Session 1 - Chemometrics for process modelling/control/monitoring

16-06-2025 - 13:00 - 14:55

Chemometrics. Methods and concepts. From one-block to multi-block analysis

Tormod Næs [1]

[1] Nofima, Ås, Norway

This presentation is a discussion of basic concepts and methods in chemometrics. Focus will be on classical disciplines such as calibration/regression and classification. The talk with start with a discussion of fundamental issues such as collinearity and linearity. It will be illustrated why collinearity is a problem both for prediction and interpretation and that it plays a different role in regression and classification. Component methods will be presented as a way of solving these problems in a stable and reliable way. Further advantages of component methods will be illustrated. Prediction bias as a consequence of using component methods in regression will be discussed together with other ways of defining bias. Possible reasons for good performance of linear methods within this framework will be discussed. Component-based methods for handling non-linear relations will also be presented. The concept of components will then be discussed within the framework of multi-block analysis. The distinction between common and distinct components of the blocks is a central aspect in this context. It will be discussed how this distinction has consequences for multi-block regression analysis. The talk will end with a discussion of the basic methodologies involved in many of the standard methods used in chemometrics.

From chemical fingerprints to environmental footprints: advancing feed production through near-infrared spectroscopy, Life Cycle Assessment and Chemometrics

Jeroen Jasper Jansen [1]

[1] Radboud University, Nijmegen, Netherlands

Process Analytical Technologies has been the key technology of quality maintenance and improvement in process industry. Quality is however only one indicator of process excellence: Safety, Cost, Delivery, Maintenance and specifically Environment are strongly complementary determinants of process value. The rising societal demands on sustainability of contemporary process industry has made specifically Environmental impact increasingly relevant, demonstrable by the implementation of CSRD into national legislation in the coming years. This however creates an

interesting “collision of timelines” as the future predictions from large volumes of PAT data collide with the retrospective quantification of environmental with Life Cycle Assessment (LCA), *of data that is available at time of production.*

Aside from quality information, PAT data (e.g. NIR spectra) contain a wealth of information on aspects like provenance, which are the key inputs for LCA. The available sustainability data on ingredients may therefore also be used to *predict* the footprint of the end-product. In this way, both quality and environmental impact (and production cost) may be simultaneously predicted. This allows the producer to take control of the product footprint, like they already are used to take control of quality through PAT. We show on a case study of animal feeds, how NIR spectroscopy (1) adequately predicts all product outcomes, (2) likewise predicts ingredient provenance, thereby providing a *paperless evidence basis* for their origin and (3) makes transparent the economic balance underlying sustainable production.

A pipeline for predictive modelling using industrial time series data

Ingrid Måge [1], Marco Cattaldo [1]Alberto Ferrer [2]

[1] Nofima AS, Ås, Norway
[2] Universitat Politècnica de València, Valencia, Spain

Introduction

In many industry projects, the aim is to develop models that can predict the outcome of the process based on data from raw materials and processing. To achieve this, relevant data from the process needs to be collected, and it is usually necessary to combine data from several sensors and processing steps. Spectroscopic sensors are often used in-line to characterize raw material and product quality. These sensors are multichannel and present specific challenges when combined with other process sensor data.

The pipeline from raw process data to a useful prediction model involves many steps, with a range of choices to be made at each step. Generally, the choice of methods and parameters at each step should be based on both domain knowledge and data science expertise. But even for highly competent project teams, the choices are not straightforward and are often made pragmatically or based on intuition.

Purpose

The primary aim of this study is to present a pipeline from raw process data to a validated prediction model, and a strategy for optimising the pipeline. Our strategy involves comparing different combinations of methods and parameters at each step, using an experimental design. This approach helps identify the optimal choices, and highlights areas of the pipeline that require further attention.

Methods

The proposed pipeline consists of four main steps:

1. Select relevant time frames and variables: Identifying relevant data sources and time frames with regard to process on/off or different product categories, and performing examine each variable to understand data characteristics. Split data in training and validation set.

2. Preprocessing of individual variables: Outlier removal, smoothing time series data, and handling missing values. Spectroscopic sensor data typically need specific preprocessing followed by feature extraction.

3. Data fusion and synchronization: Combining individual time series data into cohesive data tables, adjusting for time delays and differences in time resolution.

4. Predictive modelling: Selecting an appropriate modelling method, performing variable selection, and optimizing model parameters.

Results

The pipeline was applied to an enzymatic hydrolysis process in the food industry, using data collected over five weeks. The study identified specific steps in the pipeline with high impact on model performance, which in this case was adjustment of time lags and modelling method, i.e. steps 3 and 4 of the pipeline.

Conclusions

This study demonstrates the importance of a systematic approach to building predictive models in industrial settings. Our strategy also gives insight into which parts of the pipeline that has the largest effect on the outcome and therefore deserves more attention.

Validation of Predictive Models Based on Time-Series Data for Bioprocess Monitoring and Control

Andreas Eriksson [1]Rafael Machleid [2], Izabella Surowiec [3], Olivier Cloarec [4], Pär Jonsson [5]

[1] Department of Chemistry, Umeå University, Umeå, Sweden
[2] Sartorius Stedim Biotech GmbH, Göttingen, Germany
[3] Sartorius Stedim Data Analytics AB, Umeå, Sweden
[4] Sartorius Corporate Research, Bordeaux, France
[5] Sartorius Corporate Research, Umeå, Sweden

Introduction

In the biopharmaceutical industry, mammalian systems like Chinese hamster ovary (CHO) cells are the predominant producers of therapeutic proteins like monoclonal antibodies. Techniques such as Raman or infrared spectroscopy are commonly used to monitor culture performance through estimation of key nutrient concentrations, cell culture characteristics, and critical quality attributes (CQAs). Conversely, mass spectrometry-based omics are commonly used to decipher the complex biochemical reactions and metabolic pathways that underlie the production process. However, analysis of bioprocess data coming from both spectroscopic and omics techniques struggles with the time dependence present in the data [1]. Analytes like nutrients, metabolites, and products are collinear, and they, and many CQAs are furthermore inherently correlated with elapsed process time. This dependence may lead to false positives during validation of calibration models, and time becomes a confounding factor in elucidating metabolite-CQA relationships.

Purpose

To address these time-related challenges, we have developed a novel validation approach for predictive models based on time-series data. This method detects the limitations of collinearity and false positives, enabling more robust model predictions.

Methods

Our validation method is applicable to any machine learning (ML) method that predicts process parameters. The method assesses a model's ability to predict the difference between the actual and expected response values. In essence, the method compares the deviation from the global expected value across every time point (global model) against the deviation at each given time point. If the deviations diverge, the global model's predictions are overly optimistic, making predictions at individual time points inaccurate and unsuitable for making process adjustments. Conversely, if the deviations do not diverge, the model is robust and reliable.

Results

We demonstrate the new validation method on orthogonal partial least squares (OPLS) models based on three experimental time-series data sets: metabolite concentrations and process variables from 11 different CHO clones cultivated over 14 days; mAb titer, glucose, lactate, and Raman spectra from the same cultivations as the first data set, and mAb titer and Raman spectra from 12 CHO cell cultivations run under different process conditions over 12 days, visualizing how the method differentiates between reliable and poor prediction models.

Conclusions

In summary, we have developed a method for validating predictive models based on time-series bioprocess data, allowing the user to identify overoptimistic predictions stemming from spurious correlations with time. By doing so, the method allows for more robust machine learning models and more accurate process monitoring and control.

References

Alinaghi M.; Surowiec S.; Scholze S.; McCready C.; Zehe C.; Johansson E.; Trygg J.; Cloarec O. Hierarhical time-series analysis of dynamic bioprocess systems. Biotechnology Journal **2022**, *17(12)*,1-13.

NIR hyperspectral imaging and chemometrics for hybrid sausage assessment

Victor Gustavo Kelis Cardoso [1], Giulia Gorla [2, 3], José Manuel Amigo [2, 3], Rasmus Bro [1]Daniel Halling Breiner [4], Ivan R. Perch-Nielsen [4], Frederik Nielsen [5], Marchen Sonja Hviid [4], Patrick Bowen Montague [5]

[1] University of Copenhagen, Frederiksberg, Denmark
[2] University of the Basque Country, Leioa, Spain
[3] Ikerbasque, Bilbao, Spain
[4] Danish Meat Research Institute, Taastrup, Denmark
[5] NKT Photonics A/S, Birkerød, Denmark

Consumer interest in environmentally friendly products has been increasing steadily in recent years. To meet this growing demand, hybrid sausages – combining meat and plant-based ingredients – have emerged as an alternative to reduce meat content in processed foods [1]. Additionally, these products offer an opportunity to upcycle nutritious by-products, such as fermented brewing spent grains, which are rich in fibres, support circular economy practices, and can promote functional properties in food formulations. However, the complex composition of such products poses challenges for quality assessment, especially regarding the uniform distribution of ingredients and the variability of raw materials. In this sense, the present study aims to use hyperspectral imaging (HSI) with near infrared (NIR) spectroscopy for characterization and assist in the product design [2]. Hyperspectral images were acquired for six different sausage recipes using a line scanner camera Specim FX17, ranging from 900 to 1700 nm with3.5 nmof spectral resolution. Two different light sources were also used: traditional halogen lamps and a Supercontinuum laser (NKT Photonics A/S, Denmark). The use of the Supercontinuum laser represents an innovative approach in HSI for food analysis, as it minimizes sample heating, a common drawback of halogen lamps when analysing heat-sensitive matrices. The data analysis was performed with chemometric algorithms such as principal component analysis (PCA) and multivariate curve resolution (MCR) aiming to identify systematic variations in samples, assisting to identify chemical variations within samples. Preliminary results showed the great capability of NIR-HSI systems for the characterization of such sausages, indicating tendencies related to content of meat and plant-based products in different levels. MCR also allowed to identify the distribution of such ingredients within the sausages and their relative concentrations. The use of the Supercontinuum laser showed promising results, although further investigations are still needed. These findings highlight the potential of NIR-HSI as a non-destructive tool for the characterization of hybrid meat products, contributing to the development of more sustainable and transparent food production systems.

Session 2 - Design of Experiments

16-06-2025 - **15:20 - 16:45**

Quality by design approach to improve quality and decrease cost of mRNA production

Henrik Widmark [1]

[1] Sartorius, Umeå, Sweden

Introduction

The SARS-Cov-2 pandemic has contributed to accelerated research of messenger RNA (mRNA) based vaccines. Increased interest in mRNA has been further driven by promising therapeutic applications such as cancer immunotherapies, protein replacement therapies, regenerative medicine and cellular reprogramming.

Purpose

The in vitro transcription (IVT) reaction is an enzymatic-catalyzed production of messenger RNA (mRNA) from a DNA template and is the unit operation with the highest cost of goods in the mRNA drug substance production process. The purpose of this investigation was to increase the cost-efficiency of the IVT process and reduce unwanted by-products. To accomplish this a Quality by Design (QbD) approach was used to study factors that influence process yield (in g/L), reduce the amount of impurity (dsRNA) and increase raw material efficiency.

Methods

A Design of Experiments (DOE) based framework for modelling reaction profiles will be presented showing how the model can be generalized to be used for estimating reaction profiles of mRNA constructs of different lengths.

The main steps of the investigative approach are:

* Iterative model development

* Detection of model validity issues

* Design Space (DSp) estimation

* Model based cost optimization

* Experimental verification of design space

Results

We successfully identified and experimentally verified a design space. Optimization within this design space led to an increase in cost efficiency by 44% and a reduction of impurities by 75%. Above all, we were able to push reaction boundaries, by reaching a process yield of 25 g/L corresponding to a doubling of the highest yield reported in literature, so far.

Conclusions

The work shows how data driven models together with process knowledge can help scientists to push boundaries, increase process understanding and create better processes for the biopharmaceutical industry.

General Effect Modelling (GEM) – a platform for analyses of multivariate data influenced by several qualitative and quantitative factors

Ellen Mosleth [1]Kristian H Liland [2]

[1] Ellen Mosleth, Aas, Norway
[2] Norwegian University of Life Sciences, Aas, Norway

General Effect Modelling (GEM) – a platform for analyses of multivariate data influenced by several qualitative and quantitative factors

Mosleth, Ellen F.[1] and Liland, Kristian Hovde[2*]

[1] Nofima AS, Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, 1430 Ås, Norway

[2] Faculty of Science and Technology, Norwegian University of Life Sciences, 1430 Ås, Norway

[*)] Corresponding author

Abstract

Introduction: Multivariate data influenced by one or more design variables is challenging. Different methods have been developed to handle this situation, although they are often restricted in the downstream analyses.

Purpose: The purpose of this publication is to present a flexible platform for analysing such multivariate data influenced by several factors.

Methods: We here present General Effect Modelling (GEM) as an umbrella enabling a broad platform of methods developed for multivariate response data influenced by multiple factors (1-3) as an extension of Effect plus Residual modelling (4, 5). In GE, these factors may be design factors or any other qualitative or quantitative factors influencing the data at hand. The first step in GEM is a linear model which outputs the effects of each factor, the residuals of the complete model, and Effect+Residual (ER) values where the residuals of the complete model are added to each effect. These ER values can be used for any univariate and multivariate downstream analyses suitable for the data at hand. ER modelling is presented and compared with various established methods as benchmarkings.

Results: The ER values obtained by GEM contain detailed information on the effects of each design variable and validation of the results by different methodologies that were not available by other benchmarking methods, and it has the flexibility to apply any subsequent data analyses. We demonstrate how PCA, PLS, Elastic net and neural networks elucidate different aspects of the data and their factors.

Conclusion: GEM can be considered an umbrella enabling a broader platform of methods developed for multivariate response data, encompassing other tested and untested options. GEM can be applied to any scientific field.

References:

[1] Mosleth EF, Liland KH. General Effect Modelling (GEM) -- Part 1. Method description. arXiv:240403024 **2024**.

[2] Mosleth EF, Liland KH, Dankel SN, Mellgren G, Barajas-Olmos F, M., Orozco L, et al. General Effect Modelling (GEM) -- Part 2. Multivariate GEM applied to gene expression data of type 2 diabetes detects information that is lost by univariate validation. arXiv:240403029. **2024**.

[3]. Mosleth EF, Myhr K, Vedeler CA, Berven FS, Lysenko A, Gavasso S, et al. General Effect Modelling (GEM) -- Part 3. GEM applied on proteome data of cerebrospinal fluid of multiple sclerosis and clinically isolated syndrome. arXiv:240403034. **2024**.

[4]. Mosleth EF, McLeod A, Rud I, Axelsson L, Solberg L.E., Moen B., et al. Analysis of Megavariate Data in Functional Omics. In: Brown S, Tauler, R. W, editors. Comprehensive Chemometrics,. 4. 2nd ed: Elsevier; **2020**. p. 515-67.

[5]. Mosleth EF, Vedeler CA, Liland KH, McLeod A, Bringeland GH, Kroondijk L, et al. Cerebrospinal fluid proteome shows disrupted neuronal development in multiple sclerosis. Sci Rep-Uk. **2021**;11,4087.

GEMANOVA: a multiway approach to DoE

Jokin Ezenarro [1]Daniel Schorn-García [2], Morten A. Rasmussen [1], Rasmus Bro [1], Olga Busto [3], Ricard Boqué [3]

[1] University of Copenhagen, Frederiksberg C, Denmark
[2] Stellenbosch University, Stellenbosch, South Africa
[3] Universitat Rovira i Virgili, Tarragona, Spain

Design of Experiments (DoE) often involves complex factor interactions that are challenging to capture and interpret using traditional additive models. These models, while effective for simple experimental designs, struggle to accurately represent the intricate, non-linear relationships that frequently occur in multiway datasets. Generalised Multiplicative ANOVA (GEMANOVA) addresses these limitations by extending beyond conventional ANOVA-based methods, such as ANOVA-Simultaneous Component Analysis (ASCA) and Parallel Factor Analysis with ASCA (PARAFASCA) [1]. Unlike these additive models, GEMANOVA incorporates a multiplicative framework, which allows for a direct analysis and interpretation of high-order interactions among experimental factors [2].

Despite its advantages, GEMANOVA has not yet achieved widespread adoption in the chemometrics community. A significant barrier to its broader application is the absence of standardised validation tools, which limits the interpretability and reliability of its results compared to its additive counterparts. Traditional validation techniques used for ASCA and PARAFASCA, such as cross-validation and permutation testing, are not directly transferable to GEMANOVA due to its fundamentally different mathematical structure [2]. This gap has hindered the confidence of researchers in deploying GEMANOVA for critical data analysis tasks.

To address this challenge, we propose a novel validation methodology tailored for GEMANOVA based on permutation testing; a non-parametric statistical method that enables the empirical assessment of factor significance without relying on stringent parametric assumptions. This methodology systematically permutes individual factors within the dataset while preserving the overall experimental structure dictated by the DoE. By generating a distribution of permuted loadings, we create a robust reference framework against which the original GEMANOVA model can be compared. This comparison allows for the quantification of the statistical significance of observed effects, thereby enhancing the model's reliability and interpretability.

In addition, through the quantification of the pseudovariances represented by each mode on the GEMANOVA model, we propose a way to quantify the effect of the experimental factors on the measured data; a metric that is of outmost relevance in DoE-based models. Furthermore, GEMANOVA inherits the ability of PARAFAC to handle missing data efficiently. This feature makes GEMANOVA uniquely capable of accommodating incomplete DoE designs, where certain factor combinations may be unmeasurable due to practical constraints such as resource limitations, experimental failures, or ethical considerations. This flexibility is invaluable in real-world experimental settings, where achieving a fully balanced factorial design is often impractical.

Ultimately, our work contributes to the broader goal of improving data-driven decision-making through DoE. The proposed framework not only enhances the robustness of GEMANOVA but also ensures that it can be applied with the same level of confidence as more established methods like ASCA and PARAFASCA.

[1]: Guisset S.; Martin M.; Govaerts B.; Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs. Chemom. Intell. Lab. Syst. **2019**, *184,* 44–63.

[2]: Bro R.; Jakobsen M.; Exploring complex interactions in designed data using GEMANOVA. Color changes in fresh beef during storage. J. Chemom. **2002**, *16*, 294–304.

Optimizing bottom-up proteomics sample preparation for absolute quantification of 170 plasma proteins in human plasma using a chemometric approach

Kári Arnarson [1, 2]Kristrún Ýr Hólm [1, 3], Valdís Gunnarsdóttir [1, 3], Sigríður Klara Böðvarsdóttir [1, 3], Yassene Mohammed [4], Christoph H Borchers [5], Finnur F Eiriksson [1, 2], Margret Thorsteinsdottir [1, 2]

[1] Faculty of Health Sciences, Reykjavík, Iceland
[2] ArcticMass, Reykjavík, Iceland
[3] BioMedical Center, University of Iceland, Reykjavík, Iceland

[4] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands
[5] Department Oncology, Faculty of Medicine, McGill University, Montreal, Canada

Introduction

Bottom-up proteomics utilizing ultra-performance liquid chromatography multiple reaction monitoring mass spectrometry (UPLC-MRM-MS) combined with stable isotopic labelled internal standard (SIS), is a widely used method for absolute quantification of proteins for clinical research. However, in-solution digestion of liquid biopsies often involves several steps with low reproducibility and limited throughput. To overcome these challenges, it is essential to develop more efficient and robust sample preparation techniques. A systematic approach is required to investigate how various experimental factors influence protein digestion, enabling the identification and optimization of significant experimental parameters to improve reproducibility and throughput[1].

Purpose

The aim of the study was to optimize experimental factors affecting in-solution bottom-up sample preparation using design of experiments DoE for increased throughput of absolute quantification of 170 proteins in human plasma.

Methods

A two-level full factorial (FF) design was used for experimental screening of digestion time, digestion temperature, enzyme-to-substrate (E:S) ratio and denaturing agent. Two experimental domains were generated for the FF design, with and without the addition of 10 mM Ca_2Cl to the trypsin digestion buffer. Significant experimental factors were further optimized using a central composite face (CCF) design. Both designs used endogenous peptide/SIS peptide peak area ratio for 170 proteins in a pooled human plasma as response, analyzed on Waters Acquity UPLC coupled to Xevo TQ-XS mass spectrometer. Both designs were generated using MODDE 13 and data analysis performed using Skyline, MODDE 13, SIMCA Pro-17 and RStudio.

Results

The experimental screening from the FF designs revealed that all experimental factors in the experimental domains had significant effect for multiple proteins, indicating that these factors cannot be independently adjusted to obtain optimal conditions. The FF domain without the addition of Ca_2Cl to the buffer yielded higher response compared to using a trypsin digestion buffer with Ca_2Cl. Further, using guanidine-HCl instead of urea as a denaturing agent resulted in higher response for most proteins. The CCF design was therefore generated with guanidine-HCl as denaturing agent and without the addition of Ca_2Cl to the trypsin digestion buffer. The optimized conditions for the majority of proteins were 4 hours digestion with E:S 1:1 (w/w) at 37°C. Comparing responses of 170 proteins for 4 hour digestion with E:S 1:1 (w/w) to 18 hour digestion with E:S 1:20 (w/w), the response increased for 33 proteins, decreased for 21 proteins and was not significantly impacted for 116 proteins. The protein coverage was not impacted by the change from 18 hours to 4 hours digestion time.

Conclusions

Bottom-up sample preparation was successfully optimized from 18 hours digestion time to 4 hours, without loss of protein coverage, using DoE. Significant interactions were observed between the experimental factors, highlighting the importance of using DoE for optimization of bottom-up proteomic sample preparation.

References

(1) Thorsteinsdóttir, U. A.; Thorsteinsdóttir, M. Design of Experiments for Development and Optimization of a Liquid Chromatography Coupled to Tandem Mass Spectrometry Bioanalytical Assay. *J Mass Spectrom* **2021**, *56* (9), e4727. https://doi.org/10.1002/jms.4727.

Robust process design through MODDE

Ásta Rós Sigtryggsdóttir [1]

[1] Alvotech, Reykjavik, Iceland

Alvotech is dedicated to expanding access to essential biologic medicines globally by developing and manufacturing high-quality, cost-effective biosimilars. This initiative aims to ensure that life-saving treatments are available to a broader patient population. The implementation of Quality by Design (QbD) and Design Space principles is critical in this endeavor. These principles emphasize the importance of understanding and controlling Critical Quality Attributes (CQA) and Critical Process Parameters (CPP). Establishing a robust control strategy is essential for successful process validation, ensuring that the biosimilars produced meet stringent quality standards.

This presentation will demonstrate how MODDE was applied to biotechnological process development, using case studies to highlight its role in identifying critical process parameters, optimizing yield, and ensuring process consistency. Leveraging DoE in MODDE enabled improved process understanding, reduction in variability, and greatly accelerated development timelines. The insight gained from these approaches contributes to more reliable bioproduction methods, ultimately supporting scalable and cost-effective biotechnology applications.

Session 3 - Advanced data processing

17-06-2025 - 09:00 - 12:00

Aspects and Benefits of Applying Convolutional Neural Networks to Small Data Sets – A Viability Study

Andreas Baum [1, 2]

[1] Manufacturing Intelligence, Novo Nordisk Engineering, Virum, Denmark
[2] Technical University of Denmark, Kgs, Lyngby, Denmark

This study [1] addresses a domain-specific problem in an industrial flocculation process, initially approached using traditional image analysis. Using these results as a baseline, we introduce a novel model combining a convolutional neural network (CNN) with a Projection to Latent Structures (PLS) component. We evaluate three CNN architectures with varying degrees of regularization and compare the results with logistic regression models based on inputs obtained through traditional image analysis. By integrating a PLS-like regularization framework into the CNN, we succeed in learning latent variable representations using only 117 raw images. Furthermore, we demonstrate perfect interpretability of the latent variable space and the resulting CNN activation patterns.

[1] Baum, A.; Moiseyenko, R.; Glanville, S.; Jørgensen, T.M. Image-based Characterization of Flocculation Processes through PLS Inspired Representation Learning in Convolutional Neural Networks. *J. Chemom.* 2025, **39**, e3534. https://doi.org/10.1002/cem.3534

All sparse PCA models are wrong, but some are useful

Age K. Smilde [1], Edoardo Saccenti [2], Johan A. Westerhuis [1], Rasmus Bro [3]

[1] University of Amsterdam, Amsterdam, Netherlands
[2] University of Wageningen, Wageningen, Netherlands
[3] University of Copenhagen, Copenhagen, Denmark

There is an increased interest in sparse Principal Component Analysis (sPCA) [1,2] for reasons of simplicity and interpretability. The results discussed in this presentation have been described in a series of three papers. In the first part, we demonstrate that sPCA models have limitations with respect to factorizing sparse and noise-free data accurately when loadings are overlapping. In the second part, we show that sPCA algorithms based on deflation can generate artifacts in higher components. We also show that scores orthogonalization and the incorporation of orthogonal loadings are suitable means to avoid large artifacts. Both approaches constrain the set of possible sPCA solutions in a similar way that is both poorly characterized and understood. In the third part, we study the sPCA solution by Zou et al [2], which according to our results represent the best sPCA algorithm of those considered in the series. We provide new derivations on the model equations, the computation and interpretation of the model parameters and the selection of meta-parameters in practical cases, making sPCA an even more powerful data modelling tool.

During the presentation, results concerning multivariate factorizations other than sPCA will be also highlighted. Of special interest are the generation of artifacts when combining constraints and deflation, and the unexpected connection between sPCA and Partial Least Squares (PLS).

[1] Jolliffe, et al. Journal of computational and Graphical Statistics 12, no. 3 (2003): 531-547.

[2] Zou, et al. *Journal of computational and graphical statistics* 15, no. 2 (2006): 265-286.

A rigorous Sparse-PLS workflow using multiple imputation of missing values

Marta Bevilacqua [1]José Camacho Páez [2], Age Klaas Smilde [1], Giorgio Tomasi [1]

[1] University of Copenhagen, Copenhagen, Denmark
[2] University of Granada, Granada, Spain

Sparse partial least squares (sPLS)[1] is a regression method that combines the dimension reduction capabilities of PLS with variable selection to improve model interpretability and prediction accuracy. In this work, the method was tested on an environmental chemistry data set in which contaminant concentrations are used, together with process variables, to predict the toxicity of wastewater samples observed on in-vivo and in-vitro assays. The rationale for using sPLS was that one of the main aims of the study was to establish which variables were responsible for the toxic response.

The training set is comprised of 169 samples collected over the course of 13 weeks in 2023 and validated on a set of samples collected in the same plant, one year later, for three weeks. One additional challenge of this data set that is shared with many other practical applications is that both **X** and **y** are censored: the predictors are left censored because the concentrations can be below the analytical detection limit, while the predicted variables are right censored because no toxicity was detected at the tested ranges of concentration.

We present here a rigorous workflow for the selection of sPLS meta-parameters, such as the number of latent variables and the number of variables to keep in the model, using double cross-validation and multiple imputation to address the censored data issue.

Results will be presented also based on simulations that include several factors: ratio between number of samples and number of variables, **X** noise type and level, **y** noise level, number of rows, presence of clusters of correlated variables in **X**, level and type of correlation within each cluster, percentage of missing values. For comparison, the performance of the sPLS workflow will be compared with other related methods: group PLS, group LASSO, Elastic Net[2-4].

The calculations were run on Matlab using the MEDA package[5] and in-house scripts.

[1] K.-A. L. Cao, D. Rossouw, C. Robert-Granié, and P. Besse, “A Sparse PLS for Variable Selection when Integrating Omics Data,” *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, Nov. 2008, doi: 10.2202/1544-6115.1390.

[2] J. Camacho and E. Saccenti, “Group-wise partial least square regression,” *Journal of Chemometrics*, vol. 32, no. 3, p. e2964, 2018, doi: 10.1002/cem.2964.

[3] H. Zou and T. Hastie, “Regularization and Variable Selection Via the Elastic Net,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[4] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A Sparse-Group Lasso,” *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, Jan. 2013, doi: 10.1080/10618600.2012.681250.

[5] J. Camacho, A. Pérez-Villegas, R. A. Rodríguez-Gómez, and E. Jiménez-Mañas, “Multivariate Exploratory Data Analysis (MEDA) Toolbox for Matlab,” *Chemometrics and Intelligent Laboratory Systems*, vol. 143, pp. 49–57, Apr. 2015, doi: 10.1016/j.chemolab.2015.02.016.

Data Quality: The importance of the ‘before analysis’ domain on data modelling [BLM, ML, AI] - What can chemometricians do?

Kim H. Esbensen [1]

[1] KHE Consulting (KHEC), Copenhagen, Denmark

Data analysts/chemometricians are part of a scientific collegium covering three domains [sampling / analysis / data modelling, which are collectively responsible for ‘data quality’. The three-domain foundation has serious consequences for data modelling in the chemometrics domain – and is also shown to have serious implications for the current PAT paradigm, the foundation of which turns out to need significant reform regarding a key process sampling aspect (regardless of whether physical samples, or PAT sensor technology spectra, are extracted/acquired).

The ‘before analysis’ domain is prone to sampling errors resulting in uncertainties influencing the quality and relevance of both analysis and data analysis, data modelling, chemometrics [BLM, ML, AI]. Non-representative sampling of *heterogeneous* materials, batches, lots and process streams ‘before analysis’ contribute significantly to the total measurement uncertainty, $MU_{total} = MU_{sampling} + MU_{analysis}$. The total sampling error (TSE) can dominate over the total analytical error (TAE) by factors ranging 5, 10 or *higher*, depending on the *degree* of material heterogeneity encountered and the *specific* sampling procedure employed to produce the final analytical aliquot, which is the only material actually analysed. The analytical aliquot is the physical manifestation of transgressing the boundary from the sampling domain to the domain of analysis. It is only possible to guarantee representativity of the analytical aliquot, and thus of the analytical results with respect to the *original* target batch/lot/process stream, by invoking the necessary domain competence stipulated by Theory of Sampling (TOS). If the sources of adverse sampling effects have not been eliminated, the sampling process is *biased* and MU_{total} will always be unnecessarily inflated. TOS offers ways and means to deal actively with a potential sampling bias, which is fundamentally different from the analytical bias. Overlooking, or deliberately ignoring dealing appropriately with sampling effects constitutes a lack of due diligence, which has critical bearings on the QC/QA demands on both analysis and data analysis/modelling/chemometrics [1].

[1] Esbensen, K. (2025) Data Quality: Importance of the ‘Before Analysis’ Domain - Theory of Sampling (TOS)]. Journal of Chemometrics, 39: e70021. https://doi.org/10.1002/cem.70021

Enhancing Multiclass Classification with OPLS-HDA

Edvin Forsgren[1], Benny Björkblom[2], Johan Trygg[1, 3], Pär Jonsson[3]

[1] Computational Life Science Cluster (CLiC), Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden
[2] Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden
[3] Sartorius Corporate Research, SE-903 33 Umeå, Sweden

Multiclass data sets are becoming more common in omics sciences, drug discovery, and clinical research due to advancements in analytical platforms. However, they present significant challenges in handling and interpreting the data efficiently. Traditional methods often require breaking down multiclass comparisons into multiple two-class analyses, which is labor-intensive and complex. To address these challenges, we introduce orthogonal partial least squares-hierarchical discriminant analysis (OPLS-HDA). This innovative approach combines hierarchical cluster analysis with the OPLS-DA framework, creating a decision tree that clearly visualizes relationships between classes. OPLS-HDA simplifies the model-building process, providing a streamlined method for analyzing multiclass data sets. It offers clear insights into subtle differences between classes, enhancing the interpretability of complex data. This makes OPLS-HDA a versatile tool applicable across various fields. This advancement represents a significant step forward in multiclass data analysis, offering researchers a powerful and user-friendly solution to efficiently dissect and understand intricate data sets.

Alternative definitions of effects in path models with multidimensional blocks

Tormod Næs [1]Rosaria Romano [2], Oliver Tomic [3], Age K. Smilde [4], Kristian Hovde Liland [3]

[1] Nofima, Ås, Norway
[2] Federico II University of Naples, Naples, Italy
[3] Norwegian University of Life Sciences, Ås, Norway
[4] University of Amsterdam, Amsterdam, Netherlands

Introduction

Quantifying the effects of one node on another node in path modelling is well-defined in univariate analyses but is a more open problem when the nodes are multivariate. Attempts have been made before, though none are fully transparent and intuitive. We propose a new definition which is motivated by simple orthogonalisations and then generalised to flexible regression.

Purpose

The purpose of the study is to define an intuitive set of path effects for multidimensional blocks that also makes sense in the unidimensional case.

Methods

Three regressions are defined in the presence of an input block, an output block, and a set of blocks pointing to the output block, leading to definitions of the total effect, unique effect, interaction effect and additional effect.

Results

We will demonstrate results from simulations elucidating various aspects and real data. This also shows practical considerations for rank-deficient cases.

Conclusions

A definition is made available, and its consequences, strengths and weaknesses are demonstrated.

The Assay of Theseus - Behavioral Dynamics of xC-MS in the Real World

Jim Edwards [1]

[1] Indigo BioAutomation, Carmel, United States

Releasing high quality chromatography and mass spectrometry quantitative results at a high operational tempo with high confidence requires the active coordination of a number of agents and systems in the laboratory. The core to collaborative software automation in analytical chemistry for this type of work is “the assay” – the collection of preparation, acquisition, processing, and assessment (as well as human review/release) activities which take a set of samples and ultimately produces a set of trustworthy, valuable results. Focusing more specifically on data processing, and starting from the context of a validated assay, one key question is “when is an assay no longer ‘the same’ assay?” It’s not uncommon in the lab for an operator to make small adjustments, to address the analytical reality of running instrumentation under high throughput conditions. This begs the question: how much of a change is too much of change, potentially altering the validated state of the assay? Recent regulatory changes describe the increasing requirements in documenting the system’s verification and validation. It is becoming increasingly important to treat validation not as something that is “done” but rather, something that is always happening (and as much as possible, automatically). This talk will explore the capabilities of retrospective and real-time analytics as an aid in characterizing, documenting, and advancing “the assay” over its lifetime of utility in the lab.

Study on the interpretability of Kernel-based multivariate tools

Zina-Sabrina Duma [1]Tuomas Sihvonen [1], Sara Heikkinen [1], Satu-Pia Reinikainen [1]

[1] LUT University, Lappeenranta, Finland

Kernel-based chemometric methods are non-linear extensions of traditional multivariate statistical tools. Kernel Partial Least-Squares Regression (K-PLS) and Kernel Principal Component Regression (K-PCR) are effectively used when the relationship between input data and output is non-linear, and traditional methods exhibit inconsistent performance across the prediction range [1, 2]. However, one of the major drawbacks of nonlinear methods is their lack of interpretability; for instance, by-products such as loadings or regression coefficients do not have a meaning in the original data space. On the other hand, kernel-based multivariate methods often involve fewer learned parameters compared to deep learning approaches. With fewer parameters to tune, these models tend to be more interpretable. In this study, we assess existing methods for evaluating variable contributions to non-linear models and propose novel approaches. Our proposed indicators are based on kernel space covariance studies and the behavior of learning parameters. These methods are tested using both synthetic and real data case studies.

[1] - Duma Z. S.; Susiluoto J.; Lamminpää O.; Sihvonen T.; Reinikainen S. P.; Haario H. (2024). KF-PLS: Optimizing kernel partial least-squares (K-PLS) with Kernel Flows. *Chemometrics and Intelligent Laboratory Systems*, *254*, 105238.

[2] - Duma Z. S.; Sihvonen T.; Susiluoto J.; Lamminpää O.; Haario H.; and Reinikainen S. P. (2024) "Kernel-Based Retrieval Models for Hyperspectral Image Data Optimized with Kernel Flows," *2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, Helsinki, Finland, 2024, doi: 10.1109/WHISPERS65427.2024.10876476.

Session 4 - Deep learning, machine learning and chemometrics

17-06-2025 - 13:00 - 14:35

Clinical Applications of Ensemble Machine Learning Methods for Multiplex Mass Spectrometry Interpretation

Stephen R. Master [1]Miao He [1], Taylor T. Wild [1], Rebecca D. Ganetzky [1]

[1] Children's Hospital of Philadelphia, Philadelphia, United States

Highly multiplexed data streams, such as those derived from multianalyte mass spectrometry assays, are a natural area for the application of modern machine learning techniques to clinical diagnosis. We have recently undertaken the application of ensemble tree-based machine learning methods (random forest and gradient boosted trees) to address two clinical questions: first, whether machine learning can reliably automate the interpretation of a multiplexed plasma protein N-glycan profiling assay for the diagnosis of congenital disorders of glycosylation CDG); second, whether multiplex amino acid and acylcarnitine data can accurately predict the presence of inborn errors of energy metabolism (IEEM) in patients presenting for emergent treatment of lactic acidosis. Historical data for both relevant clinical problems were collated and randomly partitioned into training and test cohorts. Model training and tuning was performed was performed using the training sets, and optimized models were then applied to the previously unanalyzed test cohorts. For N-glycan data obtained from a clinical qTOF-based profiling method, traditional requires requires expert interpretation by a trained biochemical geneticist. Our machine learning models were highly effective at discriminating normal from abnormal samples (ROC-AUC = 0.979) as well as in subtyping of different forms of CDG (normal vs. PMM2 vs. non-PMM2 Type I vs. Type II, Hand-Till AUC=0.967). Variable importance analysis demonstrates a mixture of known informative analytes as well as analytes previously not linked to diagnosis and subtyping. In the second example, patients presenting emergently with hyperlactatemia on whom amino acid and acylcarnitine analysis was performed were utilized to construct a model predicting the presence of an inborn error of energy metabolism, which is associated with higher mortality. After model construction and optimization consistent with recent guidelines from and IFCC working group on machine learning for clinical laboratory use [1], the resulting model was assessed on an independent test set. The AUC-ROC of the resulting gradient boosted tree model was 0.81, which is significantly better than peak lactate alone (AUC-ROC 0.81 vs. 0.56, P=0.027) in predicting IEEM. Taken together, we believe that these results demonstrate the potential clinical utility of machine learning for interpreting highly multiplex clinical assay results across a diverse set of medical applications.

[1] Master S.R.; Badrick T.C.; Bietenbeck A.; Haymond S. Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group. Clin Chem. **2023**, *69*, 690-698.

Identification of plasma proteins for breast cancer diagnosis by integrating targeted proteomics with clinical data

Kristrun Yr Holm [1, 2]Yassene Mohammed [3, 4], Finnur Eiriksson [1, 5], Christoph H Borchers [4], Sigridur Klara Bodvarsdottir [2], Margret Thorsteinsdottir [1, 5]

[1] Faculty of Pharmaceutical Sciences, University of Iceland, Reykjavik, Iceland
[2] Biomedical Center, University of Iceland, Reykjavik, Iceland
[3] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands
[4] Department Oncology, Faculty of Medicine, McGill University, Monreal, Canada
[5] ArcticMass, Reykjavik, Iceland

Breast cancer (BC) is the most prevalent cancer in women and ranks as the second leading cause of cancer-related deaths. The key to improve survival rates lies in the early detection of BC, highlighting the critical role of screening methods. Blood-based biomarkers may offer an alternative non-invasive strategy to improve BC screening, with better sensitivity than the routinely used X-ray mammography. For this to become a reality, analytical methods for the identification and quantification of BC biomarkers need to be improved. This study aims to identify protein biomarkers with diagnostic and prognostic value in human plasma by multivariate analysis of targeted proteomics data to improve early BC diagnosis.

An absolute quantification of 131 proteins was performed on 270 well-defined Icelandic biobank-based plasma samples from 135 BC patients and 135 healthy controls. Among the BC patients, one-third were BRCA2 mutation carriers. Absolute quantification of the proteins was conducted using PeptiQuant human protein kit with UPLC-MRM-MS analysis. Prior to analysis, the plasma samples were proteolytically cleaved with trypsin, internal standards were added to the peptide mixture, and concentrated by solid-phase extraction, using a liquid-handling robot. Data analysis was conducted using MassLynx and Skyline and SIMCA Pro-17, R studio, and Python for statistical analysis, multivariate data analysis, and machine learning. Principal component analysis (PCA) and predictive models, including orthogonal partial least squares discriminant analysis (OPLS-DA) and logistic regression, were developed and validated through cross-validation.

The targeted proteomic assay was successfully implemented for the absolute quantification of 131 proteins in human plasma samples with precision and accuracy for calibration standards and quality controls within 20% relative standard deviation. Out of the 131 proteins, 98 proteins were quantifiable in the Icelandic bio-bank plasma samples, surpassing the lower limit of quantification. The samples were analyzed in eight batches, each containing matched pairs of cases and controls. Following data acquisition, the data was normalized, and minimal batch effects were corrected using ComBat, ensuring accurate comparisons across all samples in downstream analysis. Differential protein expressions were evaluated using PCA and OPLS-DA, logistic regression models, and statistical analysis to assess variations in protein concentration between cases and controls. Considering the heterogeneous nature of BC, incorporating clinicopathological variables such as BC subtype, tumor size, histological grade, receptors, and age appear to be important for enhancing model specificity. We identified several proteins that were significantly upregulated in BC cases, particularly in those with positive nodal metastasis, tumors over 20 mm, and high histological grade. Subtype-specific protein differences were observed in Luminal B, HER2, and triple-negative BC. Additionally, protein variations were observed in BC patients with a germline BRCA2 mutation and in BC patients with tumors of sporadic origin.

Targeted proteomics using UPLC-MRM-MS shows potential for identifying and quantifying protein biomarkers in human plasma for the diagnosis of BC, particularly when categorized according to clinical data such as BC subtype or the presence of germline mutation.

Prediction Uncertainty Quantification in PLS Regression: A Gaussian Process Framework for Observation-Specific Prediction Bias and Prediction Error Variance Estimation

Carl Emil Eskildsen [1, 2]Kevin Giraldo [2], Georgios Papadopoulos [2, 3], Mauricio Barahona [3], Molly Stevens [1, 2]

[1] Department of Materials, Department of Bioengineering, Imperial College London, London, United Kingdom
[2] The Kavli Institute for Nanoscience Discovery, University of Oxford, Oxford, United Kingdom
[3] Department of Computing, Imperial College London, London, United Kingdom

Introduction: Latent variable regression models like partial least squares (PLS) are pivotal in chemometrics but lack robust, observation-specific uncertainty quantification [1]. Conventional PLS error estimation, rooted in ordinary least squares (OLS), assumes unbiased predictions, leading to unreliable confidence intervals [2]. Accurate uncertainty assessment is critical in e.g. medical diagnostics and process control, where decisions hinge on prediction accuracy and risk [1].

Purpose: This study introduces a Gaussian process (GP)-based framework to estimate observation-specific prediction bias (systematic errors) and prediction error variance (random dispersion) in PLS, addressing OLS limitations and enhancing reliability in high-stakes applications.

Methods: A two-constituent model system (analyte: fructose, interferent: sucrose) was prepared in a 5×5 full-factorial design with triplicates and analysed using Raman spectroscopy (75 calibration measurements). A PLS model linked spectral data to fructose concentrations, compressing measurements into latent scores. These scores and associated PLS prediction errors trained a GP model to map prediction bias and prediction error variance across the latent space. A total of 37 validation samples (each with 30 replicates) were used to validate the GP-derived observation-specific uncertainty estimates against empirical observations.

The GP model employs two key functions [3]:

1. A ***mean function*** identifying systematic prediction biases (e.g., consistent over/underestimation in specific latent space regions).

2. A ***variance function*** quantifying prediction error dispersion, capturing uncertainties from model instability and measurement noise.

Results: The GP framework revealed spatially structured prediction biases across the PLS latent space: the PLS model exhibited positive prediction bias in low-interferent/low-analyte and high-interferent/high-analyte regions, and negative prediction bias in high-interferent/low-analyte and low-interferent/high-analyte regions. Bias magnitudes diminished toward the latent space center (Figure 1A). Empirical validation confirmed these spatial trends, with GP-derived bias correction reducing the mean squared error (MSE) of validation samples by 75% compared to uncorrected PLS predictions.

For prediction error variance, GP estimates aligned with theoretical expectations, predicting higher uncertainty at latent space extremes (Figure 1B). However, empirical variances (from validation samples) did not exhibit such behaviour. The empirically observed variance showed minimal dispersion (0.5E-6 to 2E-6). Meanwhile, the GP-estimated variances (~1.8E-6), remain within a reasonable range relative to the empirical values.

Conclusions: This GP framework provides spatially resolved prediction bias and prediction error variance estimates, significantly enhancing PLS reliability. By explicitly quantifying observation-specific uncertainties, the method supports confident decision-making in applications requiring precise predictions, such as medical diagnostics.

References

[1] Skou P.B.; Tonolini M.; Eskildsen C.E.; Berg F. van den; Rasmussen M.A. Unbiased prediction errors for partial least squares regression models: Choosing a representative error estimator for process monitoring. J. Near Infrared Spectrosc. 2023, 31, 186–195. doi:10.1177/09670335231173139.

[2] Eskildsen C.E.; Næs T. Sample-Specific Prediction Error Measures in Spectroscopy. Appl. Spectrosc. 2020, 74, 791–798. doi:10.1177/0003702820913562.

[3] Rasmussen C.E.; Williams C.K.I. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA, 2006; ISBN 0-262-18253-X.

IPA under the light: Explainability study (XAI) of Deep CNN for near-infrared spectroscopy

Florent Haffner [1]Marion Lacoue-Negre [1], Aurélie Pirayre [2], David Gonçalves [1], Maxime Moreaud [1]

[1] IFP Energies nouvelles, Solaize, France
[2] IFP Energies nouvelles, Rueil-Malmaison, France

Inception for Petroleum Analysis (IPA) [1] is a deep convolutional network inspired from state-of-the-art computer vision architectures. Based on several computational blocks, IPA showed improved performance, compared to PLSwithout depending on complex pre-processing operations. The network begins with three stacked convolutions, followed by a multi-branch module consisting of four different paths that are concatenated. Therefore, the model learns complex but complementary, internal representations. Digging inside eXplainable Artificial Intelligence (XAI) can bring understanding on complex models and help to envisage deep networks as robust and performant alternatives to traditional PLS modelling. Two studies using gradient-weighted class activation mapping (Grad-CAM) [2,3], already have been applied to shallow CNNs, extending these methods seems essential for DL to earn the confidence of the chemometrics community. The application of XAI methods [4,5] for NIR analysis can be greatly improved and is crucial for many practical applications such as oil characterization, agriculture, and environmental monitoring.

This new framework is built on a two-step analysis applied to IPA. Explainability has been positioned at several points within the architecture to demonstrate the complementarity of each computational blocks. First, it was essential to understand the features globally influential for the model, considering the chemical aspect necessary for the property

of interest to impartially judge the quality of the deep CNN. Then, layer-wise analysis was carried out to determine the contribution of each layer to the overall model performance.

Feature importance analysis showed that the PLS and IPA put importance on the same spectral regions, with the PLS's interest is restricted and diffuse, IPA's interest is smoother and more moderate. The layer-wise analysis demonstrated the performance of IPA comes from representation learned by its complementary multi-branch layers inspired by the Inception model. It has been demonstrated, without any prior knowledge or pre-processing, the network focus on relevant spectral region in terms of physico-chemical in relation to the full range of cetane number. In addition, this XAI methodology illustrated these types of models are not only performant but also more robust than shallower CNNs and PLS. Specifically, IPA better identifies the important spectral regions in relation to the molecular composition of the product that influence the cetane number, such as aromatic content as well as the length of the hydrocarbons molecules.

References:

[1] Haffner F. et al.; IPA: A deep CNN based on Inception for Petroleum Analysis. Fuel **2025**, *379*, 133016.

[2] Yang J. et al.; An interpretable deep learning approach for calibration transfer among multiple near-infrared instruments. Computers and Electronics in Agriculture **2022**, *192*, 106584.

[3] Pasos D.; Mishra P.; An automated deep learning pipeline based on advanced optimisations for leveraging spectral classification modelling. Chemometrics and Intelligent Laboratory Systems **2021**, *215*, 104354.

[4] Lundberg S. et al.; Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering **2018**, *10*, 749-760.

[5] Selvaraju, R. et al.; Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision **2020**, *128*, 336-359.

Quantitative BIG DATA and Chemometrics: Green, Safe and Understandable AI for Natural Science and Technology

Harald Martens [1]

[1] Idletechs AS, Trondheim, Norway

Introduction: There is an urgent need for better methodology and culture to convert today's and tomorrow's torrents of technical measurements into rapid, reliable and relevant information. For better understanding and performance, we need to combine our old theoretical and practical KNOWNS with unexpected UNKNOWNS from new measurements. This must be done for maximal usefulness but with minimal computational demands, minimal uncertainty and minimal human ***black box*** alienation.

I believe chemometricians can and should contribute to meet these challenges [1], [2], [3], [4]. This lecture outlines how we work in this respect in Idletechs AS and our Big Data Cybernetics group at NTNU in Trondheim.

A material *fact* is that apples tend to fall downwards, not upwards. To say that apples often fall up is immaterial *fiction.* But even fiction ends up in LMMs based Chat GPT. Compared to e.g. data from the IMMATERIAL world of natural language, multi-channel measured data from the MATERIAL world are usually quite well-structured - luckily.

Purpose: Highly informative, but overwhelming streams of Quantitative BIG DATA are created by high-speed, multi-channel measuring devices in modern natural science and technology.

Examples: *Continuous process monitoring of real-world industrial input materials, production units and output products by fixed Thermal-, RGB- or hyperspectral NIR cameras. Environmental and agricultural monitoring of ground regions by multichannel RGB/VNIR/SWIR/LIDAR/InSAR imaging from repeated passages of satellites or drones. Large sets of human profile data from medical imagers, multichannel MS/NMR/IR/NIR spectrometers, multi-sensor ECG or EEG combinations etc.*

Such high-speed direct measurements from the real world are always cognitively overwhelming, due to the sheer amount of data and to confusing selectivity problems from overlapping causal signals. But they can often give very useful information.

Methods: We chemometricians already have metamodeling- and subspace methods to merge the essence of prior knowledge with the essence of variation patterns in measured data. And we have pre-processing methods to reduce data complexity, e.g. to linearize responses, to balance different variables, to identify and separate the spectral variation patterns of chemical, physical and instrumental nature, to remove hyperspectral image shadows, and to compensate for spatial motions between images or spectral peaks.

Results: Methodologies from Chemometrics and related science cultures have recently been extended to handle overwhelming streams of Quantitative BIG DATA from science and technology. But they need further work and require big changes in software implementation.

Conclusion: When a stream of high-dimensional measurements is well structured, minimalistic chemometrics gives *green, safe and understandable* machine learning.

[1] Martens, H.; Factor analysis of chemical mixtures. Non-negative factor solutions for spectra of cereal amino acids. Analytica Chimica Acta **1979**, *112*, 423-442.

[2] Martens, H.; Quantitative Big Data: Where chemometrics can contribute. Journal of Chemometrics **2015**, *29*(11), 563-581.

[3] Martens, H.; Causality, Machine Learning and Human Insight. Analytica Chimica Acta **2023**, *1277*.

[4] Martens, H.; A Greener, Safer, and More Understandable AI for Natural Science and Technology. J. Chemometrics **2025,** 39, 2 (*27 pages*).

Session 5 - Chemometrics in ”omics” technologies

18-06-2025 - **09:30 - 12:15**

Chemometrics in untargeted Lipidomics: Applications and cautions

Laura Goracci [1]Stefano Bonciarelli [2], Gabriele Cruciani [1]

[1] University of Perugia, Dep. of Chemistry, Biology and Biotechniology, Perugia, Italy
[2] Mass Analytica, Saint Cugat del Vallés , Spain

Lipids play a crucial role in cellular structure and function, including cell signaling, membrane plasticity, and trafficking. Alterations in the lipid composition of cells, tissues, or organelles have been associated with a large number of diseases, including inflammation, cancer, and degenerative or metabolic disorders. Therefore, lipidomics, which can be defined as the large-scale study of lipid species and their related networks and metabolic pathways [1], is now considered a stand-alone discipline, although it can also be classified as a branch of metabolomics.

Among the analytical approaches applied to lipidomics, high-resolution mass spectrometry (HRMS) currently represents the most widely used technique. Advances in HRMS instrumentation now provide unprecedented sensitivity, mass accuracy, and resolving power.[2] In addition, separation techniques such as liquid chromatography (LC) can be conveniently coupled with MS analyzers to achieve prior sample separation, which reduces ion suppression phenomena and increases the signal-to-noise ratio for low-abundance analytes. LC-HRMS analyses are especially useful in untargeted approaches based on global lipid profiling, which requires the collection of hundreds (if not thousands) of features from complex mixtures such as liquid biofluids or tissue extracts.

In untargeted lipidomics, multivariate statistical analysis is widely used,[3] typically applied to "fat" matrices. The number of collected samples (objects) is often limited due to associated costs or the rarity of the material, while recent technologies have increased the number of detected features (variables). However, multivariate statistical methods are often applied without fully considering the nature and quality of the original data matrix. Additionally, the use of statistical models for prediction is often hindered by the large number of steps that a software tool must apply to obtain reliable results. Finally, while multivariate statistical analysis plays a pivotal role in global lipid profiling approaches due to its effectiveness in data reduction, interpretation, and classification, the application of lipidomics to biomarker discovery (trend analysis) benefits from cluster analysis.

In this presentation, we will report examples of the application of multivariate statistical analysis in LC-HRMS-based lipidomics, highlighting cautions that should be more widely addressed in the scientific community. Finally, we will describe tools we developed in the Lipostar software [4] to facilitate the use of chemometrics in LC-HRMS-based lipidomics, both for global lipid profiling and biomarker discovery.

[1] Wenk M.R. Lipidomics: New Tools and Applications, Cell, **2010**, *143*, 888 – 895.

[2] Tito Damiani T., Stefano Bonciarelli S., Thallinger G.G., Koehler N., Krettler C.A., Salihoğlu A.K., Korf A., Pauling J.K., Pluskal T., Ni Z., and Goracci L. Software and Computational Tools for LC-MS-Based Epilipidomics: Challenges and Solutions Anal. Chem. **2023**, *95,* 287-303.

[3] Griffin J.L., Liggi S., and Hall Z., in Lipidomics: Current and Emerging Techniques, ed. W. Griffiths and Y. Wang, The Royal Society of Chemistry, **2020**, pp. 25-48.

[4] Goracci L., Tortorella S., Tiberi P., Pellegrino R.M., Di Veroli A., Valeri A., and Cruciani G. Anal. Chem. **2017**, *89,* 6257-6264.

Statistical validation of multivariate treatment effects in longitudinal study designs

Guro F. Giskeødegård [1]Torfinn Støve Madssen [1], Age Smilde [2], Jose Camacho [3], Anders Hagen Jarmund [1], Johan Westerhuis [2]

[1] Norwegian University of Science and Technology, Trondheim, Norway
[2] University of Amsterdam, Amsterdam, Netherlands
[3] University of Granada, , Spain

Introduction

Multivariate extensions of repeated measures linear mixed models, such as repeated measures ANOVA simultaneous component analysis (RM-ASCA+) and linear mixed model-principal component analysis (LiMM-PCA), can be used for analyzing longitudinal studies with multivariate outcomes. However, there are no gold standards to assess the statistical validation of the observed effects of such models.

Purpose

The aim of our study was to compare different strategies for statistical validation of multivariate treatment effects in longitudinal study designs.

Methods

Using real and simulated data, we here perform an empirical comparison of different strategies for assessing statistical significance in these frameworks: permutation tests, the global log-likelihood ratio (GLLR) test, and nonparametric bootstrap confidence intervals for the estimated multivariate effects. Power curves were used to examine the statistical power of the different tests in detecting time-treatment interactions with varying effect sizes.

Results

Our results show that both the permutation tests and the GLLR-test can be used to statistically test the presence of a time-treatment interaction effect for multivariate data, however the GLLR approach will be sensitive to the number of included principal components in LiMM-PCA. The bootstrap confidence interval approach generally shows good statistical power, but has inflated type 1 error rates under certain conditions. This makes it unsuitable for the purpose of hypothesis testing in its present implementation, although it may still be useful for exploratory purposes.

Conclusion

Overall, our results show that the power of the tests for assessing multivariate effects in longitudinal studies is dependent on characteristics of the dataset, and it is important to be aware of the strengths and weaknesses of the different validation procedures.

Parallel targeted and untargeted metabolite analysis of mouse plasma samples using a benchtop multi-reflecting time of flight mass spectrometer

Ross Chawner [1]J. Kirk [1], A. King [1], L. Gethings [1], I.D. Wilson [2]

[1] Waters Corporation, Wilmslow , United Kingdom
[2] Liverpool University, Liverpool, United Kingdom

Metabolomic profiling of biological matrices can be targeted (quantifying known metabolites) or untargeted (identifying unknown molecules). Traditionally, these analyses are performed separately on different MS platforms due to their specific performance attributes. However, advancements in instrument sensitivity and scanning capabilities now allow both approaches on a single platform, reducing resource and sample volume requirements.

The Xevo MRT's fast-scanning capabilities enable the targeting of 65 amino acids and internal standards while generating high-quality untargeted data for biomarker discovery. Of these, 20 amino acids were fully quantifiable against stable isotopically labelled standards (SILS), and the remaining 40 were quantified against the SILS based on retention time. Amino acids were quantified over a 5–1000 µM range, with calibration curve linearity ($R^2 > 0.99$) and 67% of QC standards within 15% of the nominal concentration.

Simultaneously, untargeted data were processed using MZmine software for peak picking and normalization. Statistical analysis, including PCA and OPLS-DA, highlighted significant features between sample groups. Combining both acquisition approaches in a single analysis showed minimal compromise on data quality, maintaining high-resolution accurate mass untargeted analysis data.

Coming full circle: an omics practitioner's perspective on improving mass spectrometry for life sciences research

Matthew Lewis [1]

[1] Bruker Daltonics GmbH & Co, Bremen, Germany

Modern bioanalytical instruments are built by teams of scientists that include practitioners of the applications for which the instruments are intended, ensuring fitness for purpose. Beyond offering a pragmatic voice in a design process otherwise dominated by theory in physics and engineering, are there specific lessons learned from years of bioanalytical method development, data analysis, and omics application in the study of health and disease that add value to the final product and the experience of its users? In other words, are the skills learned in omics sciences translatable? Here, the influence of common "omics" problem solving and data analysis approaches on bioanalytical instrument design and operation are explored.

A Chemometric Framework for Population Metabolic Profiling: Application to a multi-country molecular epidemiology study

Reika Masuda [1]Jeremy Nicholson [1, 2], Elaine Holmes [1, 3], Julien Wist [1], Paul Elliott [4]

[1] Australian National Phenome Centre, Murdoch, Australia
[2] Institute of Global Health and Innovation, Imperial College London, London, United Kingdom
[3] Faculty of Medicine, Imperial College London, London, United Kingdom
[4] School of Public Health, Faculty of Medicine, Imperial College London, London, United Kingdom

Introduction

Population phenotyping generates large, high-dimensional datasets influenced by genetic, environmental, and analytical factors. New data analysis frameworks are needed to optimize the extraction of biologically relevant information and to assess specific sources of variation within these datasets. The INTERSALT study, which collected 10,079 24-hour urine samples from 52 centres across 32 countries, exemplifies the complexity and variability inherent in such data.

Purpose

A novel data analysis framework for metabolomic studies was developed, integrating hierarchical clustering, multivariate classification, and metabolite annotation, to streamline the analysis of large-scale nuclear magnetic resonance (NMR) datasets. Applying this framework to a global dataset of urinary NMR spectra, we aimed to uncover population-specific metabolic differences and assess the influence of demographic and lifestyle factors on metabolic profiles.

Methods

Urine samples were analyzed using 600 MHz ^{1}H NMR spectroscopy with both standard 1D and 2D experiments. Data preprocessing included baseline correction and normalization to ensure spectral consistency. To manage dataset complexity, median 1D spectra were grouped by centre and sex, then analyzed using principal component analysis (PCA) to capture up to 99% of the cumulative variance. Distance matrices were computed from PCA scores, followed by hierarchical cluster analysis (HCA) to identify metabolic similarities at the population level. Orthogonal partial least squares discriminant analysis (OPLS-DA) was employed to pinpoint metabolic distinctions between populations and sex differences within populations. Finally, statistical total correlation spectroscopy (STOCSY) facilitated the annotation of key metabolites.

Results

The necessity of an efficient and reliable analysis framework became evident given the initial complexity of the raw dataset. By reducing data to median 1D spectra, we effectively captured the overarching metabolic variation while mitigating noise, chemical shift drift and outlier effects, thereby reducing computational demands.

HCA of the distance matrices derived from the PCA scores revealed distinct clustering patterns, clearly differentiating metabolic profiles across populations. In contrast, direct analysis of the full spectra resulted in more complex and less interpretable patterns due to noise and outlier interference.

To validate these findings, OPLS-DA was applied to a subset of spectra, confirming significant separation between populations and reinforcing the robustness of the framework. Using OPLS-DA loadings, STOCSY facilitated the identification of correlated spectral features, aiding in the annotation of key metabolites responsible for metabolic differences. Further biological interpretation was guided by known metabolic pathways, highlighting the potential influence of long-term dietary and environmental factors on population-level metabolic variation.

Conclusions

This chemometric framework provides a powerful and efficient approach for analyzing large-scale metabolomic datasets while minimizing noise and outlier effects. It offers a scalable solution for biomarker discovery and enhances our understanding of metabolic variation across diverse populations, influenced by environmental and dietary factors.

From misinference to chemical insight: immune cells and cardiovascular disease

Gerjen Tinnevelt [1]Teun Schuncken [1], Coen Stehouwer [2], Casper Schalkwijk [3], Jeroen Jansen [1], Kristiaan Wouters [3]

[1] Radboud University, Nijmegen, Netherlands
[2] Maastricht UMC+, Maastricht, Netherlands
[3] Maastricht University, Maastricht, Netherlands

Introduction: Low-grade inflammation and endothelial dysfunction (ED) are important for the development of cardiovascular diseases (CVD), which are the leading cause of death globally. Understanding the underlying mechanisms linking immune cell activity to ED could provide valuable insights into CVD pathogenesis and potential therapeutic targets.

Purpose: This study aimed to investigate the relationship between immune cell signatures and endothelial dysfunction in a large population-based cohort, the Maastricht Study, to enhance our understanding of the contribution of immune cells to cardiovascular diseases.

Methods: We measured circulating immune cells in 798 participants using flow cytometry, analyzing 18 proteins to identify immunological signatures associated with ED. Endothelial dysfunction was quantified using a combined Z-score of sVCAM1, vWF, E-selectin, and sICAM1. Given that many known CVD risk factors also influence the immune system, we developed a novel approach to correct multivariate flow cytometry analysis for potential confounding factors. This approach involved four steps:

1) Deflating the impact of confounders on ED using regression.

2) Converting flow cytometry data to cellular distributions via principal component analysis, followed by deflation of these distributions using sequential orthogonal partial least squares regression.

3) Predicting deflated ED from the deflated cellular distributions resulting in immune cell signatures

4) Immune cell signatures, along with potential confounders, were then used to predict ED in a separate test set.

Confounders included age, sex, waist circumference, HbA1c, LDL/HDL ratio, statin use, smoking status, systolic blood pressure, presence of CVD, and education level.

Results: Initial crude analysis showed no significant association between B-cell, granulocytes and NK cell signatures with ED. Correction for age and sex resulted in a significant associated between T-cell and monocytes signatures with ED. Further adjustment for all confounders revealed that only T-cell signatures maintained a significant relationship with ED.

Conclusion: Our study highlights the significant association between T-cell signatures and endothelial dysfunction (ED) after adjusting for multiple confounders. This finding underscores the potential role of specific immune cell signatures in the pathogenesis of cardiovascular diseases (CVD). The novel approach we developed for correcting multivariate flow cytometry analysis for confounding factors proved effective in isolating the independent relationship

between immune cell signatures and ED. These results warrant further investigation into the precise phenotype and functional role of T cells in ED, which could lead to new insights and therapeutic targets for preventing and managing CVD.

Moreover, this study introduces a novel perspective on data analysis that diverges from traditional data-driven chemometrics. Instead of focusing solely on predictive accuracy, our approach emphasizes disproving associations between variables. This method aligns more closely with academic critical thinking and inferential reasoning, providing a robust framework for understanding complex biological interactions. For example, in this work, monocytes are covarying with obesity. Although monocytes are important for developing CVD, focusing on lifestyle changes would be more beneficial than targeting monocyte activation therapeutically. In contrast, T cells show an independent mechanism in developing CVD, even in the absence of obesity. Thus, targeting T cells therapeutically might be beneficial for individuals suffering from abnormal T cell signature.

Development of a high-throughput LC-MS/MS assay for inborn errors of metabolism

Ingvi Karl Jonsson [1]Freyr Johannsson [2], Leifur Franzson [2], Margret Thorsteinsdottir [3], Jon Johannes Jonsson [2], Ottar Rolfsson [1]

[1] University of Iceland, Faculty of Medicine, Center for Systems Biology, Reykjavik, Iceland
[2] Landspitali University Hospital, Department of Genetics and Molecular Medicine, Reykjavik, Iceland
[3] University of Iceland, Faculty of Medicine, Department of Pharmaceutical Science, Reykjavik, Iceland

Introduction

Inborn errors of metabolism are a group of rare inherited disorders which can have early onset during infancy. If not diagnosed and treated, these disorders can lead to brain damage, cognitive impairment and death. The pathophysiological reason for these disorders is typically a monoenzymatic deficiency, resulting in toxic substrate accumulation. Many of these disorders have been classified as amino acidemias, organic acidemias, beta oxidation defects and more. To catch patients before disease onset, newborn screening has become a universal public health program – screening amino acids and acyl carnitines with high-throughput, infusion-based tandem mass spectrometry. Even though not detected, organic acids and N-acyl glycines can provide important information. Additional biomarkers quantified with second-tier assays such as with liquid chromatography-tandem mass spectrometry (LC-MS/MS) may be used for differential diagnosis and false positives reduction. However, these biomarkers have differing physicochemical properties, making the development of a universal LC-MS/MS assay difficult.

Purpose

Amino acids and organic acids are polar, whereas long-chain acyl carnitines and glycines are not – resulting in different LC methods needed to separate them. Our goal is to develop a single high-throughput LC-MS/MS assay to quantify these biomarkers. This will be done using 3-nitrophenylhydrazine (3-NPH) derivatization. The derivatization alters the chemical structure of analytes, reducing their polarity and making them amenable to reverse-phase liquid chromatography (RPLC).

Methods

LC-MS/MS method development is carried out using Waters Acquity UPLC system coupled to a Waters Xevo TQ-XS tandem mass spectrometer. Briefly, method development consists of MS/MS optimization, column screening and mobile phase evaluation. Additional parameters to optimize include column temperature, flow rate and mobile phase additives. Currently, 61 analytes are being evaluated for a 5-minute method after derivatization with 3-NPH. The final panel of analytes is yet to be determined. Currently, pure chemical standards are used for method development. However, our next steps include the transfer to dried blood spots. Once developed, the assay will be validated using blood spots of multiple inborn errors of metabolism.

Results

A common problem with derivatization is unspecific fragmentation: neutral loss of the original analyte and detection of only the derivatizing agent. Here, unspecific fragmentation is observed in negative electrospray ionization mode, but specific in positive mode – ideal for MS/MS quantitation. Preliminary LC development indicates preference to low pH

environment on commonly used RPLC columns. Furthermore, acyl carnitines separate effectively across the elution gradient in order of increasing chain length – indicating suitability of the method to accurately identify and quantify multiple biomarkers of beta oxidation defects.

Conclusions

While in early stages of development, preliminary results are promising. The assay may have an application to newborn screening as no second-tier LC-MS/MS screen is currently available in the national newborn screening program. Such assays benefit the healthcare system, the patient and the patients family. Second-tier LC-MS/MS assays use the initial neonatal dried blood spots, and can thus be used to reveal false positives and provide further biomarker selectivity before further intervention such as invasive and expensive neonatal testing.

Integrative Single- and Multi-Omics Data Analysis using iSODA

Yassene Mohammed [1]

[1] Leiden University Medical Center, , Netherlands

The rapid advances in high-throughput omics resulted in an increase in data generation at lower costs. These data help understanding biological processes, elucidating disease pathways, and advancing personalized medicine. Researchers are, however, confronted with the lack of a versatile software solution to harmoniously analyze single-omics and interpret multi-omics data. We have developed iSODA, a web-based application for the analysis of single- and multi-omics data. The tool emphasizes intuitive interactive visualizations designed for user-driven data exploration. Researchers can access a variety of functions ranging from simple visualization like volcano plots and PCA to advanced functional analyses like enrichment analysis and lipid saturation analysis. For integrated multiomics, iSODA incorporates multi-omics factor analysis and similarity network fusion. The ability to adapt the data on-the-fly allows for tasks, such as the removal of outlier samples or failed features, imputation, or normalization. All results are presented through interactive plots, the modular design supports extensions, and tooltips and tutorials provide comprehensive guidance for users. iSODA is accessible under http://isoda.online/.

Mathematical super-resolution in chromatography: enhancing chemical volatile profiling with accelerated GC-MS and tensor decomposition (PARAFAC2)

Beatriz Quintanilla-Casas [1], Asta Nathalie Haagensen [1], Mikael Agerlin Petersen [1], Rasmus Bro [1]

[1] University of Copenhagen, Copenhagen, Denmark

Gas Chromatography-Mass Spectrometry (GC-MS) is a cornerstone technique for volatile chemical profiling in complex mixtures [1]. However, traditional methods rely on high chromatographic resolution, often at the expense of analysis time.

In this work, we speed up the analysis time dramatically and challenge the classical notion of chromatographic resolution – typically defined by theoretical plates [2] – by integrating mathematical modelling, taking advantage of the mass spectral dimension. Specifically, we apply tensor methods, namely PARArallel FACtor analysis 2 (PARAFAC2), to extract meaningful chemical information from GC-MS data despite the reduced chromatographic separation. This method has been proven successful for efficiently extracting more chemical information in a more robust way from raw GC-MS data in an automated way, reducing thus dependence on the user [3]. Red wines, as a case study of a complex flavour mixture, were measured by a standard (40-minute) and a fast (20-minute) GC-MS method, where only the temperature ramp was altered.

Our findings suggest that optimized faster GC-MS can significantly shorten analysis time without sacrificing the quality and robustness of the extracted chemical information using tensor models. This would be a game changer, allowing transforming conventional GC-MS into a high-throughput analytical method for complex matrix characterization, without the need to switch to a fast GC instrumental setup.

ROI-PARAFAC2 - Deconvoluting only the important LC-HRMS signals

Julius Jessen Terp [1]Rasmus Bro [1]

[1] University of Copenhagen, Department of Food Science, Copenhagen, Denmark

Liquid Chromatography-High Resolution Mass Spectrometry (LC-HRMS) is an essential analytical technique in many fields, ranging from untargeted environmental analysis over metabolomics to food quality [2]. Advancements in instrumentation have resulted in higher resolution, leading to a substantial increase in data size. Additionally, the complexity of both samples analyzed, and research questions asked has escalated. As a result, the challenge of analyzing said data has grown correspondingly [6]. Current data analysis workflows rely on expert tuning of a number of parameters, each of which control individual functions, resulting in a ”black box” in terms of evaluating the validity of the results [1].

Region Of Interest (ROI) search, based on values obtained from raw data such as average peak width, noise threshold and instrument accuracy, is a way of binning the data. As the ROI-search is based on the experimental setup, this reduction in data size will not affect the mass accuracy, as the data will be binned into mass spectral ROIs [4]. Previous ROI searches have been performed on the full dataset, but in this work, the searches will be performed on selected retention time intervals before modelling. Since the masses present in each interval will vary, the number of masses included in each interval will be reduced, which in turn will speed up modelling.

The retention-time intervals are then deconvoluted using PARAllel FACtor analysis 2 (PARAFAC2), a model based on the intrinsic, (mostly) tri-linear properties of chromatographic data, but can handle small shifts and/or shape changes in the retention time dimension that arise from instrument and experiment variation [3]. This can allow for better post-analysis separation of peaks as well as reduce the number of individual and independent parameters needed to analyze the data, resulting in increased interpretability.

To evaluate the performance of the proposed workflow, data from [5] will be analyzed follow the

new approach.

References

[1] M. Aigensberger, C. Bueschl, E. Castillo-Lopez, S. Ricci, R. Rivera-Chacon, Q. Zebeli, F.

Berthiller, and H. E. Schwartz-Zimmermann. “Modular Comparison of Untargeted Metabolomics

Processing Steps”. In: Analytica Chimica Acta 1336 (Jan. 22, 2025), p. 343491.

[2] C. Aydogan. “Recent Advances and Applications in LC-HRMS for Food and Plant Natural

Products: A Critical Review”. In: Analytical and bioanalytical chemistry 412.9 (2020), pp. 1973–1991.

[3] R. Bro, C. A. Andersson, and H. A. L. Kiers. “PARAFAC2—Part II. Modeling Chromato-graphic Data with Retention Time Shifts”. In: Journal of Chemometrics 13.3–4 (1999), pp. 295–309.

[4] E. Gorrochategui, J. Jaumot, and R. Tauler. “ROIMCR: A Powerful Analysis Strategy for

LC-MS Metabolomic Datasets”. In: BMC Bioinformatics 20.1 (May 17, 2019), p. 256.

[5] G. Gürdeniz, M. G. Jensen, S. Meier, L. Bech, E. Lund, and L. O. Dragsted. “Detecting Beer

Intake by Unique Metabolite Patterns”. In: Journal of Proteome Research 15.12 (Dec. 2, 2016),pp. 4544–4556.

[6] G. Renner and M. Reuschenbach. “Critical Review on Data Processing Algorithms in Non-

Target Screening: Challenges and Opportunities to Improve Result Comparability”. In: Ana-lytical and bioanalytical chemistry 415.18 (May 15, 2023), pp. 4111–4123.

Integration of GC–MS and PTR–MS Data: A Holistic Approach to Comprehensive Volatile Analysis in Cheese and Plant-Based Alternatives

Berta Torres Cobos [1]Sylvester Holt [1], Mette Skau Mikkelsen [2], Åsmund Rinnan [1]

[1] Department of Food Science, University of Copenhagen, Rolighedsvej 26, DK-1958 , Copenhagen, Denmark
[2] FOSS Analytical A/S, Nils Foss Allé 1, Hillerød, 3400, Denmark, Hillerød, Denmark

Understanding the volatile profile of cheese and its plant-based alternatives is crucial for sensory optimization and new product development, as consumer demand for sustainable food options continues to grow. The transition toward plant-based diets is driven by increasing awareness of environmental sustainability, health considerations, and evolving regulatory policies aimed at reducing greenhouse gas emissions from food production. While plant-based cheese alternatives (PBCA) are rapidly entering the market, they often struggle to replicate the complex flavor profiles of traditional dairy cheese, impacting consumer acceptance. A detailed analysis of volatile organic compounds (VOCs) in these products is essential not only for improving sensory attributes but also for supporting manufacturers in developing appealing and sustainable alternatives that align with consumer expectations and regulatory frameworks.

Gas chromatography–mass spectrometry (GC–MS) is the reference technique for analyzing VOCs in food. However, its need for extensive sample preparation, including extraction and preconcentration, limits its feasibility for high-throughput analysis. In contrast, proton transfer reaction time-of-flight mass spectrometry (PTR-ToF-MS) enables direct, real-time VOC analysis with minimal sample preparation, aligning with green analytical chemistry principles. While PTR-ToF-MS offers high sensitivity and rapid screening, it faces challenges in compound identification within complex food matrices due to isomeric interferences. Integrating GC–MS with PTR-ToF-MS leverages their complementary strengths, enhancing VOC identification and addressing individual limitations.

This study aims to develop an integrated analytical approach by combining GC–MS and PTR–MS data through multivariate analysis techniques to achieve a more comprehensive volatile profile of cheese and PBCA. A set of 70 cheeses and PBCA, covering major cheese types from leading European producers, was analyzed using untargeted profiling via GC-MS and PTR-ToF-MS fingerprinting. GC-MS data was processed using PARADISe, a deconvolution and identification tool based on PARAllel FACtor analysis 2 (PARAFAC2), which estimates compound concentrations, elution profiles, and pure mass spectra. PTR-ToF-MS data, providing high-resolution fingerprints of volatile compounds, was explored through multivariate analysis to identify patterns and sample clustering based on volatile composition.

This integrated approach establishes a robust framework for volatile analysis in cheese and PBCA. By refining compound identification, leveraging PTR-ToF-MS fingerprinting for rapid screening, and streamlining analytical workflows, it enhances flavor characterization, supporting quality control and product innovation in the evolving food market.

Session 6 - Multivariate curve resolution and calibration

18-06-2025 - **15:20 - 15:40**

Can canonical angle measures be useful in MCR analysis?

Klaus Neymeyr [1]

[1] University of Rostock, Rostock, Germany

Multivariate Curve Resolution (MCR) methods use vector and matrix norms or inner products to measure the similarity or distance of, e.g., spectra or concentration profiles. For example, the cosine of the acute angle between two vectors is given by the absolute Euclidean inner product and can be interpreted as a correlation. Angle values are the basis for the 2D correlation analysis [1] and for more general concepts such as generalized canonical correlation analysis [2]. Acute angles or correlations between vectors can be generalized to pairs of subspaces.

C*anonical angles*, also called principal angles, measure the mutual orientation of a pair of subspaces [3]. The angular distance of two s-dimensional subspaces is determined by s canonical angles. These angles have a rich structure and allow a precise measurement of subspace distances. The *largest canonical angle* is called the *subspace angle* and has found application in chemometrics [4].

This talk introduces the geometry and analysis of canonical angles and shows how they can be used in MCR analysis. Canonical angle analysis can help to detect changes in chemical composition during a reaction and can be related to the chemical conversion. Similarities and differences with the evolving factor analysis are pointed out. Canonical angles are studied for model data and experimental spectroelectrochemical data.

References:

[1] Noda I.; Two-dimensional infrared spectroscopy. J. Am. Chem. Soc **1989**, 111(21):8116–8118.

[2] Smilde A.K., Næs T., Liland K.H.; Multiblock data fusion in statistics and machine learning: Applications in the natural and life sciences. Wiley, **2022**.

[3] Björck A., Golub G.H.; Numerical methods for computing angles between linear subspaces. Math. Comp **1973**, 27(123):579–594.

[4] Liu Y.J., Tran T., Postma G., Buydens L.M.C., Jansen J.; Estimating the number of components and detecting outliers using Angle Distribution of Loading Subspaces (ADLS) in PCA analysis. Anal. Chim. Acta **2018**, 1020:17–29.

Enhancing Pharmaceutical Manufacturing: pH Monitoring Using Raman Spectroscopy and Multivariate Curve Resolution Methods

Sara Piqueras Solsona [1]Milana Sostar [1], Iben Sacher Salinas [1], Isidro Badillo Ramirez [2], Anja Boisen [2]

[1] Novo Nordisk, Bagsværd, Denmark
[2] Technical University of Denmark, Kogens Lyngby, Denmark

Introduction:

pH control is critical in pharmaceutical manufacturing, directly impacting drug stability, solubility, and potency. While pH measurement is conceptually simple, its precise control during manufacturing processes remains challenging [1]. Good Manufacturing Practice (GMP) regulations demand high-quality standards, necessitating accurate pH monitoring. Raman spectroscopy, a rapid and non-destructive analytical technique, has emerged as a promising Process Analytical Technology (PAT) tool in pharmaceutical production. Despite its widespread use in process monitoring and quality control, Raman spectroscopy potential for pH monitoring is yet to be fully explored. The

overlapping Raman spectra of components in acid-base reactions present unique analytical challenges, necessitating advanced chemometric tools for data interpretation.

Purpose:

This study aims to develop a novel, accurate, pH measurement method for pharmaceutical manufacturing environments by integrating Raman spectroscopy with multivariate curve resolution (MCR) techniques [2].

Methods:

Raman spectra were collected from aqueous phosphoric acid solutions across a range of pH values to develop a predictive model. Titrations were performed using a METTLER TOLEDO Optimax™ 1001 equipped with a 1000 ml glass reactor, with continuous monitoring of temperature and pH. Raman spectra were acquired in triplicate using a Horiba system with a 532 nm, 300 mW visible diode laser. Data preprocessing was conducted using MATLAB 2024a, while modeling employed the MCR-ALS GUI 2.0 software [3].

Results and Discussion:

The study demonstrates the efficacy of chemometric analysis of Raman spectroscopy for pH monitoring. MCR-ALS proved to be a flexible approach, enabling the resolution of different chemical species involved in phosphoric acid titration and providing insights into the evolution of chemical components. The method's ability to incorporate external information and constraints guided the iterative process, reducing ambiguities in the bilinear curve decomposition. This approach led to chemically interpretable solutions, overcoming the limitations of traditional pH probes, such as instability and frequent calibration requirements.

Conclusions:

This work presents a novel method for pH monitoring in pharmaceutical manufacturing using Raman spectroscopy coupled with MCR-ALS. The developed models show promise for predicting pH in pharmaceutical drug products with improved accuracy and stability compared to conventional pH probes. This approach has the potential to enhance process control and quality assurance in pharmaceutical manufacturing, aligning with PAT initiatives and GMP standards.

References:

[1] Cheng, K. L.; Zhu, D. M. Sensors **2005**, 5 (4), 209−219.

[2] de Juan, A; Jaumot , J.; Tauler, R. Anal. Methods **2014**, 6 (14),4964−4976.

[3] Jaumot, J.; de Juan, A, Tauler, R, Chemometrics and Intelligent Laboratory Systems **2015**,140 ,1-2.

Comparing methods for calibration transfer

Lars Erik Solberg [1]Jens Petter Wold [1], Nils Kristian Afseth [1], Ingrid Måge [1], TIril Aurora Lintvedt [1], Katinka Dankel [1], Karen Wahlstrøm Sanden [1], Erik Tengstrand [1]

[1] Nofima AS, Ås, Norway

Introduction

The calibration transfer problem has been addressed in the scientific literature at least since the mid 1980s, yet methods are still being proposed on a regular basis, perhaps based on the claims that the problem has not yet been solved [1]. While originally formulated as a transfer between two devices, the term “calibration transfer” has been used to address changes also in the environment or in the sample composition [2]. We believe that making a distinction between these sub-problems is important as it narrows down the scope of the problem. For instance, the difference between instruments concerns properties specific to these such as a shift in wavelengths, differences in light source, wavelength resolution, etc. This problem strictly speaking does not include temperature or humidity in the environment, or changes to the sample matrix.

Many methods (see [1,3,4] for literature reviews) have been developed and the most recent of these reviews included more than 50. When presented, the methods propose a comparison with a small set of competing methods while

using a couple of datasets, often with an unclear transfer problem. Few wider comparisons between methods have been presented in the literature, (see [5] for one example).

In this work, we present such a comparison between transfer methods for the original (instrument-to-instrument) calibration transfer problem.

Purpose

Compare calibration transfer methods by using perturbations of real datasets to emulate differences between instruments. We will also apply methods to real transfer problems within this set. The comparisons will also attempt to address the question of “difficult” versus “hard” transfer problems.

Methods

A collection of existing spectroscopic datasets was subject to perturbations, each representing one instrumental difference. This creates a design-of-experiments where each combination of perturbations represents a calibration transfer problem with known “problem”.

A selection of methods was applied to these problems and compared to assess the relative difficulty of these perturbations as well as on the relative performance of methods.

Results

We are working on an article and plan to present results which will suggest a ranking of methods, perturbations and datasets.

Conclusions

We will show that there are categories of methods which, among those included, work better for the original calibration transfer problem on our selection of datasets.

References

[1] Workman, J. J. (2018). A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy. Applied Spectroscopy, 72(3), 340–365.

[2] Poerio, D. V., & Brown, S. D. (2018). Dual-Domain Calibration Transfer Using Orthogonal Projection. Applied Spectroscopy, 72(3), 378–391.

[3] Fearn, T. (2001). Standardisation and calibration transfer for near infrared instruments: A review. Journal of Near Infrared Spectroscopy, 9(4), 229–244.

[4] Mishra, P. et al. (2021). Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always. TrAC - Trends in Analytical Chemistry, 143, 116331.

[5] Igne, B. et al. (2009). Improving the transfer of near infrared prediction models by orthogonal methods. Chemometrics and Intelligent Laboratory Systems, 99(1), 57–65.

Session 7 - Miscellaneous topics

18-06-2025 - 16:15 - 17:20

MIR and NIR Hyperspectral imaging for food quality control and traceability

Federico Marini [2]Consuelo Giustizieri [1], Camilla D'Orazio [2], Federico Fancoli [2], Massimo Reverberi [1]

[1] Department of Environmental Biology, University of Rome La Sapienza, Rome, Italy
[2] Department of Chemistry, University of Rome La Sapienza, Rome, Italy

Food quality control covers various aspects, including the detection of adulteration, contamination, and fraud, as well as verifying safety, authenticity, and compliance with labeling regulations. This also includes the ability to trace the origin of food products, whether botanical, animal-based, or geographical. In this context, hyperspectral imaging—particularly in the infrared spectrum—emerges as a valuable tool for supporting producers, consumers, and regulatory bodies. This is due to its capability to combine a highly detailed spectroscopic signature at each pixel with the ability to map the spatial distribution of specific molecules or groups of substances.

Food products, however, are complex in nature, and their composition can be affected by multiple sources of variability. As a result, their characterization and authentication can benefit from integrating complementary data from various spectroscopic techniques. Considering these factors, this discussion presents the application of a combined MIR (mid-infrared) and NIR (near-infrared) hyperspectral imaging system for food quality control. The focus will be on both the distinct information provided by each platform and how data fusion strategies can effectively integrate near and mid-infrared data to enhance model performance.

Practical examples will include the characterization of rice, pasta and lentil samples, and, on a broader scale, findings from the National Agritech project. In particular, within the context of the Agritech project, the main results related to the successful application of MIR and NIR hyperspectral imaging to track the entire supply chain of specific food products, especially in the case of wheat samples, will be disseminated.

Acknowledgements:

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022).

Explorative analysis of hyperspectral images by Block Term Decomposition

Marina Cocchi [1], Alessandra Olarini [1, 2]Cyril Ruckebusch [2], Ludovic Duponchel [2]

[1] Dept. of Chemical and Geological Sciences, Univ. of Modena and Reggio Emilia , Modena, Italy
[2] University of Lille, LASIRE, Lille, France

Tensor decomposition methods have long been recognized as robust tools in chemometrics, providing effective approaches for analysing complex data structures. In hyperspectral imaging (HSI), where the simultaneous exploration of spatial and spectral domains is essential, tensor decomposition techniques are appealing since keep the 3D structure of HSI and naturally provide combined spatial and spectral information. The most diffuse tensor decomposition methods are Canonical Polyadic Decomposition (CPD) and Tucker [1]. However, CPD's strict rank-1 constraint on each factor can limit its applicability, while Tucker decomposition offers greater flexibility but is affected by rotational ambiguity. Block Term Decomposition (BTD) [2] provides a valuable middle ground. Specifically, the rank-(Lr,Lr,1)-BTD variant imposes a rank-1 constraint on one factor while allowing higher ranks for the other two factors (see Figure). This approach retains the uniqueness of CPD while better capturing complex spatial structures associated with a single spectral signature. This makes rank-(Lr,Lr,1)-BTD particularly suited for hyperspectral image analysis, nonetheless, it has been mainly explored in remote sensing context [3-4] and much less in HSI applications relevant to chemometrics community.

We are convinced that it could be suitable in addressing scenarios with minor components, components with similar spatial distributions but distinct spectral signatures, and vice versa. In this work, we focus on determining the subfactors Lr, a critical step for the effective application of this method. We investigate the use of rank-(Lr,Lr,1)-BTD for analysing benchmark hyperspectral imaging datasets, including chemical mixtures, biological fluids, and remote sensing images. These datasets span multiple spectroscopic techniques, such as UV-Vis, NIR, Raman. The results highlight the potential for rank-(Lr,Lr,1)-BTD method in hyperspectral image analysis, offering valuable insights into the efficacy of tensor-based decomposition methods for addressing the significant challenges posed by such data.

What to NOT do when calculating complex PARAFAC models

Åsmund Rinnan [1]Helene Froriep Stengade Halberg [1]

[1] University of Copenhagen, , Denmark

PARAllel FACtor analysis (PARAFAC) [1] is a common data analytical approach for handling three-way data. One type of data that behaves very well in this regard, is fluorescence spectroscopy, either as steady-state measurements through excitation-emission-matrices (EEMs) or time-resolved fluorescence through time-resolved emission spectroscopy (TRES'). In both cases, the data form three-way tensors, and behave trilinearly, i.e. as a sum of products of vectors in each direction in the tensor.

The combination of EEMs and PARAFAC has many different publications in the literature, with measurements ranging from rather pure systems with 3-4 fluorophores present, to complex samples with more than 10 fluorophores. However, very few of the published applications of PARAFAC on EEM extracts these many components. The reasoning behind this is, at least, two-fold: 1) the trust put in core-consistency [2] for model correctness, and 2) the common practice of using Direct TriLinear Decomposition (DTLD) or General Rank Annihilation Method (GRAM) for initialization.

This presentation will focus on the work by Halberg et al [3], but with some extensions and details with regards to both the core-consistency and the initialization problem.

References

1. Bro R (1997): PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems, 38: 149-171

2. Bro R, Kiers HAL (2003): A New Efficient Method for Determining the Number of Components in PARAFAC Models, Journal of Chemometrics, 17: 274–286.

3. Halberg HFF, Bevilacqua M, Rinnan Å (2024): Resampling as a robust measure of model complexity in PARAFAC models, Journal of Chemometrics: e3601

Why Gaussian Processes not used more often in Chemometrics?

Olivier Cloarec [1]

[1] Sartorius Corporate Research, , France

Gaussian Process (GP) is not widely used within the chemometrics community, although it offers a robust probabilistic framework for tackling the challenges of multivariate calibration and process monitoring. Their non-parametric nature allows for flexible modeling of complex, nonlinear relationships in high-dimensional spectroscopic and chemical data, providing both predictive accuracy and uncertainty quantification.

Within the chemometrics data domain, GP regression can be successfully applied to multivariate spectroscopic calibration, enabling the development of calibration models that handle both multivariate and functional covariates, and accommodate multidimensional responses. For example, in spectroscopic analysis, GP models have demonstrated superior performance in predicting chemical compositions (such as moisture, fat, and protein content in food samples) from high-dimensional spectral data, outperforming traditional linear methods, especially when response variables are correlated. Principal component analysis (PCA) is often integrated into GP frameworks to

decorrelate multivariate responses, simplifying the modeling of complex covariance structures. Beyond calibration, the chemometrics community leverages GPs for process monitoring, where their ability to model uncertainty supports the early detection of process deviations and faults.

Overall, Gaussian Process models should be a main instrument tin the chemometrics toolkit, driving advances in multivariate calibration, process monitoring but it is not. Why is it not widely used? This is a question I would like to ask the SSC community