutting directions (see figure a) untreated mouse tongue and b) treated mouse tongue). Several cross sections are stained for histopathological classification and the others remain unstained for the label-free spectroscopic examination. The histopathological characterization with Hematoxylin-Eosin staining yield an assignment to different tissue types as well as to different dysplastic stages. Based on this assignment tissue areas of the unstained cross sections are identified for subsequent contactless MIR-microspectroscopic investigation in reflexion mode. Spectra are recorded at up to 350 different histopathological clearly identified positions at each mouse tongue. Multivariate models with a principal component analysis (see figure c) scores and d) loadings) combined with a Bayesian discriminant analysis on the recorded spectra are built. The results are compared to the pathologist's assessment. Based on this a prediction of unclassified spectra across cancerous cross sections is done (see figure b), red/blue line). With the models healthy (red) and unhealthy (blue) regions are demarcated. Even regions which are not clearly assigned by standard histopathology can be correlated. In our approach, there should be a quick and easy decision if a tissue is healthy or dysplastic or where a surgical resection of a carcinoma could be done.



We acknowledge Dr. Inês Sequeira and Prof. Dr. Fiona M. Watt for the sample set and the preparation.

- [1] Argiris, A., et al. Head and neck cancer. *The Lancet*, 371(9625), 1695-1709 (2008)
- [2] Meyer, T., et al. Multi-modal nonlinear microscopic investigations on head and neck squamous cell carcinoma: Toward intraoperative imaging. *Head & neck*, 35(9), E280-E287 (2013)

PROCESS PLS: A NEW PATH MODELING ALGORITHM FOR HIGH DIMENSIONAL AND MULTICOLLINEAR DATA

Roel Bouman^{1,2}, Geert van Kollenburg^{1,2}, Jeroen Jansen^{1,2}

¹ *Radboud University, Department of Analytical Chemistry/Chemometrics*

² *TI-COAST*

Presenting Author: rbouman@science.ru.nl

Chemical processes are often analyzed using dimensionality reduction based techniques such as PLS and PCA. While these techniques already help in uncovering chemically relevant information, the incorporation of more domain knowledge can help in elucidating these processes even more. Consequently, path modeling methods such as PLS-PM and SO-PLS-PM, which combine dimensionality reduction and regression, have recently been applied to elucidate chemical processes.

We present the novel path modelling algorithm Process PLS. We demonstrate by use of several simulated datasets that Process PLS solves several of the key problems of PLS-PM when the latter is applied to data for which several of its key assumptions do not hold. Process PLS can estimate models with looser assumptions than PLS-PM, allowing it to be used on data which is not unidimensional, multicollinear, and which suffer from the sign ambiguity in decomposition.

Process PLS was furthermore compared to SO-PLS-PM by applying it on the well-known Val de Loire wine dataset. Results of Process PLS and SO-PLS-PM are similar, but differ in that Process PLS emphasizes direct effects when these are present. Process PLS is furthermore applicable to network topologies for which SO-PLS-PM can not be used due to its method of block incorporation. In this manner the developed Process PLS algorithm is a more broadly applicable alternative to other path modeling methods such as PLS-PM and SO-PLS-PM.

Getting more from the PLS model: application to metabolomics

Gavin Rhys Lloyd*¹, Ralf Weber¹

*g.r.lloyd@bham.ac.uk

1. Phenome Centre Birmingham, University of Birmingham, UK

Partial Least Squares Discriminant Analysis (PLS-DA) is a commonly used method for the analysis of metabolomics data sets. PLS and its variants such as Orthogonal PLS are popular in part because they provide a means for interpretation of high dimensional data and metrics for assessing the importance of features for the model. These aspects are important for metabolomics analysis where a key aim is to identify metabolites that vary in response to e.g. disease.

The PLS model can be described as an 'inverse' regression model in which a matrix of metabolite peak intensities for each sample is regressed onto a vector or matrix of group labels, sometimes called a design matrix. This is in contrast to some univariate tests, such as t-test and ANOVA that use a 'classical' approach to regression where the design matrix is regressed onto the data matrix i.e. the matrices in the regression formula have switched places. The outputs from PLS are therefore difficult to compare with univariate approaches as a different philosophy underlies each model.

To overcome this we show how that the regression coefficients of a classical regression model can be used to calculate the coefficients of the inverse regression model for the same data. We illustrate, with a metabolomics dataset, how this can be extended to PLS to extract additional metrics that are more directly comparable with univariate approaches.

Statistics in R Using Class Templates (StRUCT)

Gavin Rhys Lloyd*¹ and Ralf Weber¹

*g.r.lloyd@bham.ac.uk

1. Phenome Centre Birmingham, University of Birmingham, UK

R is a powerful statistical computing environment that is increasingly used for metabolomics analysis, partly due a drive toward open-source, reproducible software in this area. The R language allows packages to be imported from multiple sources such as CRAN, Bioconductor and GitHub to add functionality provided by the community. Whilst this allows huge flexibility this can also make it difficult to quickly produce reproducible code and workflows because packages can be incompatible or hard to combine.

To address this we have developed an R package called `struct` to facilitate the development of reproducible workflows in R. The StRUCT framework provides a set of templates that allow functionality from other R packages to be wrapped into a common format making use an object-oriented approach. The framework also allows different workflow steps to be quickly and clearly combined and a uniform approach to generating outputs such as charts and tables that facilitate consistent reporting, as well as readable, reproducible code.

To demonstrate the utility of StRUCT we have developed a second package `structToolbox` that uses StRUCT templates to provide a suite of tools (PCA, PLS etc) for carrying out multivariate analysis. The use of templates allows the toolbox to easily be expanded using functionality from other packages provided by the community.

In this poster we demonstrate some of the features provided by `struct` and the accompanying `structToolbox` for multivariate analysis of a metabolomics dataset. The pre-release versions of the `struct` and `structToolbox` packages are currently publically available on GitHub (computational-metabolomics/struct).

Detection of High Fructose Corn Syrup in Honey by Fourier Transform Infrared Spectroscopy and Chemometrics

Mercedes Bertotto, Marcelo Bello, Héctor Goicoechea, Verónica Fusca

National Service of Agri-Food Health and Quality, Talcahuano 1660, CO 1640 Argentina
(phone: 541144442272; e-mail: mbertotto@senasa.gob.ar).

The National Service of Agri-Food Health and Quality (SENASA), controls honey to detect contamination by synthetic or natural chemical substances and establishes and controls the traceability of the product. The utility of near infrared spectroscopy for the detection of adulteration of honeys with high fructose corn syrup (HFCS) was investigated. First of all, a pool of different argentinian honeys was prepared and then, several mixtures were prepared by adding different concentrations of high fructose corn syrup (HFCS) to samples of the honey pool. 237 samples were used, 108 of them were authentic honeys and 129 samples corresponded to honeys adulterated with HFCS between 1 and 10%. They were stored ambient temperature from time of production until scanning. Immediately prior to spectral collection, honeys were incubated at 40°C overnight to dissolve any crystalline material, manually stirred to achieve homogeneity and adjusted to a standard solids content (80° Brix) with distilled water. Adulterant solutions used as standard were also adjusted to 80° Brix. Samples were measured by Microscopy coupled Fourier Transform Infrared Spectroscopy in the range of 650 to 7000 cm^{-1} . The technique of specular reflectance was used, with a lens aperture range of 150 mm. A pretreatment of the spectra was performed by Standard Normal Variate (SNV). The ant colony optimization genetic algorithm sample selection (ACOGASS) graphical interface was used, using MatLab version 5.3, to select the variables with the greatest discriminating power. The data set was divided into a validation set and a calibration set, using the Kernnard-Stone (KS) algorithm. A combined method of Potential Functions (PF) together with Partial Least Squares-Discriminant Analysis (PLS-DA) was chosen. Different estimators of the predictive capacity of the model were compared, which were obtained using a decreasing number of groups, which implies more demanding validation conditions. The optimal number of latent variables was selected as the number associated with the minimum error and the smallest number of unassigned samples. Once the optimal number of latent variables was defined, the model was applied to the training samples. With the model calibrated with the training samples, the validation samples were studied. The calibrated model that combines the potential function methods and PLS-DA can be considered reliable and stable, since its performance in future samples is expected to be comparable to the one achieved for the training samples. By use of Potential Functions (PF) and Partial Least Square Linear Discriminant Analysis (PLS-DA) classification, authentic honey and honey adulterated with HFCS could be discriminated with a correct classification rate of 97.9%. The results showed that NIR in combination with the PT and PLS-DS methods can be a simple, fast and low cost technique for the detection of HFCS in honey with high sensitivity ($\text{LoD}=1\%$) and specificity.

Mid-infrared spectroscopy and multivariate analysis to characterize *Lactobacillus acidophilus* fermentation processes

Sumana Narayana^{1,*}, Line Christensen², Thomas Skov¹, Frans van den Berg¹

¹ Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 26, Frederiksberg C, Denmark

² Chr. Hansen A/S, Søndre Ringvej 22, DK-4000 Roskilde, Denmark

*corresponding author, sumana@food.ku.dk

Process Analytical Technology based sensors such as mid-infrared (MIR) spectroscopy have been gaining popularity for process monitoring in the biotech industry. The multivariate nature of MIR enables following of the process through dynamics of substrate and by-product concentrations, which directly represent the physiology of cells. Combined with multivariate statistics, MIR can be employed as a strategy for production performance mapping and detect deviations from the normal trajectory. In this study, we describe the application of MIR spectroscopy to characterize *Lactobacillus acidophilus* production. The process considered is a simple fermentation, involving conversion of sugars (glucose and lactose) into lactic acid by the biomass involved.

The focus of the study is the alternative arrangements of the spectroscopic data and application of chemometric methods - Principle Component Analysis (PCA), Multivariate Curve Resolution (MCR) and Parallel Factor Analysis (PARAFAC) to decompose it. Following modelling, a post-process fitting on the scores from the different models is performed and two key parameters, “rate constant” and “time of inflection” are extracted. Their use as process performance descriptors to characterize the dynamics of substrate consumption, product formation and batch-to-batch variations is suggested.

The simplest chemometric model, PCA, models the dominant change in biomass during the process and captures differences between batches well along the first PC. The model is too flexible and needs constraints for better description of the principle variations occurring during the fermentation process. When used with non-negativity constraints, PARAFAC resulted in two distinct dynamic profiles, interpreted as consumption of sugars and production of lactic acid. The spectral loadings however, indicate that the data at hand is not completely tri-linear, rendering PARAFAC unsuitable.

MCR was considered in both augmented (MCR(a)) and individual-batch (MCR(i)) arrangements; followed by post-processing, MCR(a) also showed two profiles similar to PARAFAC results. In contrast to PARAFAC however, the spectral loadings from MCR(a) were chemically meaningful loadings in both components. The last modelling strategy, MCR(i) followed by post-processing, was found to be the most acceptable and interpretable solution applied to the MIR data in this study. It was concluded that MCR on individual-batch data, followed by post-process fitting is the preferred strategy for MIR spectroscopic monitoring of the fermentation process considered.

Gene expression in petroleum workers exposed to sub-ppm benzene levels

Katarina M. Jørgensen^{1,2}, Ellen Færgestad Mosleth³, Kristian Hovde Liland^{3,4}, Nancy B. Hopf⁵, Rita Holdhus^{1,6}, Anne-Kristin Stavrum^{1,6}, Bjørn Tore Gjertsen⁷ and Jorunn Kirkeleit^{1,8,*}

¹ Department of Clinical Science, University of Bergen, P.O. Box 7804, N-5020 Bergen, Norway. ² Institute of Marine Research, P.O. Box 1870 Nordnes, N-5817 Bergen, Norway

³ Nofima AS, Osloveien 1, N-1430 Ås, Norway. ⁴ Faculty of Science and Technology, Norwegian University of Life Sciences, NO-1430 Ås, Norway. ⁵ Institute for Work and Health (IST), Universities of Lausanne and Geneva, CH-1066 Lausanne-Epalinges, Switzerland. ⁶ Department of Medical Genetics, Haukeland University Hospital, P.O. Box 1400, N-5021 Bergen, Norway. ⁷ Center for Cancer Biomarkers (CCBIO), Department of Clinical Science, Precision Oncology Research Group, University of Bergen, P.O. Box 7804, N-5020 Bergen, Norway. ⁸ Department of Global Public Health and Primary Care, University of Bergen, P.O. Box 7804, N-5020 Bergen, Norway

Abstract

Background:

Benzene exposure has previously been found to be linked to increased cancer incidence in offshore workers exposed to benzene concentrations below the current occupational exposure limit.

The aim of the study:

The aim of the study was to investigate the effects on gene expression of exposure to low doses of benzene during short time exposure for offshore workers. Furthermore, in many fields, data availability is often restricted, leading to small number of samples while many variables may have been measured. Rather than performing data analysis on a small data set with many variables, we here aim to present a strategy where our data set is used for validation of the conclusion of another study.

Methodology:

Genes expression changes in petroleum workers exposed to sub-ppm benzene exposure (1 ppm) were tested during three consecutive work shifts. Previous studies have shown that sub-ppm benzene levels reduces counts of circulating white blood cells and results in changes in gene expression in exposed workers. In a Chinese study on gene expression data¹, six biomarkers were identified in chronically exposed factory workers as elevated under low benzene exposure. We selected these genes in our study and validated their relation to benzene exposure in our data.

Results and implications:

Benzene exposed workers were separated from unexposed referents also in our data for four of the six genes previously proposed¹ as biomarkers of chronic sub-ppm benzene exposure. The strategy of limiting our validation to biomarkers identified in another study rather than validating the expression of a very large number of genes using a few samples, strongly improved the power of our validation and the conclusion reached from our study.

Refr

1. Schiffman et al (2018), PLoS ONE, 13, e0205427

A Comparison of ANNs, SVMs, and XGBoost in Challenging Classification Problems

Barry M. Wise, Donal O'Sullivan and Manuel A. Palacios

Eigenvector Research, Inc. bmw@eigenvector.com

Many methods have been developed to classify samples based on a multivariate response. In the realm of linear methods, Partial Least Squares Discriminant Analysis (PLS-DA) has become a very widely used method with chemical spectroscopic data. Numerous non-linear methods have also evolved, starting with Artificial Neural Networks (ANNs) which became popular in the 1980s and then Support Vector Machines (SVMs) whose use became widespread in the 1990s. More recently, the XGBoost implementation of boosted regression trees has generated a lot of interest, particularly in machine learning competitions.

In this work we compare these methods on a number of difficult classification problems. By “difficult” we mean that our correct classification rate is less than 90% using PLS-DA, our benchmark method. Data sets used include NMR for breast cancer detection, Excitation Emission Fluorescence for cervical cancer detection, hyper spectral image for crop identification, and a number of others.

Tuning of the meta-parameters in the non-linear methods is considered. Up front compression, using Principal Components Analysis (PCA) and PLS-DA was studied at length. Visualizations of the decision surface were developed in order to better understand the performance of the non-linear methods. Error rates are presented as well as observations about usability.

On some of the problems the overall performance of the non-linear methods was not much better than PLS-DA, on others there was significant improvement. The overall winner was SVM-DA, though this likely problem set dependent.

Experiments with complex numbered multivariate data analysis

**Olof Svensson¹, Amina Souihi², Álvaro Díaz-Bolado³, Anders Sparén¹,
and Mats Josefson^{1,4}**

1, AstraZeneca Gothenburg, Sweden; 2, Umeå University, Sweden; 3, ViaSat, Switzerland
4, AstraZeneca, SE-431 83, Mölndal, Sweden, e-mail: mats.josefson@astrazeneca.com

Pharmaceutical products may be sensitive to water and therefore, it is important to measure and control the water content of a formulation during manufacturing and storage of pharmaceuticals. The traditional method for determining water content is Karl Fischer titration, but this method is both time consuming and sample destructive. Microwave spectroscopy using resonators is a fast and non-destructive alternative to Karl Fischer titration. Microwave measured data are generally evaluated in a univariate way by calculating the ratio between the frequency shift and the peak width broadening for a resonance mode using the amplitude of the microwave signal. A calibration between the moisture content and the microwave ratio is then calculated using e.g. 1st, 2nd or 3rd degree polynomials. On the other hand, microwave measurement equipment generally provides both amplitude and phase information of the measured signal. This additional information could be exploited by modifying the current multivariate data analysis techniques to handle complex measured variables.

In this work, we explored the use of multivariate analysis to evaluate both real and complex numbered spectra to achieve a level of understanding during interpretation of what is happening with loadings, scores and predictions for the complex numbered case using both singular value decomposition and NIPALS versions of complex multivariate algorithms.

Simulated and experimental data was used to investigate and understand how the multivariate methods work and should be interpreted for data consisting of complex numbers both for audio and microwave spectroscopy. The multivariate methods were also used to analyze real microwave spectroscopy data from a moisture content calibration and a study to investigate the interaction between water and material movement using resonant microwave spectroscopy measurements.