Table of contents

From Vision to Emission: Bridging Optical and LIBS Worlds in Spectroscopic Imaging

Ruggero Guerrini[1], Federico Marini[2], Herreyre Nicolas[3,4], Vincent Motto-Ros[3], **Ludovic Duponchel**[1]

[1]Univ. Lille, CNRS, UMR 8516 – LASIRE – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, 59000, France

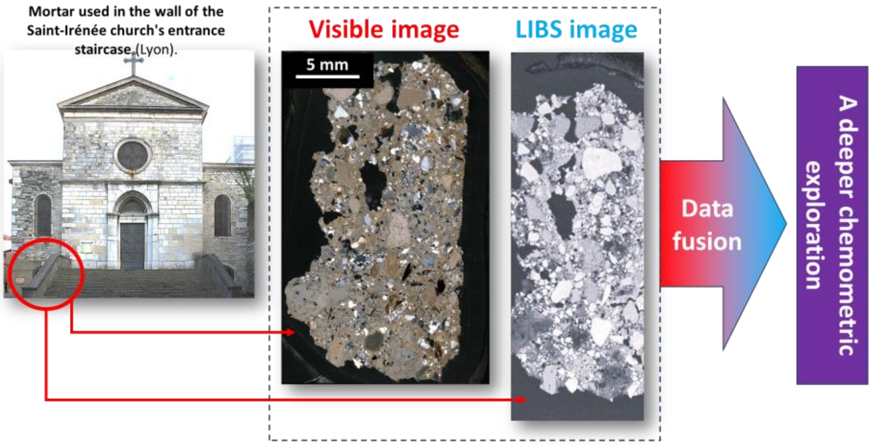
[2]Department of Chemistry, Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, 00185, Italy

[3]Institut Lumière Matière UMR 5306, Université Lyon 1. CNRS, Villeurbanne, 69622, France

[4]Archéologie et Archéométrie, UMR 5138, Univ. Lyon 2-CNRS-Univ. Lyon 1, Maison de l'Orient et de la Méditerranée, 7 rue Raulin, 69007 Lyon, France.

Keywords: LIBS hyperspectral imaging, visible image, data fusion, registration, color spaces.

Spectroscopic imaging is now more than ever at the heart of analytical chemistry. It enables us to thoroughly investigate complex samples by combining, in a single acquisition, molecular or elemental spectral analysis with spatial information. Over time, each spectroscopic technique has developed its own imaging modality, typically through the combination of a spectrometer and a microscope or another focusing system. Without being exhaustive, we can cite infrared, Raman, MALDI, and LIBS imaging, each offering unique characteristics in terms of spatial resolution, sensitivity, speed, and chemical insight. The sheer volume of spectra generated by these instruments quickly led us to develop chemometric approaches to explore and extract the most unbiased chemical information possible. In fact, it is now quite rare to see published work in spectroscopic imaging without an accompanying chemometric component. One might assume, then, that everything is progressing optimally. However, the results we will present in this work stem from an observation that applies broadly across all spectroscopic imaging techniques. Most of these instruments are also capable of acquiring a visible image of the sample under investigation. This is often used to observe the precise area to be analyzed or to ensure the optical system is correctly aligned for spectral acquisition. Yet despite the rich information contained in these visible images, they are rarely used in subsequent data processing, only the spectroscopic data are typically exploited. The aim of this presentation is to demonstrate how data fusion between a visible image and the spectroscopic data cube can significantly enhance sample exploration. We will introduce aspects related to image registration as well as colorimetric space transformations. To illustrate our approach, we will focus on the characterization of an ancient mortar using LIBS imaging.


Mortar used in the wall of the Saint-Irénée church's entrance staircase (Lyon).
Visible image
LIBS image
5 mm
Data fusion
A deeper chemometric exploration

Assessing of cooking quality level of boiled cassava by machine learning classification and NIR hyperspectral imaging

K. Meghar[1] Y. Janati Idrissi[2]

[1] UMR Qualisud, CIRAD, Montpellier, France, karima.meghar@cirad.fr

[2] Geology and Sustainable Mining Institute (GSMI), Université Mohammed VI Polytechnique (UM6P), Marocco.

Keywords: Cassava, Water absorption, Cooking quality, Machine learning classification and Hyperspectral imaging.

1 Introduction

Cassava roots provide a major calorie source for nearly 800 million people across the Americas, Asia, and Africa. They are eaten boiled, steamed, fried, or processed into foods like gari, fufu, eba, and attiéké. For boiled cassava, consumers prefer varieties with short cooking time. Water absorption and texture being key quality traits to evaluate cooking quality (1,2).

In this study, the cooking quality of cassava roots was assessed based on water absorption after 30 minutes of boiling (WAB30%). Until now, WAB30% has been evaluated using an ancestral and time-consuming method described by tran et al (1).

Hyperspectral imaging (HSI) is a fast and non-destructive technique that allows for the prediction and visualization of the spatial distribution of major components within a sample. Meghar et al (3) demonstrated the potential of HSI as a high-throughput tool for visualizing the heterogeneity in the distribution of dry matter content. However, their results showed weak performance in quantifying of WAB30%.

The aim of this study is to evaluate the potential of HSI combined with machine learning for classifying cassava genotypes into good (WAB30% ≥ 12%) and poor (WAB30% < 12%) cooking quality categories.

2 Material and methods

A total of 2175 images (samples) from 121 genotypes were analysed. After image correction and selection of regions of interest (ROIs), near-infrared average spectra were extracted and spectral pre-treatments was applied to minimize adverse effects. Various classification methods (PLS-DA, SVM, KNN, RF) were tested.

3 Results and discussion

Random Forest (RF) classification achieved the highest classification rates compared to the other methods with an accuracy of 87.54%, a specificity of 83.09%, and a sensitivity of 91.99%. The confusion matrices reveal strong overall classification performance for both training and test sets. In training, the model achieved high accuracy with 1328 TN and 1329 TP, with few FP and FN errors. Test-set results (280 TN, 310 TP, 57 FP, 27 FN) show slightly reduced but still solid performance. Overall, the model generalizes well, with expected minor decreases when applied to unseen data.

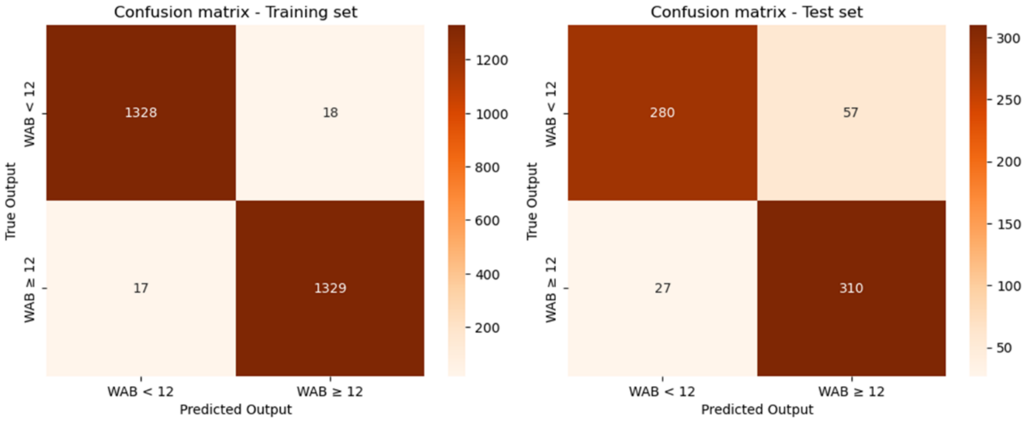


Figure 1**:** Confusion matrices for training and test sets using RF model.

The model was applied to classify/predict each pixel of fresh cassava root images to obtain information on the distribution of the WAB30 parameter within the samples at the pixel level. Two genotypes were selected to illustrate WAB30% distribution (Figure 2). Based on their WAB30% values, these genotypes represent good (CR63), and poor (FalsaReina) cooking quality.

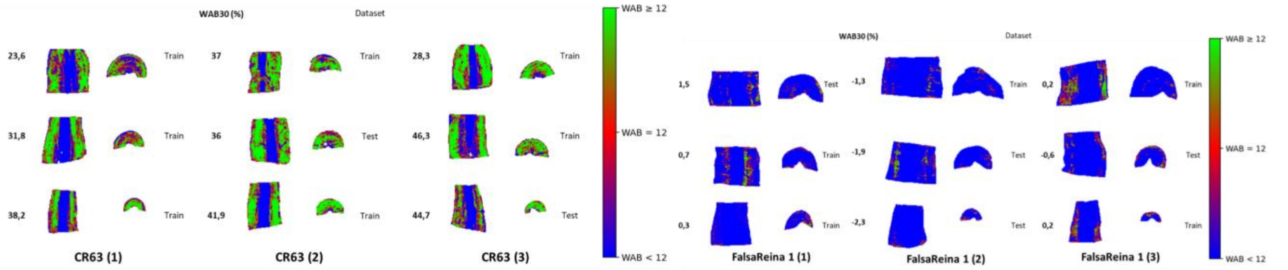


Figure 2: Classification maps of fresh cassava roots using the RF model: Visualization of WAB30 parameter distribution across: good (CR63), and poor (FalsaReina) cooking quality Genotypes.

4 Conclusion

This study shows that hyperspectral imaging can effectively classify cassava roots based on cooking quality, particularly their WAB30%. The best performance came from a Random Forest Classifier combined with SNV pretreatment. This model achieved 87.54% accuracy in distinguishing good vs. poor cooking quality. The method offers a fast, non-destructive alternative to traditional water absorption measurements. Future work should include larger, more diverse datasets and advanced analytical approaches to improve robustness for breeding and quality assessment.

5 References

1. Tran T, Zhang X, Ceballos H, Moreno JL, Luna J, Escobar A, et al. Correlation of cooking time with water absorption and changes in relative density during boiling of cassava roots. Int J Food Sci Technol. 2021;56(3):1193‑ 205.

2. Franck H, Christian M, Noël A, Brigitte P, Joseph HD, Cornet D, et al. Effects of cultivar and harvesting conditions (age, season) on the texture and taste of boiled cassava roots. Food Chem. 2011;1(126):127‑ 33.

3. Meghar K, Tran T, Delgado LF, Ospina MA, Moreno JL, Luna J, et al. Hyperspectral imaging for the determination of relevant cooking quality traits of boiled cassava. J Sci Food Agric. 26 mai 2023;jsfa.12654.

Bio-Collector from Moroccan Biomass for Flotation of Low-Grade Phosphate

Hasnaa Hilmi hasnaahilmi@gmail.com Cadi Ayyad University UCA, Faculty of Sciences Semlalia (FSSM), Applied Chemistry and Biomass Team, Department of Chemistry & Development, Marrakesh, Morocco

Abdelmoughit Abidi abidiabdelmoughit@gmail.com Cadi Ayyad University UCA, Faculty of Sciences Semlalia (FSSM), Applied Chemistry and Biomass Team, Marrakesh, Morocco

Abdelrani Yaacoubi ayaacoubi@uca.ac.ma Cadi Ayyad University UCA, Faculty of Sciences Semlalia (FSSM), Applied Chemistry and Biomass Team, Department of Chemistry & Development, Marrakesh, Morocco

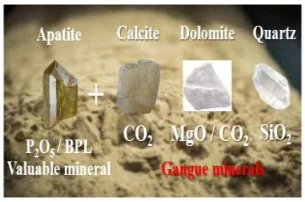
Khalid El Amari k.elamari@uca.ac.ma *Cadi Ayyad University, UCA, Faculty of Sciences and Technologies (FSTM), Laboratory of Georessources, Geoenvironment & Civil Engineering, Marrakesh, Morocco*

Abdelaziz Baçaoui bacaoui@uca.ac.ma Cadi Ayyad University UCA, Faculty of Sciences Semlalia (FSSM), Applied Chemistry and Biomass Team, Department of Chemistry & Development, Marrakesh, Morocco

Keywords: Low-grade phosphate ore, DOE, bio-based collector, Waste Biomass.

1 Introduction

Phosphate ore is a sedimentary rock rich in phosphate minerals and is an important raw material used in the chemical industry. Morocco holds nearly 70% of the world's phosphate reserves, making it a key resource for fertilizer production. Phosphate ore contains apatite as the valuable mineral, while calcite, dolomite, and quartz are the main gangue minerals.



It is classified into three categories:

- High-grade (HT): greater than 30% P_2O_5 ➔ **Quality requirement**
- Medium-grade (MT): (20-30%) P_2O_5 ➔ **Flotation process**
- Low-grade (BT): less than 20% P_2O_5 ➔ deposits are challenging due to: similarity of the surface properties and complex mineralogical particles.

2 Theory

$$Y = b0 + b1A * (X1A) + b1B * (X1B) + b2A * (X2A) + b2B * (X2B) + b3A * (X3A) + b3B * (X3B) + b3C * (X3C) + b4A * (X4A) + b4B * (X4B) + b4C * (X4C) + b5A * (X5A) + b5B * (X5B) + b5C * (X5C) + b5D * (X5D)$$

where Y is the studied response, which could be any of the seven responses abovementioned; Xi is the investigated factor (i varies from 1 to 5); A is the domain delimited by levels 1 and 2 of the factor Xi; B is the domain delimited by levels 2 and 3 of the factor Xi; C is the domain delimited by levels 3 and 4 of the factor Xi; D is the domain delimited by levels 4 and 5 of the factor Xi; biA is the Xi effect in the domain A; biB is the Xi effect in the domain B; biC is the Xi effect in the domain C; and biD is the Xi effect in the domain D

3 Material and methods

- Three oils extracted from waste biomass (OOC, POC, and COC) were evaluated as flotation collectors, with emphasis on the effect of saturated and unsaturated fatty acid content on the flotation performance of low-grade Moroccan phosphate ore.

- A representative phosphate ore sample, provided by the Office Chérifien des Phosphates (OCP Group, Morocco), was quartered to obtain the sample used in bench-scale flotation tests.

4 Results and discussion

Table 1 shows the main chemical composition of representative sedimentary phosphate, indicating that it is a typical calcareous and siliceous ore

Table 1: Chemical composition of the sedimentary phosphate ore

Compositions	P_2O_5	BPL	CaO	SiO_2	MgO	CO_2	Al_2O_3	Fe_2O_3	S	K_2O	TiO_2
Content (%)	20.3	44.35	40.48	25.95	2.06	8.26	1.16	0.32	1.38	0.39	0.10

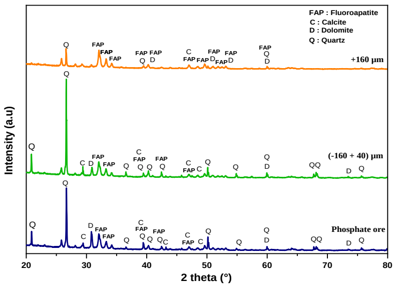


Fig. 1: Mineralogical analysis for sedimentary phosphate ore, the +160 µm, and the -160 µm size fraction.

The flowsheet of phosphate ore flotation experiments is presented in Fig.2

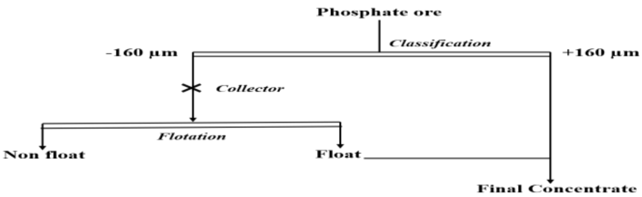
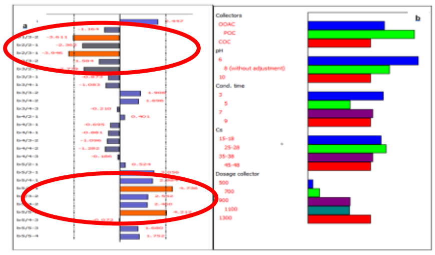


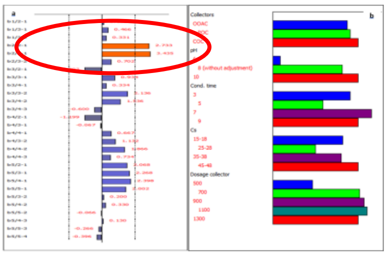
Fig. 2: Flowsheet of phosphate ore flotation experiment

The experimental design, applied for the size fractions flotation (-160, +40 µm), was performed here using the software package NemrodW_OPEX_2007. Table 3 displays the selected factors to study their effects on flotation efficiency. The response variables were the float yield; P_2O_5, CO_2, and SiO_2 grades; and the float recoveries R (P_2O_5), R (CO_2), and R(SiO_2).

Coded variable	Factor	Number of levels	Levels
X1	Collectors	3	OOC POC COC
X2	pH	3	6 8 10
X3	Conditioning time	3	3 5 7
X4	Solid concentration	4	15 25 35 45
X5	Dosage collector	5	500 700 900 1100 1300

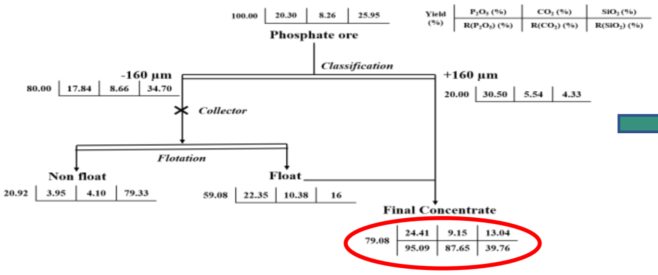


Graphical study of the factors' effects on the P_2O_5 grade response. a) Differences in the weight of the different levels and b) graphical study of the total effects



Graphical study of the factors' effects on the SiO_2 grade response. a) Differences in the weight of the different levels and b) graphical study of the total effects

5 Conclusion



P_2O_5 grade is upgraded from 20% to 24% using direct flotation with a recovery of 95% using 1300 g/t of POC at pH 6

Jchemo: Chemometrics and machine learning on high-dimensional data with Julia

M. Lesnoff[1,2,3]
[1]SELMET, Univ Montpellier, CIRAD, INRAe, Institut Agro, Montpellier, France
[2]CIRAD, UMR SELMET, Montpellier, France
[3]ChemHouse Research Group, Montpellier, France
E-Mail: matthieu.lesnoff@cirad.fr

Keywords: Chemometrics, Machine learning, Julia language, Toolbox.

1 Introduction

Julia (https://julialang.org) is a programming language designed for high performance. It is an open-source project made available under the MIT license. The language tries to tackle the "two-language problem" referring to the fact that many scientific codes are prototyped in a slow but flexible language (to test an idea quickly) but then have to be moved to a faster (e.g., C++) but less flexible language for practical applications. Julia allows fast computations with simple and easily readable coding. Works on Julia began in 2009. Julia's syntax is now considered stable, since version 1.0 in 2018 (actual version December 2025: 1.12.2), with many registered available packages and a very active users' forum (https://discourse.julialang.org).

The proposed poster will present Jchemo [1] (https://github.com/mlesnoff/Jchemo.jl), a Julia package (tool-box) dedicated to chemometrics and machine learning in general. Why did I decide to switch in 2021 from the language R to Julia for my chemometrics works? Trying to run a PLSR (25 LVs) with n = 1e6 samples and p = 500 variables with my function crashed systematically my R working sessions (with a I9 Intel processor). With the same computer and function written in Julia, the computation took 8 seconds. Why did I choose Julia compared to Matlab? Since Julia is free.

2 Theory

Julia programs automatically compile to efficient native code via LLVM, and support multiple platforms (Windows, MacOs, Linux etc.). Julia uses multiple dispatch as a paradigm, making it easy to express many object-oriented and functional programming pattern. The official IDE recommended for Julia users is Visual Studio Code (https://code.visualstudio.com).

3 Material and methods

Jchemo was built initially around partial least squares regression (PLSR) and discrimination (PLSDA) methods and their non-linear extensions, in particular locally weighted PLS models (kNN-LWPLS-R & -DA; e.g., [2]). The package has then been expanded with many other methods of dimension reduction, regression, discrimination, and signal (e.g., spectra) preprocessing.

Why the name Jchemo? Since it is oriented towards chemometrics, in brief the use of biometrics for chemistry data. But most of the provided methods are generic and can be applied to other types of data. The package has two related projects: JchemoData.jl (a container package of data sets used in the examples; https://github.com/mlesnoff/JchemoData.jl) and JchemoDemo (a pedagogical environment; https://github.com/mlesnoff/JchemoDemo).

Beside usual chemometrics methods (signal preprocessing, PCA, PLS etc.), multi-block methods are available for dimension reduction (e.g., MBPCA, ComDim, rCCA, etc.) and regression/discrimination (MBPLS, ROSAPLS, SOPLS, etc.) models. Various ridge and sparse models are proposed, as many nonlinear models useful for heterogeneous data (kernels-based models, kNN, RF). The syntax of Jchemo is very consistent between all the functions and can therefore be learned and used easily by non-specialists of programming.

The Jchemo functions are organized between: *transformation operators* (e.g., PCA models), *predictors* (e.g., PLSR/PLSDA models), and *utility functions*. Ad'hoc pipelines (chains of models) can also easily be built. In Jchemo, a pipeline is a chain of K models: either a set of K transformers, or a set of $K – 1$ transformers and a final predictor.

4 Results and discussion

The fitting of an ad'hoc pipeline is illustrated below. The example considers the "LWR" algorithm of Naes et al. [3] that consists in a preliminary global PCA on the data and then a kNN locally weighted multiple linear regression (kNN-LWMLR) on the global PCA scores:

```
model1 = pcasvd(; nlv = 25)                        # transformation operator
model2 = lwmlr(; metric = :eucl, h = 2, k = 200)   # predictor
model = pip(model1, model2)  # final pipeline (more than 2 models can be specified)
fit!(mod, X, Y)
pred = predict(mod, Xnew).pred
```

Efficient generic (i.e., the same for all models) functions allow to tune the models, by test-set validation or cross-validation. For instance, for the first, a grid-search for a gaussian KPLSR model is implemented by:

```
kern = [:krbf] ; gamma = [100, 1, .1, 0.001]
pars = mpar(kern = kern, gamma = gamma)  # the grid
nlv = 0:30
model = kplsr()
res = gridscore(model, Xcal, Ycal, Xval, Yval; score = rmsep, pars, nlv)
```

5 Conclusion

Jchemo is registered on the official Julia package repository (equivalent of the CRAN for R), It is an easy tool that can handle most of the most frequent needs in chemometrics, as well as at a baseline or a research level. The development of the package is active, the package being regularly updated by new functions (for instance, sparse PLS functions have been recently added).

6 References

[1] M. Lesnoff. Jchemo: Chemometrics and machine learning on high-dimensional data with Julia. 2021, https://github.com/mlesnoff/Jchemo. UMR SELMET, Univ Montpellier, CIRAD, INRA, Institut Agro, Montpellier, France.

[2] M. Lesnoff, M. Metz, J.M. Roger. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. *Journal of Chemometrics* n/a, e3209, 2020.

[3] T. Naes, T. Isaksson, B. Kowalski. Locally weighted regression and scatter correction for near-infrared reflectance data. Analytical Chemistry 664–673, 1990.

Identification de faces latérales et basales du talc par factorisation en matrices non-négatives en imagerie hyperspectrale Raman.

L. Govohetchan[1*], A. Razafitianamaharavo[1], F. Villieras[1], T. Kauffmann[2], J. F. L. Duval[1], D. Chapron[2], M. Offroy[1]

[1] Université de Lorraine, CNRS, LIEC, F-54000 Nancy, France

[2] Université de Lorraine, CNRS, LMOPS, Metz, France

*leon.govohetchan@univ-lorraine.fr

Mots-clés : Résolution multivariée de courbes, algorithme NMF, imagerie hyperspectrale Raman, MT-SVD, talc.

1 Introduction

Le talc est un minéral appartenant à la famille des phyllosilicates. Il est composé de magnésium, de silicate et d'un groupement doublement hydroxylé, de formule structurale théorique $Si_4\ Mg_3\ O_{1\ 0}\ (OH)_2$. Ce minéral présente des faces latérales hydrophiles et des faces basales hydrophobes. Toutefois, ces différentes surfaces ne sont pas directement accessibles en imagerie hyperspectrale Raman, en raison d'une résolution spatiale limitée à l'échelle du micromètre ainsi que de la limite intrinsèque de la spectroscopie moléculaire, en matière de sélectivité spectrale. Nous proposons ainsi une approche de décomposition bilinéaire de type factorisation par matrice non-négative (Non-negative Matrix Factorization en anglais, NMF) permettant de repousser ces contraintes et d'identifier différentes faces du minéral.

2 Théorie

Notre approche **NMF** [1] vise à décomposer une matrice positive $\mathbf{D}$ par le produit de deux matrices positives **W** et **H**, de dimensions réduites. L'objectif est de minimiser l'erreur $\mathbf{E} = \mathbf{D} - \mathbf{WH}$ en résolvant un problème d'optimisation non convexe sous contraintes de positivité. Les matrices obtenues, dites matrice des concentrations pures (**W**) et matrice des spectres purs (**H**), sont calculées au moyen d'un algorithme optimisation itératifs basé sur des mises à jour multiplicatives.

3 Matériel et méthodes

L'identification des faces latérales et basales du talc a été menée à partir d'un échantillon constitué de 30 microgrammes de talc déposés sur un support en or préalablement nettoyé à l'éthanol et séché à l'air. La caractérisation a été effectuée à l'aide d'un spectromètre Raman LabRAM HR Evolution utilisant une longueur d'onde d'excitation de 532 nm. Une zone d'analyse de 55 × 43 pixels a été définie à l'aide du microscope optique (objectif ×50) couplé au spectromètre (voir N°1 de la Figure), puis un cube de données spectrales a été acquis sur cette zone, avec un pas de 0,5 µm dans la direction x et y, sur une plage spectrale allant de 100 à 4000 $cm^{-\ 1}$. Une table motorisée a permis le déplacement de l'échantillon. Le prétraitement des données, une étape presque obligatoire [2] a été réalisé à l'aide du filtre de Whittaker ($\lambda = 100$, $p = 10^{-4}$) de la Toolbox MATLAB 2024, ainsi qu'avec la méthode MT-SVD [3]. La séparation des signaux sources a ensuite été effectuée par notre approche NMF.

4 Résultats et discussion

La Factorisation par Matrice Non-négative (NMF), basée sur les formules de mises à jour multiplicatives a permis de déterminer les sites hydrophobes caractéristiques des faces basales mais surtout des sites hydrophiles caractéristiques des surfaces latérales du talc. Avec les signatures spectrales extraites par la NMF, nous notons pour certaines d'entre elles des modes de vibrations intenses des liaisons T–O–T, Si–O, TO_4 et Si–O–Si, qui confèrent principalement un caractère hydrophobe à la surface du talc analysée par imagerie hyperspectrale Raman. En outre, la vibration de la molécule H-O-H est quant à elle uniquement observée sur une signature spectrale particulière caractéristique des faces latérales leur conférant un caractère hydrophile (cf. Figure).

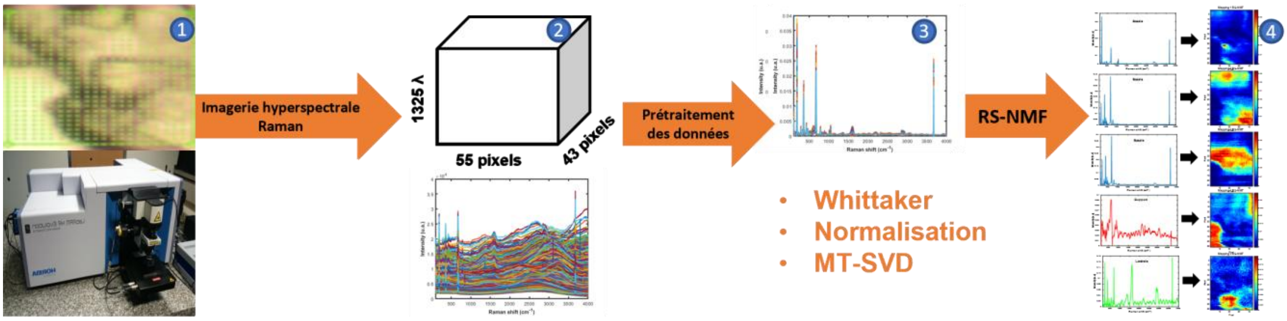


Figure : identification des faces latérales et basales du talc. (1) surface du talc analysée par l'imageur Raman ; (2) visualisation des données ; (3) prétraitement de données spectrales ; (4) extraction des signaux purs et des cartographies par RS-NMF.

5 Conclusion

Les limites de détection de l'imageur hyperspectral Raman ont été repoussée par une stratégie d'analyse et un traitement chimiométrique rigoureux, avec une particularité : une des premières applications de notre approche NMF sur un échantillon minéral.

6 Références

[1] M. Offroy et al, Enhanced Raman hyperspectral imaging using RS-NMF: a novel Regularized Sparse Non-Negative Matrix Factorization for spectral unmixing, Chemometrics and Intelligent Laboratory Systems, 105602 (2025). https://doi.org/10.1016/j.chemolab.2025.105602.

[2] M. Offroy, L. Duponchel, Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry, Analytica chimica Acta 910 (2016) 1-11 https://doi.org/10.1016/j.aca.2015.12.037.

[3] M. Haouchine et al, Handle Matrix Rank Deficiency, Noise, and Interferences in 3D Emission−Excitation Matrices: Effective Truncated Singular-Value Decomposition in Chemometrics Applied to the Analysis of Polycyclic Aromatic Compounds, ACS Omega 2022, 7, 23653-23661.

Image deconvolution to improve MCR solutions

Adrián Gómez-Sánchez, Raffaele Vitale, Cyril Ruckebusch

Dynamics, Nanoscopy Chemometrics (Dynachem), Laboratory of Advanced Spectorscopy, Interactions, Reactivity a,d Environement (LASIRe) CNRS U Lille France

Keywords: Shift-invariant convolution, Bilinear matrix factorization, Spectral unmixing, Rotational ambiguity

1 Introduction

Diffraction of light is a fundamental limit to spatial resolution in optical microscopy. In theory, diffraction can be modelled by the response of a focused optical imaging system to a point object, *i.e*. a convolution kernel shifted over all locations of a 2D image. In practice, diffraction translates into a blur in digital images, and deconvolution is the inverse operation that aims at correcting blur to provide an enhancement of the image spatial resolution (better contrast in smaller features) [1]. Multivariate curve resolution (MCR) which aims at spectral unmixing is also an inverse problem, formulated in terms of an ambiguous bilinear matrix factorization. Using proper spatial constraints can reduce the amount of ambiguity in the results yielded by the MCR analysis of spectral images [2]. In this presentation, we discuss an alternative approach that combines image deconvolution with MCR to improve the quality of the solutions obtained from diffraction-limited spectral images.

2 Theory

For spectral images, the MCR model can be written as in Equation 1, where the rows of **D** contain spectra, the columns of **C** the (unfolded) concentration maps and the columns of **S** the pure spectra of the individual components underlying the measured spectral pixels:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{T} \qquad \text{Eq.1}$$

When the convolution model is linear and shift-invariant, one way to solve the inverse deconvolution problem is then to introduce a Toeplitz deconvolution matrix **H** into the bilinear unmixing decomposition model itself. Additionally, a downsampling matrix **A** can be introduced, enabling better contrast in **X**, as shown in Equation 2:

$$\mathbf{D}=\mathbf{AHXS}^{T} \qquad \text{Eq. 2}$$

The MCR model in Equation 1 can be solved by alternating least squares (ALS), fixing one of the two unknown matrices (**C** or **S**) and solving the (convex) subproblem of minimizing the LS criterion with respect to the other one, iteratively, until convergence [3]. On the other hand, solving Equation 2 requires adapting this ALS scheme, because calculating **X** is now an ill-posed problem and penalized LS approaches should be considered. Additionally, scenarios where the blurring kernel in **H** is unknown (blind deconvolution) can be considered.

3 Results and discussion

We start with the direct formulation of the convolution model. We illustrate in Figure 1 the effect of image convolution and blurring on the amount of rotational ambiguity of the MCR factorization of a simulated spectral imaging dataset.

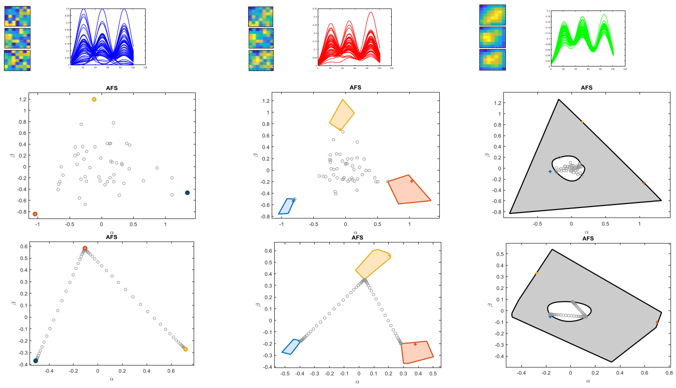


*Figure 1: **Image convolution yields more ambiguous MCR factorizations.** The figure displays, from left to right, the areas of feasible solutions obtained from simulated datasets with increasing degrees of spatial convolution and blurring.*

From the simulation in Figure 1, it can be clearly anticipated that when tackling the inverse problem in Equation 2, image deconvolution should somehow reduce such an ambiguity. In Figure 2, we provide the outcomes obtained from a three-component fluorescence image when implementing the decomposition in Equation 2 into the MCR-ALS framework and compare them to those resulting from the conventional approach.

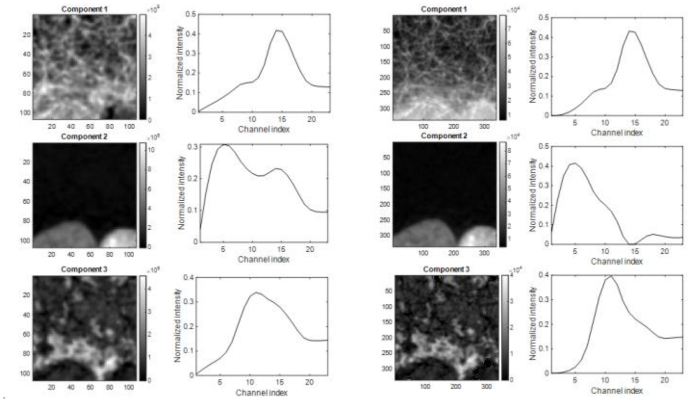


*Figure 2: **Embedding deconvolution into MCR-ALS.** Comparison of the results obtained from a three-component fluorescence image when exploiting the factorization models in Equation 1 (left panel) and Equation 2 (right panel), respectively.*

The main take-home message of this presentation will be the following: beyond spatial resolution enhancement, spectral unmixing solutions can be enhanced when performing image deconvolution.

3 References

[1] Eilers, P.H.C.; Ruckebusch, C. Fast and simple super-resolution with single images. Scientific Report. 2022, 12.

[2] Ghaffari, M.; Hugelier, S.; Duponchel, L.; Abdollahi, A.; Ruckebusch, C. Effect of image processing constraints on the extent of rotational ambiguity in MCR-ALS of hyperspectral images. Analytica Chimica Acta. 2019, 23.

[3] R. Tauler. Multivariate curve resolution of second-order data. Chemometrics and Intelligent Laboratory Systems. 1995, 133.

Exploiting the Complexity of MALDI Data through Chemometrics: From Pixel to Model

Yohann Clément[1], Baptiste Riou[1], Pierre Lanteri[1], Delphine Arquier[1], Justine Massias[2], Marie-Laure Plissonier[2], Arnaud Chaumot[3], Oliver Geffard[3], Davide Degli Esposti[3], Sophie Ayciriex[1]

[1] Univ Lyon, CNRS, Université Claude Bernard Lyon 1, Institut des Sciences Analytiques, UMR 5280, 5 rue de la Doua, F-69100, Villeurbanne, France

[2] Lyon Hepatology Institute, Lyon, France

[3] INRAE, UR RiverLy, Ecotoxicology Team, F-69625, Villeurbanne, France

Keywords: Madi Imaging, Data pretreatments, t-SNE, bisecting kmeans, PLS-DA

1 Introduction

Matrix-Assisted Laser Desorption/Ionization mass spectrometry imaging (MALDI-MSI) has become a powerful tool for the spatially resolved investigation of metabolomes and lipidomes in both biomedical and environmental contexts. However, the intrinsic complexity of MALDI data—characterized by high dimensionality, instrumental noise, mass drift, spectral redundancy, and non-linear relationships—requires robust and carefully structured chemometric approaches.

2 Material and methods

Particular emphasis is placed on critical preprocessing steps, including pixel-wise quality control, *m/z* recalibration based on matrix-derived anchor peaks, strict spectral alignment onto the manufacturer's mass grid, post-alignment peak picking, and deisotoping. This strategy significantly reduces instrumental variance and ensures both intra- and inter-image comparability, which is a prerequisite for reliable multivariate data analysis.

Normalized datasets are subsequently explored using descriptive multivariate methods (PCA, t-SNE, UMAP), allowing the direct visualization of anatomical and chemical structures through score images, as well as through clustering approaches (K-means, bisecting K-means, hierarchical and density-based methods). Finally, supervised models such as PLS-DA and kernel PLS-DA are applied to identify discriminant ions associated with contaminant exposure while explicitly accounting for the non-linear nature of MALDI data.

3 Results and discussion

Here, we present a comprehensive chemometric workflow dedicated to MALDI imaging data, spanning from raw data preprocessing to supervised modeling. The approach is illustrated through an ecotoxicological study investigating cadmium (Cd) exposure in *Gammarus fossarum*, a freshwater sentinel organism. MALDI-MSI acquisitions were performed using a timsTOF fleX

MALDI-2 mass spectrometer (Bruker), enabling the combined analysis of metabolomic and lipidomic information at the tissue level.

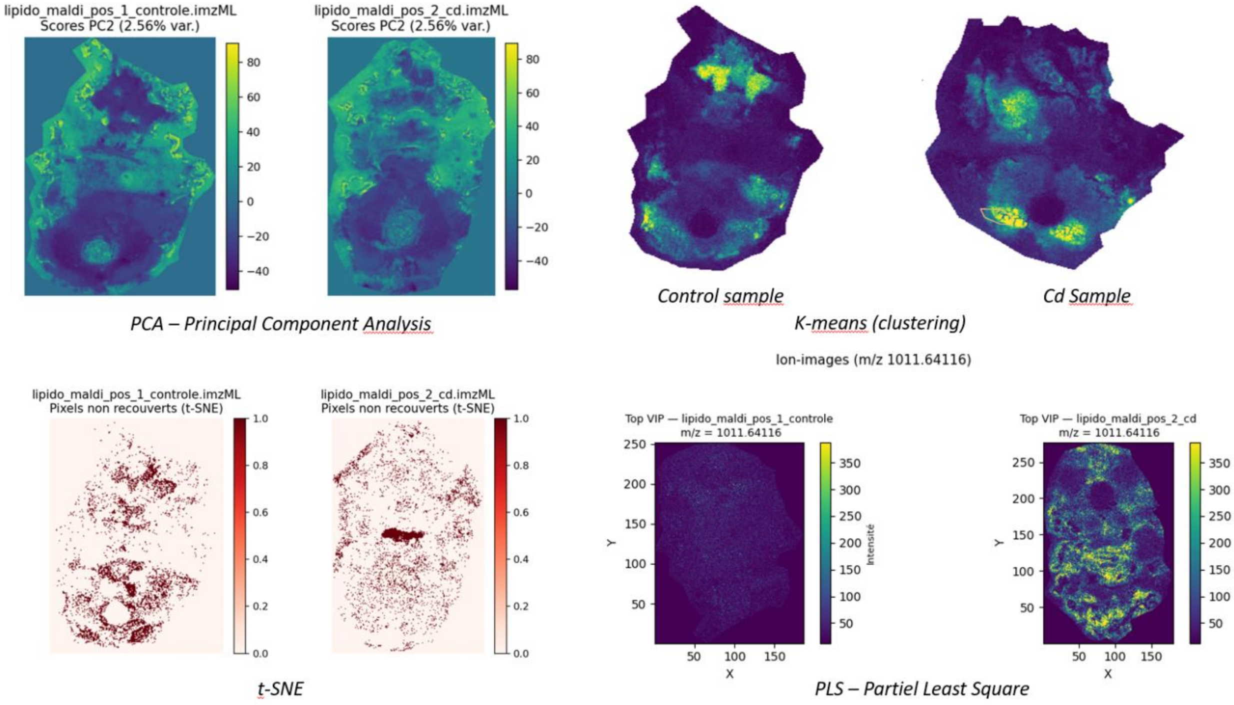

lipido_maldi_pos_1_controle.imzML
Scores PC2 (2.56% var.)
lipido_maldi_pos_2_cd.imzML
Scores PC2 (2.56% var.)
80
60
40
20
0
−20
−40
PCA – Principal Component Analysis
Control sample
Cd Sample
K-means (clustering)
Ion-images (m/z 1011.64116)
lipido_maldi_pos_1_controle.imzML
Pixels non recouverts (t-SNE)
lipido_maldi_pos_2_cd.imzML
Pixels non recouverts (t-SNE)
1.0
0.8
0.6
0.4
0.2
0.0
t-SNE
Top VIP — lipido_maldi_pos_1_controle
m/z = 1011.64116
Top VIP — lipido_maldi_pos_2_cd
m/z = 1011.64116
Y
X
250
200
150
100
50
350
300
Intensité
PLS – Partiel Least Square

Fig.1 : Data Processing Pipeline for MALDI Mass Spectrometry Imaging

4 Conclusion

Overall, this integrated framework highlights the central role of chemometrics in the quantitative and interpretable exploitation of MALDI imaging data, and demonstrates its relevance for applications in health and environmental sciences, from exploratory analysis to advanced modeling and chemical image reconstruction.

5 References

6 McDonnell, L. A., Heeren, R. M. A., *Imaging mass spectrometry.*
Mass Spectrometry Reviews, 26(4), 606–643 (2007).

7 Caprioli, R. M., Farmer, T. B., Gile, J. *Molecular imaging of biological samples: Localization of peptides and proteins using MALDI-TOF MS.***Analytical Chemistry**, 69(23), 4751–4760 (1997).

8 Alexandrov, T., *MALDI imaging mass spectrometry: statistical data analysis and current computational challenges.***BMC Bioinformatics**, 13(Suppl 16), S11 (2012).

9 Broadhurst, D., Kell, D. B., *Statistical strategies for avoiding false discoveries in metabolomics and related experiments.* **Metabolomics**, 2, 171–196 (2006).

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

Optimization of an innovative Non-Target Screening workflow for occupational multi-exposure assessment: from sample preparation to data processing

K. Mtitou[1] A. Martin Remy[1] G. Antoine [1] S. Ndaw[1] N. Grova[1] B. Habchi [1]

INRS (French National Research and Safety Institute for the Prevention of Occupational Accidents and Diseases), Department of Toxicology and Biomonitoring, Vandoeuvre-lès-Nancy, France, Kaoutar.mtitou@inrs.fr

Keywords: Non-target screening, Occupational exposure, Urine sample preparation, Data processing, Feature extraction

1 Introduction

Workers are often exposed to complex mixtures of chemicals, making the assessment of occupational poly-exposures particularly challenging. Traditional target analytical methods, while highly precise, are limited to detect a predefined set of compounds. In contrast, broader approaches such as suspect and non-target screening (SS/NTS) provide a comprehensive view of the metabolome without prior assumptions, offering a promising strategy for identifying both exposure and effect biomarkers [1]. Despite this potential, SS/NTS approaches face significant challenges, particularly in sample preparation and data processing, which remain critical steps in the metabolomic workflow.

2 Theory

SS/NTS workflow requires careful optimization of each analytical step to ensure reliable outcomes. In urinary metabolomics, pre-acquisition normalization, especially specific gravity (SG) adjustment, is effective in reducing dilution-related variability [2]. Moreover, the selection and optimization of extraction parameters in software such as MS-DIAL have a significant impact on signal detection and annotation [3]. In this study, we focused on these two key steps, optimizing both urine sample preparation and data-extraction parameters to enhance the robustness of the resulting metabolomic data.

3 Material and methods

Urinary sample preparation: A pooled urine sample was supplemented with 40 chemical standards at four concentrations (0, 5, 20, and 100 µg/L). Samples were prepared with or without SG normalization, and for both three dilution factors (2, 5, and 10) were evaluated to assess matrix effects. At the end of the preparation step, a mixture of three external standards were added at the same concentration in all samples. Liquid chromatography high-resolution mass spectrometry (LC-HRMS, VION IMS-QTOF, Waters) data were acquired, and 40 endogenous and the 40 exogenous compounds were manually extracted in MZmine to assess the influence of SG normalization and dilution factor on metabolite detection.

Optimization of data processing parameters: Following the determination of the optimal sample preparation conditions, seven individual urine pools and one quality control (QC) sample were each supplemented with 17 exogenous compounds at five concentration levels (0, 1, 5, 10, and 20 µg/L). Manual extraction of 17 exogenous and 20 endogenous compounds in MZmine was performed as a reference dataset. These results were then used to evaluate and optimize automatic feature-extraction parameters in MS-DIAL.

4 Results and discussion

Comparison of the external standards across SG normalization and dilution factors revealed that applying SG normalization in combination with a 1:10 dilution produced the greatest analytical stability, yielding markedly improved signal reproducibility and minimal retention-time drift. Under these conditions, the external standards showed the lowest variation (CV = 3%) and the highest chromatographic consistency (**Figure 1a)** confirming this workflow as optimal for minimizing matrix effects and reducing inter-individual variability while preserving relatively similar sensitivity for low-abundance exogenous substances. In parallel, the optimization of MS-DIAL parameters (minimum intensity, smoothing level, peak width) resulted in an optimal processing configuration that increased both the number of detected molecules and the number of accurately integrated peaks compared with the default parameters (**Figure 1b**).

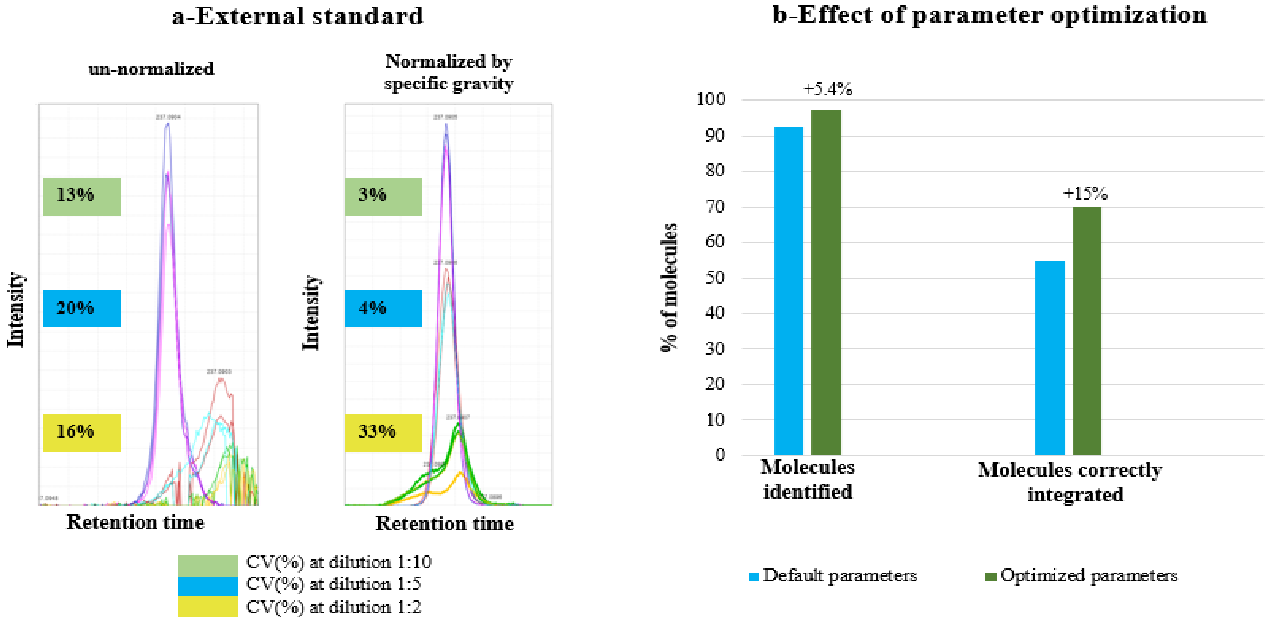


Figure 1: (a) Evaluation of urinary sample preparation conditions **(b)** Effect of MS-DIAL parameter optimization on feature detection and peak integration.

5 Conclusion

These findings point out the importance of optimizing each step in the SS/NTS workflow to generate robust and reliable data. They further highlight the ability of the method to detect a wide range of endogenous and exogenous compounds in a single run. This work represents an initial step toward applying SS/NTS to larger occupational cohorts, allowing the assessment of multiple exposures and enhancing its usefulness for identifying both exposure and effect biomarkers.

6 References

1. Habchi, B. and A. Rémy, *un nouvel outil pour l'analyse des polyexpositions chimiques professionnelles : la métabolomique non ciblée* 2023.

2. Rosen Vollmar, A.K., et al., *Normalizing Untargeted Periconceptional Urinary Metabolomics Data: A Comparison of Approaches.* Metabolites, 2019. **9**(10).

3. Wang, X.C., et al., *A comparison of feature extraction capabilities of advanced UHPLC-HRMS data analysis tools in plant metabolomics.* Anal Chim Acta, 2023. **1254**.

Analyte-informed multivariate calibration

R. Nikzad-Langerodi[1]

[1] Software Competence Center Hagenberg, Softwarepark 32a 4231 Hagenberg, Austria, ramin.nikzad-langerodi@scch.at

Keywords: Multivariate Calibration, Tikhonov Regularization, Domain Adaptation

1 Introduction

Partial least squares regression (PLS) is a foundational pilar of multivariate calibration in analytical chemistry. PLS "learns" the relationship among the input variables and between inputs and the target variable(s) from data alone. Purely data-driven models, however, tend to overfit, especially in data deprived settings with few samples and many variables often yielding models that exhibit poor generalization. Data-preprocessing has traditionally been employed to improve the robustness of multivariate calibrations by infusing the chemometric modeling pipeline with domain knowledge. Such pre-processing is usually aimed at removing sources of variability not related to the response variable, e.g., light scattering in spectroscopy. Approaches to inform chemometric models about (latent) input patterns that cause the variation in the response, however, are as yet largely missing.

In the present contribution we revisit the Tikhonov regularization framework introduced by Nikzad-Langerodi *et al.* 2018 [1] and study its application for "analyte-informed" multivariate calibration, where the Tikhonov matrix encodes the pure/net signal of the target analyte. We test our approach on simulated data, where standard PLS fails to generalize under alteration of the correlation structure of analyte and interferent. Our results indicate that robustness of multivariate calibrations can be drastically improved when the (correct) net analyte signal is imposed on the first PLS weight vector.

2 Theory

The Tikhonov regularization approach by Nikzad-Langerodi *et al.* 2018 [1] computes a *m x 1* penalized PLS weight vector **w** by solving

$$\min_{\mathbf{w}} \|\mathbf{X} - \mathbf{y}\mathbf{w}^T\|_F^2 - \lambda\, \mathbf{w}^T \mathbf{\Gamma} \mathbf{w} \qquad (1)$$

for negative values of λ[1]**.** **X** and **y** denote the *n x m* input matrix holding *n* samples measured in *m* variables, and the *n x 1* vector holding the corresponding response values, respectively. When the Tikhonov matrix is chosen such that $\mathbf{\Gamma}=\mathbf{a}\mathbf{a}^{\mathbf{T}}$ with **a** denoting the *m x 1* vector holding the net analyte signal (NAS), the solution to the optimization problem in Eq. (1) will be increasingly biased towards the direction of **a** as λ increases. The minimizer is given by

[1] Eq. (1) is a zero-bounded, convex optimization problem if **Γ** is a positive semi-definite Tikhonov matrix and $\lambda < 0$.

$$\mathbf{w}^* = (\|\mathbf{y}\|_2^2\mathbf{I} - \lambda\boldsymbol{\Gamma})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2}$$

if λ is chosen such that the Hessian $\mathbf{H} = (\|\mathbf{y}\|_2^2\mathbf{I} - \lambda\boldsymbol{\Gamma})$ is positive semi-definite (p.s.d.).

3 Material and methods

Implementation of analyte-informed PLS (aPLS) was carried out using a standard PLS1 NIPALS framework with scikit-learn compatible API (see https://github.com/B-Analytics/diPLSlib), whereas λ in Eq. (2) was tuned using 7-fold cross-validation to attain maximal dot product between $\frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$ and $\frac{\mathbf{a}}{\|\mathbf{a}\|_2}$, i.e., alignment between $\mathbf{w}$ and $\mathbf{a}$. The net analyte signal $\mathbf{a}$ was computed as follows: 1) Removal of the analyte information from the spectral matrix $\mathbf{X}_{\text{res}} = \mathbf{X} - \mathbf{y}\mathbf{s}^T$ with $\mathbf{s}$ denoting the pure analyte spectrum. 2) Mean centering of the residual matrix $\mathbf{X}_{\text{res}} \coloneqq \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{X}_{\text{res}}$ and 3) computation of the orthogonal complement of $\mathbf{s}$ w.r.t. $\mathbf{X}_{\text{res}}$, i.e., $\mathbf{a} = \mathbf{s} - \text{proj}_{\mathbf{X}_{\text{res}}}(\mathbf{s})$.

4 Results and discussion

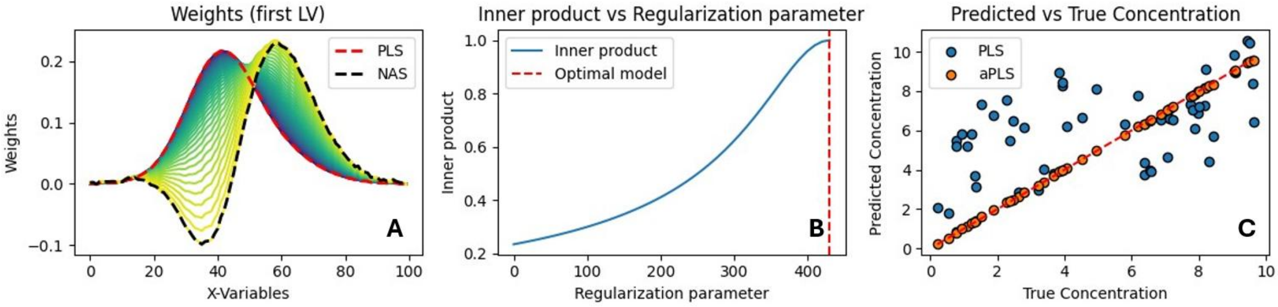


Figure 1. Analyte-informed PLS. A) PLS weights at increasing regularization (blue to yellow). B) Inner product between weights and NAS at increasing regularization. The red line indicates the optimal choice of the regularization parameter. C) Generalization in some target domain, where correlation between analyte and interferent has been removed. Blue: PLS, red: aPLS.

Figure 1A shows how net analyte structure is gradually imposed on the first PLS weight vector from a simulated dataset with increasing regularization. The optimal value of the regularization parameter was found by maximizing the dot product between the PLS weights and the (normalized) net analyte signal (Figure 1B). aPLS exhibited drastically improved generalization when the correlation between analyte and interferent concentration was removed from the test data (Figure 1C).

5 Conclusion

Imposing *a priori* knowledge about the data generating process e.g., (pure/net analyte signal) implicitly on latent variable models remains a largely unexplored avenue in the realm of multivariate calibration and holds promise for increasing the robustness of such models.

6 References

[1] Nikzad-Langerodi, Ramin, et al. "Domain-invariant partial-least-squares regression." *Analytical chemistry* 90.11 (2018)

Acknowledgements

The research was supported by the Austrian ministries BMIMI, BMWET and the State of Upper Austria in the frame of the SCCH COMET competence center INTEGRATE (FFG 892418) ant the competence center CHASE (FFG 901897).

RS-NMF: Regularized sparse Non-negative Matrix Factorization algorithm for enhanced spectral unmixing in Raman hyperspectral imaging

M. Offroy[1], A. Ayadi[1], L. Govohetchan[1], J. Ayoub[2], T. Hancewicz[3], L. Duponchel[4], M. Marchetti[2]

[1] Université de Lorraine, CNRS, LIEC, F-54000 Nancy, France.

[2] Université Gustave Eiffel, MAST-CPDM, F77454 Marne-la-Vallée, France.

[3] TMH Associates, Whitehall, Pennsylvania, USA.

[2] Univ. Lille, CNRS, UMR 8516, LASIRE-Laboratoire avancé de spectroscopie pour les interactions, la réactivité et l'environnement, F-59000, Lille, France.

Keywords: RS-NMF, Non-Negative Matrix Factorization, Raman hyperspectral imaging, Multivariate Curve Resolution, spectral unmixing, selectivity problems, chemometrics.

1 Introduction

Molecular spectroscopy, especially Raman hyperspectral imaging, is crucial for non-destructive analysis of heterogeneous samples. However, data analysis is often complicated by spectral overlap, making the identification of pure chemical components challenging. Signal unmixing methods have been developed to address this, yet they face inherent limitations (rotational ambiguity, noise sensitivity) when dealing with increasing sample complexity. This work introduces RS-NMF (Regularized Sparse Non-negative Matrix Factorization) [1], a novel algorithm designed to overcome these analytical challenges for spectral unmixing in Raman hyperspectral imaging.

2 Theory of the Regularized Sparse Non-negative Matrix Factorization

For the application to hyperspectral imaging, the experimental data cube is unfolded into a matrix **X(**$n\times m$) containing variables (spectral wavelengths) in the rows and sample (pixels) in the columns. NMF aims to approximate this positive matrix as the product of two non-negative matrices, **W** (pure spectra profiles) and **H** (pure concentration profiles), following the bilinear model:

$$\mathbf{X} = \mathbf{WH} + \mathbf{E} \tag{1}$$

where **E**(nxm) is the error matrix. The core of the RS-NMF proposed is a minimization of a cost function based on the Frobenius norm with regularization (Eq. 2), which quantifies the difference between the original matrix and its approximation. The proposed RS-NMF algorithm originates from the category of gradient descent algorithms, for which 'simple' multiplicative update rules have been derived.

$$\Psi\mathrm{c}\ (\mathbf{X} \,||\, \mathbf{WH}) = ||\mathbf{X} - \mathbf{WH}||^{\mathbf{2}} + \alpha \mathrm{J}_1(\mathbf{W}) + \beta J_2(\mathbf{H}) \tag{2}$$

3 Material and methods

The RS-NMF method was evaluated using two distinct datasets: (i) an experimental dataset (named "emulsion") for comparative analysis [2], and (ii) a simulated dataset (with 7 components) specifically designed to test for collinearity challenges in spectral unmixing. RS-NMF results were systematically compared against the well-established MCR-ALS method [3].

4 Results and discussion

The application of the RS-NMF method on the complex experimental "emulsion" dataset resulted in the identification of one additional chemical component compared to the conventional MCR-ALS approach. Furthermore, on the simulated dataset containing 7 chemical components, RS-NMF demonstrated superior performance by successfully recovering all 7 spectral signatures, while MCR-ALS failed to extract the full set of sources under identical initialization and constraint settings (Figure 1). These results confirm that RS-NMF pushes back the limits of source signal unmixing. Future work will focus on implementing additional constraints (e.g., unimodality), improving initialization strategies, coupling the method with pre-processing techniques (e.g., MT-SVD [4]), and extending the methodology to tensor data, thus opening new applications.

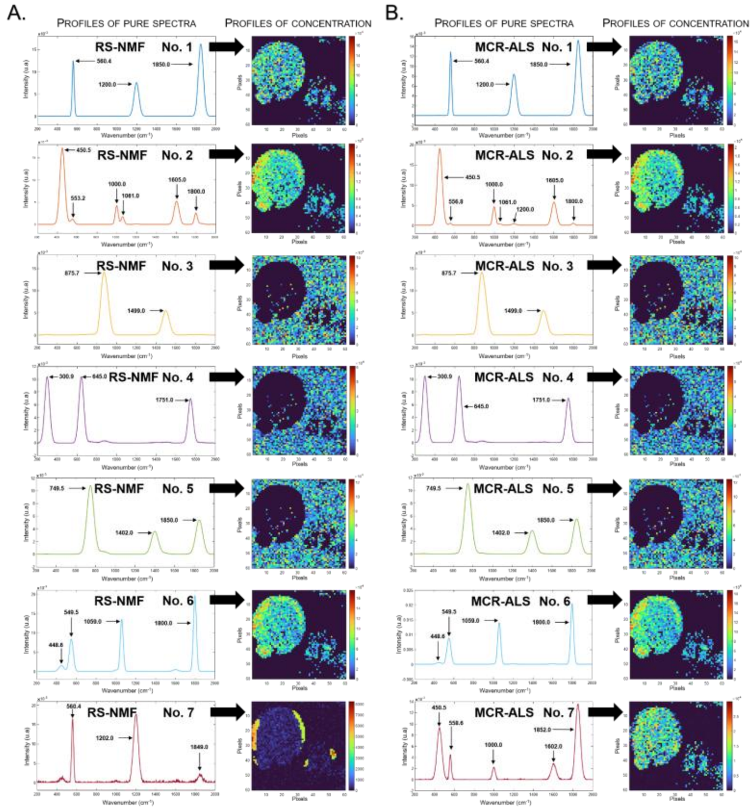


Figure 1 – Comparison of pure spectral components and concentration profiles extracted on the simulated sample: (A) RS-NMF results. (B) MCR-ALS results.

5 References

[1] M. Offroy, A. Ayadi, L. Govohetchan, J. Ayoub, T. M. Hancewicz, L. Duponchel, M. Marchetti, Enhanced Raman hyperspectral imaging using RS-NMF: a novel Regularized Sparse Non-negative Matrix Factorization for spectral unmixing, Chemometrics and Intelligent Laboratory Systems, 269, 105602, 2026.

[2] J.J. Andrew, T.M. Hancewicz, Rapid Analysis of Raman Image Data Using Two-Way Multivariate Curve Resolution, Appl. Spectrosc. 52 (1998) 797–807.

[3] A. de Juan, R.Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review, *Analytical Chimica Acta* 1145, 59-78, 2021. https://doi.org/10.1016/j.aca.2020.10.051

[4] Haouchine, C. Biache, C. Lorgeoux, P. Faure, M. Offroy, Handle Matrix Rank Deficiency, Noise, and Interferences in 3D Emission–Excitation Matrices: Effective Truncated Singular-Value Decomposition in Chemometrics Applied to the Analysis of Polycyclic Aromatic Compounds, ACS Omega 7, 23653–23661, 2022. https://doi.org/10.1021/acsomega.2c02256.

Does Data Conversion Matter? Assessing Its Impact in LC–MS Untargeted Metabolomics

Remy De Boni[1] Arnaud Salvador[1] Thomas Brunet[1] Valentina Calabrese[1] Delphine Arquier[1] Olivier Geffard[2] Arnaud Chaumot[2] Davide Degli-Esposti[2] Sophie Ayciriex[1] Pierre Lanteri[1] Yohann Clement[1]

[1] Universite Claude Bernard Lyon 1, CNRS, ISA, UMR5280, 5 rue de la Doua, F-69100 Villeurbanne

[2] INRAE, UR RiverLy, Laboratoired'écotoxicologie, Villeurbanne F-69625, France

Keywords: LC–MS, data conversion, mzML, centroiding, metabolomics

1 Introduction

Untargeted LC–MS metabolomics produces highly complex datasets characterized by extreme dimensionality, strong multicollinearity, heteroscedasticity, and structured noise. Such properties make chemometric methods indispensable for data exploration, modeling, and interpretation. While substantial efforts have been devoted to feature detection, normalization strategies, and multivariate statistical modeling, the impact of upstream data conversion — particularly profile-to-centroid transformation — remains largely overlooked from a chemometric perspective. Yet, data conversion constitutes the first operation that reshapes raw instrumental signals into numerical objects amenable to multivariate analysis, thereby directly conditioning variance structure, correlation patterns, and latent variable representations. In modern LC–MS workflows, raw data are typically acquired in profile mode and stored in vendor-specific formats, then converted into centroided, community-standard formats such as mzML [1] prior to downstream processing. Several centroiding strategies coexist, including wavelet-based algorithms (CWT), vendor-embedded methods (Vendor), both from MSconvert (Proteowizard [2]), proprietary manufacturer converters (e.g., Sciex® MS Data Converter) or open-source initiative such as MSnbase [3]. These algorithms rely on distinct signal-processing assumptions, but their consequences on the multivariate structure of metabolomics datasets are rarely evaluated explicitly. In this work, we adopt a chemometric viewpoint to assess whether centroiding choices act as a hidden source of variability and bias in untargeted LC–MS metabolomics.

2 Material and methods

Biological samples (n=17) were obtained from an ecotoxicological study focusing metabolites in *Gammarus fossarum*. LC–MS analyses were performed on a ZenoTOF 7600 system (Sciex®), and raw profile data were stored in .wiff format. Identical datasets were subsequently converted into mzML using three widely employed centroiding approaches: Continuous Wavelet Transform, vendor-based centroiding, and Sciex proprietary algorithms, called further respectively CWT, Vendor and Sciex. Chemometric analyses were applied directly to the converted datasets to characterize how centroiding strategies propagate into multivariate space. Exploratory data analysis was conducted using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize global variance structure and sample organization. Supervised modeling was performed using Partial Least Squares Discriminant Analysis (PLS-DA) to investigate how centroiding affects discriminant variance and the identification of influential regions in variable space.

3 Results and discussion

Initial inspection revealed substantial differences in file size, point density, and intensity distributions between centroiding strategies (Figure 1A, B and C), indicating marked discrepancies in retained information content. These differences are translated directly into altered variance structures. PCA score plots (Figure 1D) consistently showed clustering driven by the centroiding algorithm itself, independently of biological condition, highlighting the strong influence of data conversion on the dominant sources of variance. Non-linear embeddings further emphasized algorithm-specific structuring of the data manifold. PLS-DA models confirmed that centroiding strategies profoundly reshape the latent discriminant space. Depending on the conversion method, different regions of the variable space dominated class separation, and the overall geometry of the latent components was substantially modified. CWT-based centroiding resulted in aggressive data sparsification due to intrinsic filtering steps, whereas vendor-based and Sciex algorithms preserved a higher density of data points but introduced major shifts in intensity scaling, spanning several orders of magnitude. Therefore, apparent discrimination patterns could be partially or predominantly driven by conversion-induced effects rather than biological variation.

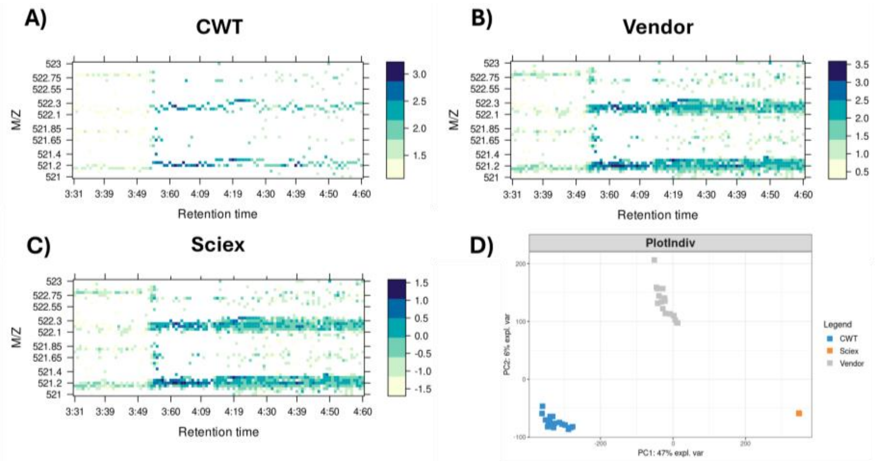


Figure 1 : A,B,C) LC–MS maps of the same sample displayed over an identical retention time–m/z region, obtained using three different centroiding strategies (intensity scale displayed in log10). D) PCA score plot of HP metabolite datasets obtained using three different centroiding approaches

4 Conclusion

Taken together, these results demonstrate that data conversion is not a neutral technical preprocessing step but a critical chemometric determinant of LC–MS metabolomics workflows. Centroiding strategies directly shape variance distribution, correlation structure, and latent variable representations, thereby influencing feature detection, statistical modeling, and biological interpretation. This work highlights the necessity of explicitly integrating data conversion into chemometric validation frameworks and urges the metabolomics community to reconsider centroiding choices as an integral component of model robustness, reproducibility, and interpretability in addition to other factors, such as peak picking algorithms [4].

5 References

[1] Martens, Lennart et al. "mzML--a community standard for mass spectrometry data." Molecular & cellular proteomics : MCP vol. 10,1 (2011): R110.000133. doi:10.1074/mcp.R110.000133

[2] Chambers, Matthew C et al. "A cross-platform toolkit for mass spectrometry and proteomics." Nature biotechnology vol. 30,10 (2012): 918-20. doi:10.1038/nbt.2377

[3] Gatto, Laurent, and Kathryn S Lilley. "MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation." Bioinformatics (Oxford, England) vol. 28,2 (2012): 288-9. doi:10.1093/bioinformatics/btr645

[4] Aigensberger, Markus et al. "Modular comparison of untargeted metabolomics processing steps." Analytica chimica acta vol. 1336 (2025): 343491. doi:10.1016/j.aca.2024.343491

Simple Uncertainty Quantification Methods in Neural Networks for Spectral Data Using ViT: Application to Mango Dry Matter Prediction

Khadija Lamdibih1, Florent Abdelghafour1, Metz Maxime2

1 ITAP, Univ. Montpellier, INRAE, Institut Agro, 34196 Montpellier, France

2 Pellenc ST, Applied Research Group, 84120 Pertuis, France

Keywords: Uncertainty Quantification, SWAG, MC Dropout, ViT-1D, VIS–NIR Spectroscopy

1 Introduction

Deep learning has demonstrated strong predictive performance for spectroscopic and chemometric applications. However, its adoption in routine analysis remains limited due to concerns related to model interpretability and prediction reliability, particularly when models are applied outside their calibration domain. In spectroscopic contexts, variations in measurement conditions, seasons, or biological properties may lead to unreliable predictions if uncertainty is not properly quantified. Conventional neural networks provide only point estimates, which may be misleading in such situations. This motivates the integration of uncertainty quantification (UQ) methods to complement predictions with confidence estimates and improve trust in deep chemometric models.

2 Theory

Uncertainty quantification aims to characterise the reliability of model predictions by estimating predictive variability. Practical approximation methods such as Monte Carlo (MC) Dropout and Stochastic Weight Averaging Gaussian (SWAG) provide efficient alternatives to fully Bayesian neural networks. MC Dropout estimates uncertainty by performing stochastic forward passes with dropout activated at inference, while SWAG approximates a Gaussian distribution over network weights using snapshots collected during training. These approaches allow estimation of epistemic uncertainty with limited computational overhead.

3 Material and methods

VIS–NIR spectra of mangoes collected across multiple seasons and cultivars were used to predict dry matter content. The dataset includes measurements acquired under varying agronomic conditions, providing a realistic evaluation scenario. Spectra were preprocessed using standard chemometric techniques and concatenated to form the input signal.

A one-dimensional adaptation of the Vision Transformer (ViT-1D) was employed to model the spectral data. Two uncertainty quantification strategies, MC Dropout and SWAG, were applied to the model. Predictive performance and uncertainty calibration were assessed using metrics such as RMSE, coefficient of determination (R^2), and coverage of confidence intervals.

4 Results and discussion

Both uncertainty quantification methods provided meaningful estimates of predictive uncertainty. MC Dropout yielded stable predictions but tended to underestimate uncertainty, resulting in limited coverage. In contrast, SWAG achieved improved predictive accuracy and more coherent uncertainty calibration, particularly for intermediate learning rates and larger patch sizes. The results highlight the sensitivity of Transformer-based models to hyperparameters and domain shifts, while demonstrating the potential of UQ methods to enhance prediction reliability in spectroscopic regression tasks.

5 Conclusion

This study demonstrates that integrating uncertainty quantification with a ViT-1D architecture improves the robustness and practical relevance of deep learning models for spectroscopic applications. MC Dropout and SWAG both provide valuable confidence estimates, with SWAG offering better calibration in the studied context. These results confirm the importance of uncertainty-aware modelling for deep chemometrics, particularly when models are applied outside their calibration domain.

6 References

[1] P. Mishra, D. Passos. A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit. Chemometrics and Intelligent Laboratory Systems, 212 (2021) 104287.

[2] P. Mishra, D. Passos, F. Marini, J. Xu, J. M. Amigo, A. A. Gowen, J. J. Jansen, A. Biancolillo, J. M. Roger, D. N. Rutledge, A. Nordon. Deep learning for near-infrared spectral data modelling: Hypes and benefits. TrAC Trends in Analytical Chemistry, 157 (2022) 116804.

[3] M. Maxime, K. Lamdibih, J.-M. Roger, D. Esteve, R. Bendoula, F. Abdelghafour. Simple methods for uncertainty estimation in neural networks applied to spectral data processing: A case study on mango dry matter prediction. Chemometrics and Intelligent Laboratory Systems, 267 (2025) 105532.

[4] Y. Gal, Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016.

[5] W. J. Maddox, T. Garipov, P. Izmailov, D. Vetrov, A. G. Wilson. A simple baseline for Bayesian uncertainty in deep learning. Advances in Neural Information Processing Systems (NeurIPS), 2019.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR), 2021.

[7] P. Fu, Y. Wen, Y. Zhang, L. Li, Y. Feng, L. Yin, H. Yang. SpectraTr: A novel deep learning model for qualitative analysis of drug spectroscopy based on transformer structure. Journal of Innovative Optical Health Sciences, 15 (3) (2022) 2250021.

The MultANOVA framework to analyze designed high-dimensional data: an outperforming alternative to PLS-DA, ASCA, rMANOVA and VASCA

B. Mahieu[1] V. Cariou[2]

[1] Rue de la Géraudière, CS 82225, 44322 Nantes, benjamin.mahieu@oniris-nantes.fr

[2] Rue de la Géraudière, CS 82225, 44322 Nantes, veronique.cariou@oniris-nantes.fr

Keywords: High dimensional data, Experimental design, Variable Selection, Partial-Least-Squares Discriminant-Analysis, Anova Simultaneous Component Analysis.

1 Introduction

Analytical platforms produce high-dimensional data, typically arising from experimental designs, and one of the main concerns for practitioners remains the evaluation of the effects corresponding to the factors and their interactions. In this configuration, analyses encounter the pitfall of the numerical superiority of variables over observations. Since these conditions prevent the application of MANOVA on such data, several alternatives were developed: PLS-DA [1] for a single factor and ASCA [2], rMANOVA [3], and VASCA [4] for multiple factors. However, these methods suffer from limitations: they involve substantial computation time to test significance, lack formal tests for pairwise comparison of levels of a significant factor, and, for ASCA and rMANOVA, do not provide means for selecting or ranking variables. To address this issue, the unified MultANOVA framework was recently proposed [5]. After presenting MultANOVA, a simulation study and real data analysis demonstrate how it outperforms existing methods.

2 Theory

MultANOVA encompasses four main stages. First, to test the significance of factors and interactions, a Fisher test is performed for each response variable and an FDR correction is applied. The smallest corrected p-value is retained as the MultANOVA p-value. Second, if a factor or an interaction effect is significant, pairwise testing of the levels are performed by means of Multiple Least Squares Difference tests (MultLSD), which operates similarly to MultANOVA, but uses LSD tests. Third, to visualize the discrimination structure, the standard Canonical Discriminant Analysis (CDA) is revisited leading to Diagonal CDA. Finally, the variable-wise Fisher test statistics and their corresponding p-values can be used to rank and select variables, respectively.

3 Simulation study design

A simulation study was conducted to compare the computational time and statistical performances (control of alpha risk and power) of MultANOVA with those of ASCA, VASCA and rMANOVA. To this end, effect sizes, p/n ratios, proportions of missing data, error correlations, and the presence of other significant factors and/or interactions in the model were manipulated throughout the simulations. The situation of p/n<1 was also considered, which made it possible to include standard

MANOVA in the comparison scheme. Using a similar protocol, the results of the MultANOVA's embedded variable selection was compared to this of VASCA and MultLSD tests investigated. The variable ranking from the MultANOVA framework was compared to PLS-DA's VIP on spectroscopic data. Finally, the entire MultANOVA workflow was applied to metagenomics data.

4 Results and discussion

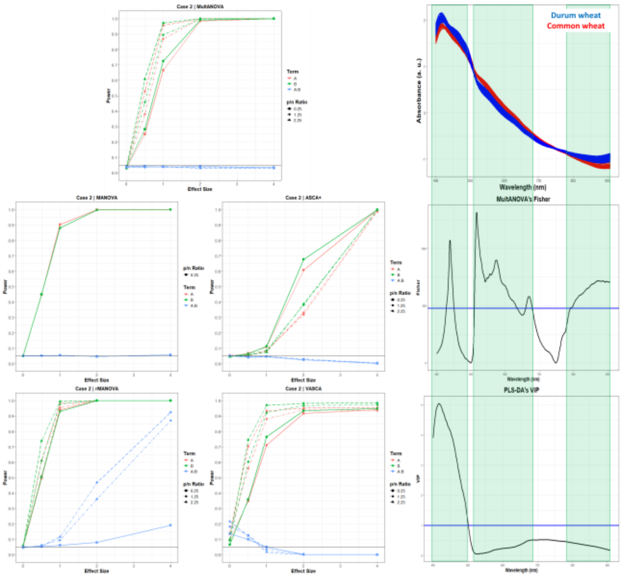


Figure 1 – Subset of the simulation results comparing MultANOVA tests with MANOVA, ASCA, rMANOVA and VASCA where effects A and B are under the alternative and their interaction under the null hypothesis (left) and variable ranking of MultANOVA and PLS-DA with SNV preprocessed spectra (right).

Figure 1 (left) shows that MultANOVA (on top) is the most powerful method that maintains control of the alpha risk and that exhibits the closest behavior to standard MANOVA when applicable, while being 100 times faster than ASCA to compute. Figure 1 (right) shows that MultANOVA's variable ranking is consistent with a visual inspection of the spectra, as opposed to PLS-DA. Additional results (and theory) will be presented during the communication.

5 Conclusion

The MultANOVA framework is proposed as an alternative to existing chemometric methods to analyze designed high dimensional data. The benefits of the MultANOVA framework are several: straightforward, fast to compute, better statistical properties than existing methods and a comprehensive consistency. All the methods developed within this framework are implemented in the MultANOVA R Package: CRAN: Package MultANOVA.

6 References

[1] Barker M, Rayens W. Partial least squares for discrimination. J Chemom. 2003;17(3):166-173. doi:10.1002/cem.785.

[2] Thiel M, Féraud B, Govaerts B. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. J Chemom. 2017;31(6):e2895. doi:10.1002/cem.2895.

[3] Engel J, Houthuijs KJ, Vasiliou V, Charkoftaki G. Regularized Multivariate Analysis of Variance. In: Comprehensive Chemometrics. Elsevier; 2020:479-494. doi:10.1016/B978-0-12-409547-2.14577-9.

[4] Camacho J, Vitale R, Morales-Jiménez D, Gómez-Llorente C. Variable-selection ANOVA Simultaneous Component Analysis (VASCA). Wren J, ed. Bioinformatics. 2023;39(1):btac795. doi:10.1093/bioinformatics/btac795.

[5] Mahieu B, Cariou V. MultANOVA Followed by Post Hoc Analyses for Designed High- Dimensional Data: A Comprehensive Framework That Outperforms ASCA, rMANOVA, and VASCA. J Chemom. 2025;39(7). doi:10.1002/cem.70039.

Study of the techno-functional properties of plant-based proteins – DATAVEG project

Caera O’Neill[1], Laurence Dujourdy[2, 3], Aurélie Lagorce[1], Camille Loupiac[1]

[1] UMR PAM, Institut Agro, Université Bourgogne Europe, INRAE, F-21000 Dijon, France, caera.oneill@agrosupdijon.fr

[2] Institut Agro Dijon, Direction Scientifique, Cellule d’Appui a` la Recherche en Sciences des Données, F-21000 Dijon, France

[3] Laboratoire d’Informatique de Bourgogne (LIB), Data Science, Université Bourgogne Europe, F-21000 Dijon, France

Keywords: alternative proteins, infrared spectroscopy, functionality, bibliometrics, chemometrics, gelation; sustainability.

1 Introduction

The transition to sustainable food systems underscores the need to study the techno-functional properties of plant-based proteins. To meet the rising demand for plant-derived alternatives, their physicochemical properties must be characterised and optimised for food applications. The DATAVEG project evaluates the processability of plant protein sources (soy, pea, potato, gluten) using advanced methods and multi-criteria analysis, aiming to develop harmonised characterisation tools for new product formulation. It brings together complementary expertise in food biochemistry, process engineering and industrial application through a multidisciplinary collaboration. The study is organised into **three parts**: a **bibliometric review** of plant protein characterisation methods; an **analysis of physico-chemical and functional data** from the DATAVEG project; and **perspectives on advanced chemometric approaches** linking technical and functional properties. This structured approach aims to identify best practices for assessing plant protein performance and to support their use in developing alternative food products.

2 Material and methods

2.1. Bibliometric research

Bibliometrics is the quantitative analysis of academic publications used to evaluate knowledge production and dissemination. It relies on statistical and computational methods to measure research productivity, impact, thematic trends and collaboration networks. Among available bibliographic databases, the Web of Science (WoS) was selected for this study due to its widespread use in academic research. Data analysis was conducted with the Bibliometrix R package [1], which offers functions for extracting and analysing bibliographic information. This tool also enables the visualisation of results, helping researchers efficiently explore scientific literature in their field.

2.2. Physico-chemical methods of samples and multivariate analysis

Seven plant protein powders were studied in the DATAVEG project: wheat gluten, potato (S200, S300), soy (concentrate, isolate) and pea (concentrate, isolate) proteins, referred to respectively as Potato 200, Potato 300, Soy Concentrate, Soy Isolate, Pea Concentrate, Pea Isolate and Gluten. They were characterised by nitrogen content, spectroscopic methods, thermal denaturation and protein fraction analysis. Functional properties including gelation, water-holding capacity and viscosity were assessed, and statistical differences in protein content were analysed using Kruskal–Wallis and Dunn’s tests in RStudio [2]. Spectral data were processed with Quasar software [3], where average spectra served as references for extended multiplicative signal correction (EMSC)

[4] applied to both NIR and FTIR data. Principal Component Analysis (PCA) was then performed on the corrected spectra, and the PC1 and PC2 loadings were extracted.

3 Results and discussion

3.1. Bibliometric study

The study analysed 18,416 articles to identify trends and emerging topics in protein ingredient research. Using publication and citation analyses, with emphasis on co-occurrence networks, it revealed strong links between plant protein functionality and antioxidant or textural properties, while fewer studies addressed physicochemical characterisation and functionality together..

3.2 Multivariate analysis

Principal Component Analysis was used for data exploration and visualisation. In NIRS, four components explained 95.7% of the variance, with less distinct clustering of isolates than in FTIR. The two potato powders clustered together, while the others were more dispersed, with Pea Concentrate showing high variability. Gluten was clearly separated, indicating complementarity between datasets. Spectral analysis identified characteristic peaks linked to lipids, proteins, sugars and moisture (figures 1a and 1b).

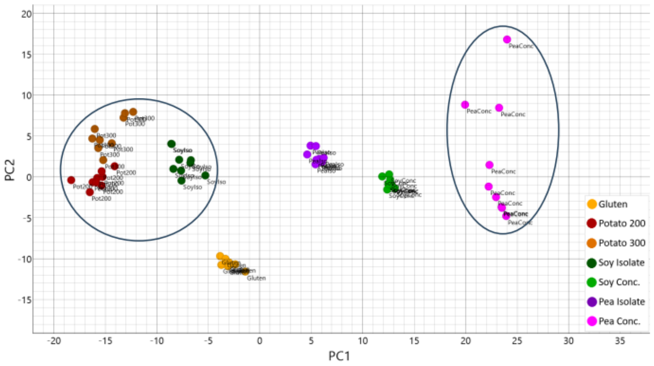


Figure 1a – Scores plot from the NIR Spectra. The components of most importance were PC1 explaining 72.4% of the variance, and PC2 accounting for 12.3% of the total explained variance.

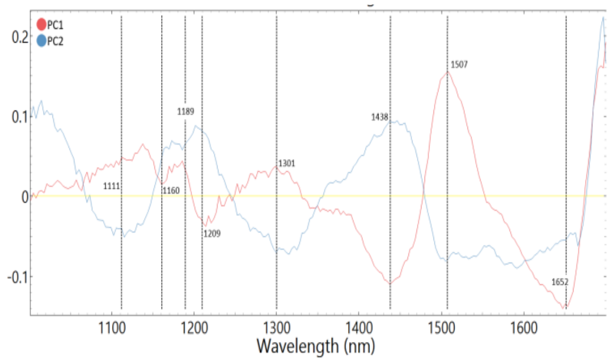


Figure 1b – loadings plot from the NIR Spectra. The red line represents PC1, while the blue line represents PC2. The bands of interests reflect the peaks of interest as interpreted from the pre-processed spectra

4 Conclusion

The results highlight the complexity of linking microscopic characterisation with functional properties, requiring advanced analytical tools. Structuring data in a database combined with chemometrics can help classify proteins and predict their suitability for specific uses. Future work will focus on gel rheology, texturised protein production, and multiblock analysis to integrate DATAVEG data.

5 References

[1] M. Aria, C. Cuccurullo. Bibliometrix: An R-tool for comprehensive science mapping analysis, *Journal of Informetrics* 2017, 11(4), pp. 959-975.

[2] Posit team (2025). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/

[3] J. Demsar et al., Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research* 2013, 2349−2353.

[4] A. Jochemsen et al.. Exploring the use of extended multiplicative scattering correction for near infrared spectra of wood with fungal decay, *Chemometrics and Intelligent Laboratory Systems* 2024, 252, pp.105187.

Improving Model Interpretability in Metabolomics by Assessing Variable Importance Stability via Resampling

J. Boccard[1] S. Rudaz[1]

[1] School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland

Keywords: variable importance, parameter stability, resampling, bootstrap, permutations.

1 Introduction

As a cornerstone of knowledge discovery in metabolomics, multivariate analysis enables the evaluation of relationships between variables, such as measured signals, and observed objects or samples, thereby facilitating the deciphering and deeper understanding of the processes under study. Matrix factorization methods are extensively employed to uncover trends and relevant groupings of observations, but also to highlight potentially related variables based on their contributions to model components. However, the inherent high dimensionality of metabolomic datasets raises questions about the reliability of the coefficients obtained and efficient solutions for this purpose are needed.

2 Material and methods

A novel method is proposed to assess the stability of Variable Importance in Projection from Partial Least Squares regression models, a common criterion widely used in metabolomics to highlight relevant subsets of variables. It combines bootstrap resampling and permutations of subsets of variables to offer an effective and versatile tool based on a stability index and a diagnostic plot. The proposed strategy leverages the full set of variable importance values collected across bootstrap replicates to construct empirical distributions, both authentic and permuted, thereby enhancing the robustness of the assessment. The different steps of the workflow for Variable Importance Stability Assessment (*VISA*) are summarized in Figure 1.

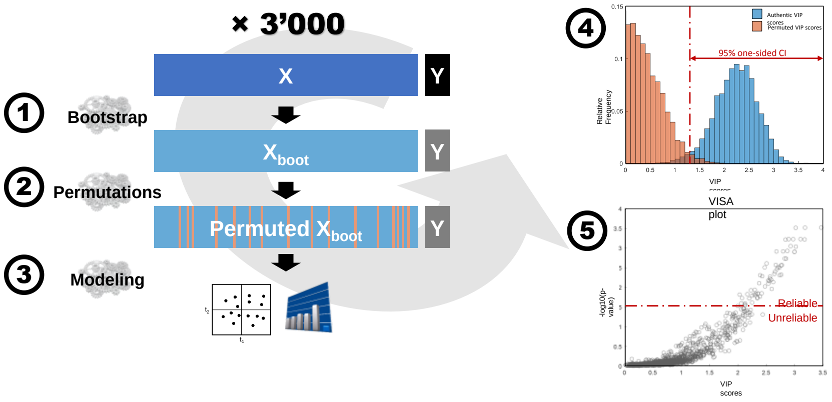


Figure 1. VISA workflow. (1) Bootstrap resampling, (2) Variable permutations, (3) PLS model fitting and VIP scores calculation, (4) VIP scores relative frequency histogram, (5) Diagnostic visualization.

3 Results and discussion

Results from representative real case studies illustrate the potential of the *VISA* method for evaluating the reliability of meaningful variables, and remove uninformative signals in metabolomic datasets resulting from different experimental configurations. A comparison benchmark with established approaches, namely VIP Threshold of One (*VIP>1*) and Bootstrap with One-Variable-at-a-Time Permutation Testing (*OVAT*), highlighted its merits, emphasizing its ability to provide more stable subsets of informative variables, improving the interpretability of metabolomics studies.

Table 1 – Datasets and PLS model parameters and important variables subsets. Important subset size is reported as absolute and relative values (%).

Dataset	Size	Original model parameters				Important subset size			Reference
	(obs x vars)	LVs	R^2X	R^2Y	Q^2Y_{CV}	VIP>1	OVAT	VISA	
SQA1	210 x 210	3	0.202	0.603	0.397	80 (38.1%)	68 (32.4%)	51 (24.3%)	Olesti et al. [1]
SQA2	210 x 210	1	0.063	0.177	0.026	70 (33.3%)	7 (3.3%)	1 (0.5%)	Olesti et al. [1]
CKD	125 x 1’035	3	0.293	0.849	0.722	380 (36.7%)	462 (44.6%)	387 (37.4%)	Gagnebin et al. [2]
AW1	16 x 970	2	0.357	0.988	0.917	375 (38.7%)	260 (26.8%)	177 (18.2%)	Boccard et al. [3]
AW2	16 x 970	2	0.266	0.980	0.785	353 (36.4%)	56 (5.8%)	25 (2.6%)	Boccard et al. [3]

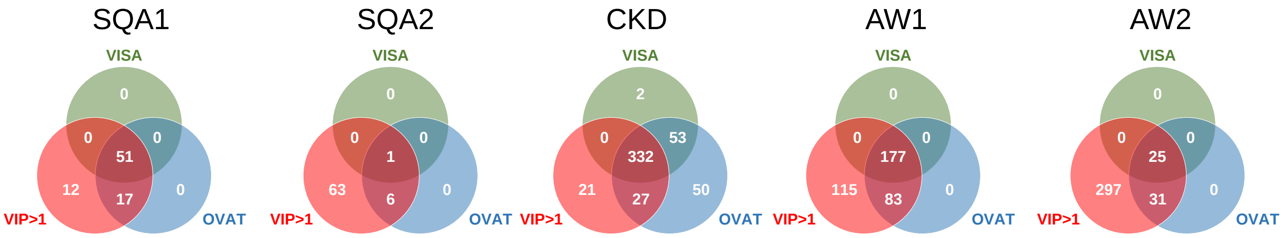


Figure 2. Overlaps of variables subsets considered important.

4 Conclusion

Because it is computationally efficient and does not require assumptions about data distribution, the proposed *VISA* method constitutes a straightforward, generic and relevant approach that is well suited to the needs of a wide range of applications. The broad adoption of this type of methodology will undoubtedly help to achieve more consistent and reproducible results, ultimately advancing the understanding of metabolic patterns.

5 References

[1] Olesti, E., Boccard, J., Rahban, R., Girel, S., Moskaleva, N.E., Zufferey, F., Rossier, M.F., Nef, S., Rudaz, S., González-Ruiz, V. Low-polarity untargeted metabolomic profiling as a tool to gain insight into seminal fluid, *Metabolomics* 19(6), 2023.

[2] Gagnebin, Y., Pezzatti, J., Lescuyer, P., Boccard, J., Ponte, B., Rudaz, S. Toward a better understanding of chronic kidney disease with complementary chromatographic methods hyphenated with mass spectrometry for improved polar metabolome coverage, *J. Chromatogr. B* 1116, 9-18, 2019.

[3] Boccard, J., Kalousis, A., Hilario, M., Lanteri, P., Hanafi, M., Mazerolles, G., Wolfender, J.L., Carrupt, P.A., Rudaz, S. Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in Arabidopsis thaliana, *Chemometr. Intell. Lab. Syst.* 104(1) 20-27, 2010.

An overview of LC-MS, GC-MS and NMR metabolomic data from Leucojum aestivum bulbs (Amaryllidaceae)

Rosella Spina * [1]

[1] Laboratoire Agronomie et Environnement, UMR 1121 – Université de Lorraine, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

*Speaker

Contribution of polarized Raman spectroscopy imaging and chemometrics for the analysis of residual scratches on polyethylene terephthalate surfaces.

M. Haouchine[1,2], M. Pecora[*2], S. Hupont[1], M. Ponçot[1], C. Gauthier[2] and I. Royaud[1]

[1] Université de Lorraine, CNRS, Institut Jean Lamour, F-54000 Nancy, France. *marina.pecora@ics-cnrs.unistra.fr

[2] Université de Strasbourg, CNRS, Institut Charles Sadron, F-67000 Strasbourg, France.

Keywords: Raman Spectroscopy, Polarization, Chemometrics, Scratch, Amorphous polymers

1 Introduction

Polymers are widely used due to their versatility but are often subjected in service to mechanical and environmental stresses that can lead to surface damage such as scratching. Scratch-induced degradation affects both functional and aesthetic properties and involves complex deformation mechanisms depending on polymer microstructure and loading conditions [1]. Understanding these mechanisms at the molecular scale is therefore essential. Confocal Raman spectroscopy is a powerful technique for probing local molecular orientation and structural changes induced by mechanical loading. However, only a limited number of studies have applied Raman spectroscopy directly to polymer tribology, mostly focusing on chemical or post-mortem effects rather than deformation mechanisms [2-3]. Multivariate chemometric methods, such as principal component analysis (PCA) and multivariate curve resolution (MCR), offer effective tools to analyze complex Raman datasets and extract structural information [4]. In this context, the present study aims to characterize scratch-induced deformations in amorphous Polyethylene terephthalate (PET) using polarized confocal Raman microspectroscopy combined with chemometric workflow.

2 Material and methods

An amorphous commercial PET plate was machined into samples for scratch testing. Scratches were performed under ambient conditions using the Micro Visio Scratch device [5]. Test parameters were selected to generate a fully plastic, permanent scratch without brittle damage, using a spherico-conical tip (R=243µm) under a normal load of 5 N, corresponding to a representative strain of ~12% [6]. Post-mortem analysis was carried out using confocal Raman microspectroscopy equipped with a 785 nm laser (14.5 mW). Raman hyperspectral maps were acquired using two spatial resolutions with ×10 and ×100 objectives to probe both global and local deformation features. Six polarization configurations were employed to assess molecular orientation effects. The resulting hyperspectral datasets were unfolded, preprocessed (baseline correction, smoothing, and normalization), analyzed using PCA and MCR-ALS, and subsequently refolded to extract spatially resolved structural information.

3 Results and discussion

PCA and MCR-ALS analysis of polarized Raman data reveals clear molecular-scale signatures of scratch-induced deformation in amorphous PET. Low-resolution experiments have shown that the

Vertical-Horizontal (VH) configuration offers the highest sensitivity. High-resolution MCR score maps show a strong spatial heterogeneity of deformation, with the most pronounced structural changes localized at the scratch edges, while the groove center exhibits intermediate behavior (Figure 1a and 1b). MCR loadings highlights significant intensity variations of several Raman bands between scratched and unscratched regions, without detectable band shifts (Figure 1c), indicating that the deformation is dominated by molecular orientation. Given the testing conditions, strain-induced crystallization is unlikely. On the contrary, these spectral features suggest the formation of a nematic mesophase under high local plastic strain, as reported for PET under uniaxial deformation [7]. Overall, scratching acts as a localized orientational field, inducing molecular rearrangements analogous to tensile stretching but confined to the near-surface region.

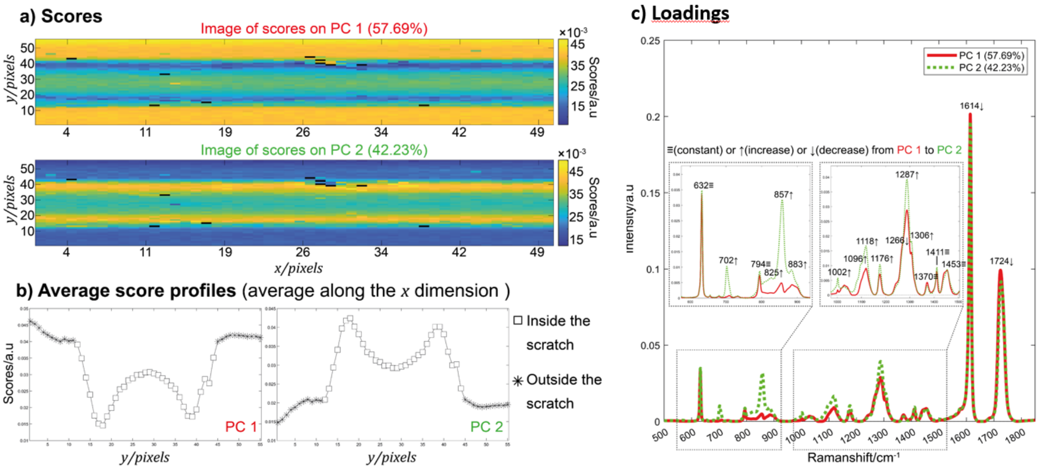


Figure 1 – (a) Image of MCR scores on PC1 (top) and PC2 (bottom). (b) Average along x/pixels for PC1 and PC2 score profiles as a function of y/pixels dimension. (c) PC1 (red) and PC2 (green) MCR loadings.

4 Conclusion

This study demonstrates that polarized confocal Raman microspectroscopy combined with chemometric analysis is a powerful approach to probe scratch-induced molecular deformation in amorphous PET. Polarization was shown to be essential for detecting structural changes, with the VH configuration providing the highest sensitivity. High-resolution mapping revealed a heterogeneous deformation field, with maximum macromolecular reorganization at the scratch edges. The absence of Raman band shifts, together with pronounced band intensity variations, indicates that deformation is governed by chain reorientation and the formation of an orientation-induced mesophase.

5 References

[1] Pelletier H, Durier AL, Gauthier C, Schirrer R. Viscoelastic and elastic–plastic behaviors of amorphous polymeric surfaces during scratch. Tribol Int. 2008 Nov;41(11):975–84.

[2] Briscoe BJ, Stuart BH, Rostami S. A Fourier transform Raman spectroscopy study of the crystallization behaviour of poly (ether ether ketone)/poly (ether imide) blends. Spectrochim Acta Part Mol Spectrosc. 1993 May;49(5–6):753–8.

[3] Stuart BH. The application of Raman spectroscopy to the tribology of polymers. Tribol Int. 1998 Nov;31(11):687–93.

[4] De Juan A, Tauler R. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review. Anal Chim Acta. 2021 Feb;1145:59–78.

[5] Gauthier C, Lafaye S, Schirrer R. Elastic recovery of a scratch in a polymeric surface: experiments and analysis. Tribol Int. 2001 July;34(7):469–79.

[6] Tabor D. The hardness of metals. New York: Oxford University Press; 2000. 175 p. (Oxford classic texts in the physical sciences).

[7] Forestier E, Combeaud C, Guigo N, Sbirrazzuoli N, Billon N. Understanding of strain-induced crystallization developments scenarios for polyesters: Comparison of poly(ethylene furanoate), PEF, and poly(ethylene terephthalate), PET. Polymer. 2020 Aug;203:122755.

APPORT DE LA CHIMIOMETRIE DANS LE SUIVI EN LIGNE DU POTENTIEL OXYDANT DES AEROSOLS

M. Haouchine[1] L. Ndouta [1] D. Rousset[1]

[1] Département Métrologie des Polluants, Institut National de Recherche et de Sécurité (INRS), 54500 Vandœuvre-lès-Nancy, France

Mots clés : Potentiel oxydant, aérosols, mesure en ligne, chimiométrie, ACP, MCR-ALS

1 Introduction

La directive (UE) 2024/2881 introduit le potentiel oxydant (PO) des aérosols (PM) comme nouvel indicateur du suivi de la qualité de l'air, reflétant mieux leurs effets sanitaires[1]. Sa mesure permet en effet d'intégrer simultanément l'influence des propriétés physico-chimiques des PM, ainsi que les effets agonistes ou antagonistes susceptibles d'émerger entre leurs différents constituants[2]. Parmi les méthodes acellulaires de mesure du PO[3], l'essai au 1,4-dithiothréitol (DTT) repose sur le suivi spectroscopique en UV-Visible de son oxydation catalysée par certaines espèces présentes dans les PM. Actuellement, majoritairement réalisée hors ligne, cette mesure peut entraîner des pertes de réactivité, soulignant la nécessité de développer des techniques de mesure en ligne[4].

La mesure du PO est indirecte et le plus souvent univariée, basée sur la formation du 2-nitro-5-thiobenzoate (TNB^-) mesuré à 412 nm. L'extension de la détection à l'ensemble du spectre UV-Visible, combinée à des outils de chimiométrie, permettrait toutefois d'exploiter l'ensemble des signatures spectrales, de gérer la quantité importante des données, de mieux discriminer les contributions des espèces et d'améliorer la robustesse de la mesure. Cette communication présente une preuve de concept de l'apport des approches multivariées, via l'analyse en composantes principales (ACP)[5] et la résolution de courbes multivariées (MCR)[6], illustrée par un jeu de données à chimie relativement simple (catalyse de l'oxydation du DTT par les ions Cu^{2+}),

2 Matériels et méthodes

Six solutions de $CuSO_4$ à 0,05, 0,1, 0,2, 0,5, 1 et 10 µM ont été préparées et utilisées pour catalyser l'oxydation du DTT. Le DTT et la DTNB ont été injectés aux concentrations de 20 µM et 150 µM, respectivement. Le prototype utilisé, décrit précédemment[7], repose sur une stratégie de mesure en « end point ». Ce compromis permet un échantillonnage de l'air et une analyse en ligne continus. Les données brutes ont d'abord été prétraitées afin de réduire le bruit instrumental par lissage de Savitzky-Golay et de corriger les effets de diffusion de la lumière par soustraction de l'absorbance à 600 nm. Un modèle ACP a ensuite été construit et utilisé pour le calcul de deux distances multivariées, la distance d'Hotelling (T^2) et les Q-résidus, en vue de la détection des valeurs aberrantes. Après correction et nettoyage des données, un modèle MCR a été calculé avec une optimisation par moindres carrés alternés (ALS), sous contrainte de non-négativité sur les deux dimensions. Le nombre de composantes pures a été fixé à deux sur la base de la connaissance du système chimique.

3 Résultats et discussion

L’exploitation univariée des données (Figure 1A) met en évidence l’évolution du signal en réponse aux injections de $CuSO_4$. Toutefois, la présence de nombreux points aberrants liés aux contraintes des mesures en ligne souligne les limites de cette approche. L’ACP a permis d’identifier et d’exclure les spectres aberrants à partir des distances T^2 et des Q-résidus, conduisant à un jeu de données filtré. Les scores des deux premières composantes principales d’un nouveau modèle ACP, construit sur ce jeu de données filtré, ont mis en évidence la présence de deux groupes d’individus attribuables aux deux espèces absorbantes du système, la DTNB et la TNB. À partir de ces données filtrées, la modélisation MCR-ALS a permis de reconstruire deux composantes pures, associées respectivement à la TNB et à la DTNB et d’obtenir leurs profils d’évolution temporelle (Figure 1B) et leurs signatures spectrales pures (Figure 1B’). Les profils cinétiques reconstruits ne présentent plus d’anomalies liées aux perturbations expérimentales observées dans l’analyse univariée (ex. absence de variation abrupte à la fin de la 2ème injection). Cette stabilité améliore la robustesse des résultats de mesure. En effet, en l’absence de valeurs aberrantes, si l’on moyennait chaque plateau (blancs ou échantillons), le coefficient de variation serait plus faible, car les valeurs extrêmes ne viennent plus altérer la moyenne et l’écart-type. Ainsi, les profils cinétiques reconstruits par MCR-ALS offrirait une estimation plus fiable du PO.

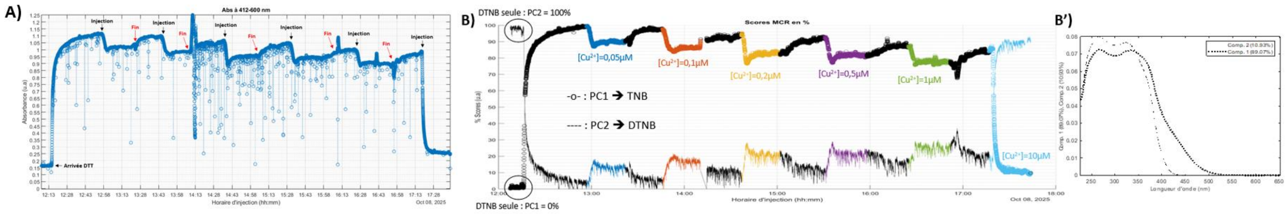


Figure 1 – Résultats de l’analyse univariée (A) et de la modélisation MCR-ALS (B et B’).

4 Conclusion

Cette étude met en évidence les limites de l’exploitation univariée des mesures de PO, particulièrement sensible aux perturbations inhérentes aux dispositifs de mesure en ligne. L’approche multivariée, combinant ACP et MCR permet un filtrage efficace des valeurs aberrantes et une séparation robuste des contributions spectrales de la DTNB et de la TNB. En perspective, l’automatisation de l’interprétation des profils MCR par calcul des dérivées ainsi qu’une réflexion sur le calcul du PO à partir des absorbances mesurées en « end point » sont en cours.

5 Références

[1] W. J. Gauderman. *et al.*, “Association between Air Pollution and Lung Function Growth in Southern California Children,” *Am. J. Respir. Crit. Care Med.*, vol. 162, no. 4, pp. 1383–1390, Oct. 2000, doi: 10.1164/ajrccm.162.4.9909096.

[2] P. S. J. Lakey *et al.*, “Chemical exposure-response relationship between air pollutants and reactive oxygen species in the human respiratory tract,” *Sci. Rep.*, vol. 6, no. 1, p. 32916, Sept. 2016, doi: 10.1038/srep32916.

[3] J. T. Bates *et al.*, “Review of Acellular Assays of Ambient Particulate Matter Oxidative Potential: Methods and Relationships with Composition, Sources, and Health Effects,” *Environ. Sci. Technol.*, vol. 53, no. 8, pp. 4003–4019, Apr. 2019, doi: 10.1021/acs.est.8b03430.

[4] M. Ghanem, L. Y. Alleman, D. Rousset, E. Perdrix, and P. Coddeville, “Experimental factors influencing the bioaccessibility and the oxidative potential of transition metals from welding fumes,” *Environ. Sci. Process. Impacts*, vol. 26, no. 5, pp. 843–857, May 2024, doi: 10.1039/D3EM00546A.

[5] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933, doi: 10.1037/h0071325.

[6] R. Tauler, A. Smilde, and B. Kowalski, “Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution,” *J. Chemom.*, vol. 9, no. 1, pp. 31–58, 1995, doi: 10.1002/cem.1180090105.

[7] T. Audoux, M. Ghanem, J-B. Lily, E. Perdrix, L. Y. Alleman, and D. Rousset, “Evaluation en laboratoire d’un dispositif de mesure en ligne du potentiel oxydant des aérosols.,” 2025, doi: 10.25576/ASFERA-CFA2025-43912.

STUDY OF THE BEHAVIOR OF REACTIVE FLUIDS APPLIED TO THE CASE OF THERMODYNAMIC CYCLES

Characterization of formic acid by Raman spectroscopy

Philippe ARNOUX[1], Sarena LOULHA[1], Silvia LASALA[2], Olivier HERBINET[1], Marc OFFROY[2]
[1] Université de Lorraine, CNRS, LRGP, F-54000 Nancy, France, philippe.arnoux@univ-lorraine.fr
[2] Université de Lorraine, CNRS, LIEC, F-54000 Nancy, France, marc.offroy@univ-lorraine.fr

Keywords: reactive fluids; formic acid; Raman spectroscopy; chemometrics

1 Introduction

Reactive fluids introduced in thermodynamic cycles have a great potential to intensify energy processes. This is explored by the REACHER ERC project. In the framework of this project, the characterization of formic acid as a reactive fluid has been studied by Raman spectroscopy.

The Raman spectra of formic acid in gas phase were obtained between 70 °C and 160 °C. Firstly, a qualitative analysis based on the literature as well as the results obtained by the Gaussian software allowed to highlight the equilibrium between the monomer and the dimer which guarantees the reactivity of the fluid. Secondly, different methods of spectra processing were applied to improve the quality of the spectra for a quantitative analysis. The sensitivity of the correction parameters was evaluated by statistical analysis. The evolution of the intensity according to different parameters was also studied.

2 Theory

The characterization of the fluid under study is based on Raman spectroscopy. Raman Spectroscopy is a non-destructive chemical analysis technique which provides detailed information about the chemical structure of a fluid. This is a non-destructive method for characterizing the molecular composition and structure of a fluid. This method relies on the Raman effect, a physical phenomenon of inelastic scattering of light interacting with the medium. This frequency shift corresponds to an exchange of energy between the incident and scattering laser photons.

3 Material and methods

Gas-phase measurements were performed using a KAISER RXN2 Raman spectrometer. The laser wavelength used was 532 nm. The analyzed species was 99% pure formic acid supplied by Carlo Erba Reagent. The fluid was introduced in liquid form into a stainless steel reactor. The reactor has three openings through which the Raman probe passes, along with a thermocouple for temperature monitoring and a valve connected to a vane pump.

4 Results and discussion

Globally, 10 gas-phase Raman spectra were acquired by varying the temperature between 70 °C and 160 °C. The spectra of the crude derivative and the derivative corrected by the Savitzky-Golay algorithm were calculated. This analysis of spectra consists of identifying the peaks visible on the obtained spectra and associating them with a vibration linked to a bond of the formic acid molecule, and possibly determining their belonging to the dimer or the monomer. We were thus able to identify 13 peaks based on the shape of the derivative. The derivative allowed us to identify the superposition of the first two peaks in this case.

In addition to the noise treated by the Savitzky-Golay algorithm, the obtained spectra exhibit a baseline deviation related to fluorescence. The "Detrend" baseline correction was applied using a second-order polynomial. Following this baseline correction, two isobestic points were identified: the first at approximately 1685 cm^{-1} on the C=O elongation band, the second at approximately 3310 cm^{-1} after the O-H elongation band. The increase in temperature corresponds to a shift in the equilibrium from the dimer to the monomer.

5 Conclusion

The qualitative analysis conducted during the REACHER research project revealed the presence of formic acid monomer and dimer in the gas phase, through the assignment of peaks corresponding to the Raman spectra, as well as through the presence of isobestic points. The quality of the spectra was improved using chemometric methods. These methods can complement signal processing performed with the Savitzky-Golay algorithm or baseline correction.

References

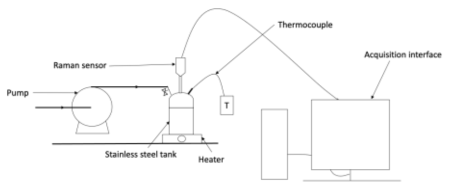
[1] Olbert-Majkut, Adriana, et al. « Raman Spectroscopy of Formic Acid and Its Dimers Isolated in Low Temperature Argon Matrices ». Chemical Physics Letters, vol. 468, nº 4‑ 6, janvier 2009, p. 176‑ 83. https://doi.org/10.1016/j.cplett.2008.12.011

[2] Bertie, John E., et Kirk H. Michaelian. « The Raman Spectra of Gaseous Formic Acid - h 2 and - d 2 ». The Journal of Chemical Physics, vol. 76, nº 2, janvier 1982, p. 886‑ 94. https://doi.org/10.1063/1.443061

[3] D.O. Wasik et al., Multiscale modeling of dimerization thermodynamics of formic acid, Fluid Phase Equilibria 594 (2025) 114356

[4] Lasala et al., Application of thermodynamics at different scales to describe the behaviour of fast reacting binary mixtures in vapour-liquid equilibrium, Chemical Engineering Journal Volume 483, 1 March 2024, 148961

Acknowledgments: This work has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation program (grant agreement No 101040994).



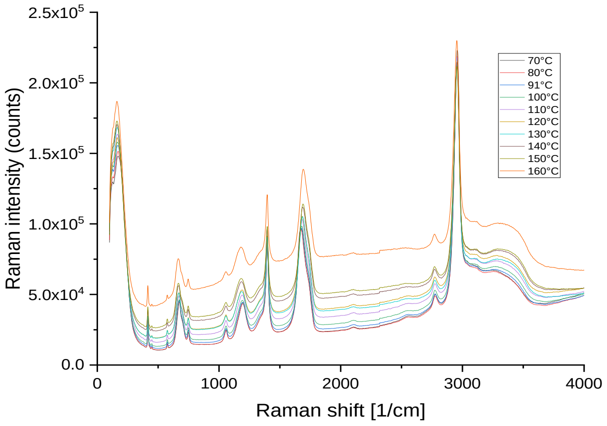


Fig. 1 – Experimental set-up

Fig. 1 – Raman spectrum of formic acid between 70°C and 160°C

CHIMIOMÉTRIE XXVème
ENSIC LIEC SFdS UNIVERSITÉ DE LORRAINE cnrs
17-19 FÉVRIER 2026
CAMPUS GRANVILLE, NANCY

Optimization of Black Elastomers Classification by 2D Fluorescence Spectroscopy and Variable Selection for Industrial Recycling

N. Caillol[1-2], J. Peyrelon-Braud[1], J. Martini[1] , M. Rey-Bayle[1],
[1] IFPEN, rond-point de l'échangeur 69390 Solaize
[2] Axel'One, rond-point de l'échangeur 69390 Solaize

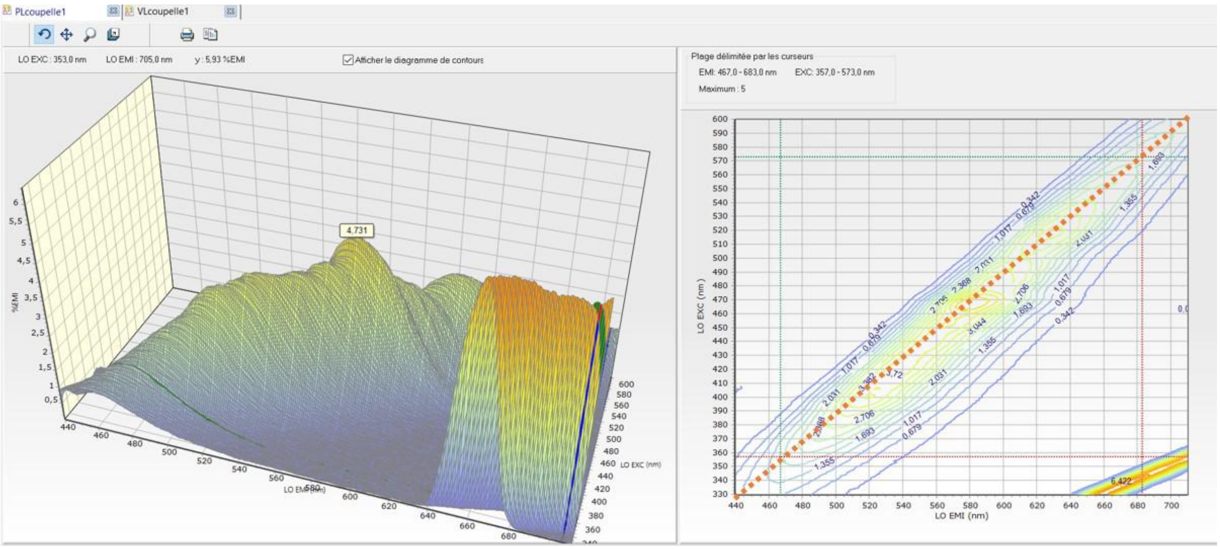
Keywords: 2D Fluorescence, ML, Machine Learning, Successive Projections Algorithm (SPA), Classification, Tire Recycling, Variable Selection.

1 Introduction

In an industrial context where tire recycling represents a major environmental and economic challenge, this study focuses on developing automatic classification methods to differentiate black elastomers derived from heavy-duty (HD) and light-duty (LD) vehicle tire granulates. The main objective is to optimize the collection and analysis of spectral data obtained through 2D fluorescence spectroscopy to improve discrimination between these two classes of black elastomer while significantly reducing acquisition time.

2 Material and methods

A Safas-Xenius spectrometer was used for acquiring the 2D fluorescence signal.



PLcoupelle1 VLcoupelle1

LO EXC : 353,0 nm LO EMI : 705,0 nm y : 5,93 %EMI

☑ Afficher le diagramme de contours

Plage délimitée par les curseurs
EMI: 467,0 - 683,0 nm EXC: 357,0 - 573,0 nm
Maximum : 5

Figure 1 – example of a tire sample's 2D-exitation/emission fluorescence signal. Plotted as a 3D-plot and a contour plot + dotted red line correspond to a synchronous spectrum

Different approaches were explored to achieve this goal through both signal acquisition strategies and mathematical variable selection tools:

1. Complete acquisition of excitation-emission matrices (EEM), providing exhaustive spectral mapping but requiring a high acquisition time of 22 minutes per sample.
2. Use of synchronous spectra, a simplified method reducing acquisition time to just 1 minute while maintaining good spectral information quality.
3. Manual variable selection, visually identifying the most discriminative excitation/emission wavelengths, but with limitations in terms of optimization.
4. Application of the Successive Projections Algorithm (SPA)[1], enabling automatic and efficient selection of discriminative variables, with acquisition time reduced to one second per sample.

3 Results and Conclusions

The spectral data from these approaches were processed using two classification algorithms: Partial Least Squares Discriminant Analysis (PLS-DA) and Support Vector Machines (SVM). The results show that all methods, except univariate selection (Welch's test), achieved perfect classification on this small number of classes (100% accuracy). However, the SPA method stood out for its robustness and ability to enhance class separation, with superior performance compared to manual selection:

- SVM Margin: 0.70 (SPA) versus 0.44 (manual).
- Mahalanobis Distance: 10.5 (SPA) versus 7.2 (manual).

These results validate the use of spectrofluorimetry combined with the SPA algorithm as a fast, robust, and high-performing method for characterizing and discriminating black elastomers. This work paves the way for promising industrial applications, particularly the automation of sorting on recycling lines.

Future perspectives include an in-depth chemical analysis to identify spectral signatures specific to tire formulations and further development of SPA expertise to extend its application to more complex multi-class problems.

4 References

[1] Mário César Ugulino Araújo et al. “Successive Projections Algorithm for Variable Selection in Spectrofluorescence Data”. Analytica Chimica Acta 455 (2001), p. 1-12.

Multiblock Analysis of Metabolomics Data: Application to Mindfulness Meditation and Healthcare Student Well-Being

Célie Da Silva[1,2], Mathieu Galmiche[1,2], Oriane Strassel[1,2], Sergey Girel[1,2], Isabel Meister[1,2], Camille Piguet[3,4], Françoise Jermann[5], Julien Boccard[1,2], Serge Rudaz[1,2]

[1]School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland

[2] Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland

[3]Psychiatry Department, Faculty of Medicine, University of Geneva, Switzerland

[4]General Pediatric Division, Geneva University Hospital, Switzerland

[5]Department of Psychiatry, Geneva University Hospitals, Geneva, Switzerland

Celie.Dasilva@unige.ch

Keywords: chemometrics, multiblock, metabolomics, mindfulness

1 Introduction

Multiblock approaches are increasingly applied to integrate heterogeneous molecular and phenotypic data, and to capture system-level relationships. These methods are also very useful in longitudinal intervention studies, where they enable the assessment of sustained biological changes over time. In the context of mindfulness-based interventions (MBIs) like Mindfulness-Based Cognitive Therapy for Life (MBCT-L), multiblock analysis provides a relevant way to link psychological outcomes with molecular responses, addressing the need for effective stress-reduction strategies in healthcare student[1,2].

2 Material and methods

We applied supervised chemometric approaches within a multiblock framework to integrate longitudinal metabolomic and lipidomic data from healthcare students across three time points. Data blocks were generated from blood samples using LC-MS, and features were annotated at Level 1 (confirmed) and Level 2 (putatively annotated). The resulting datasets included 1'156 lipids across 194 samples and 494 metabolites across 197 samples, covering all time points for each participant. This structure allows for the joint analysis of repeated measurements, with individuals shared across data blocks. Following preprocessing to remove non-informative variables and apply appropriate variable transformations, psychological questionnaire scores were incorporated into the integrated analysis. This approach enabled us to deconvolute molecular signatures associated with the MBCT-L intervention and to link observed molecular modulations to psychological outcomes.

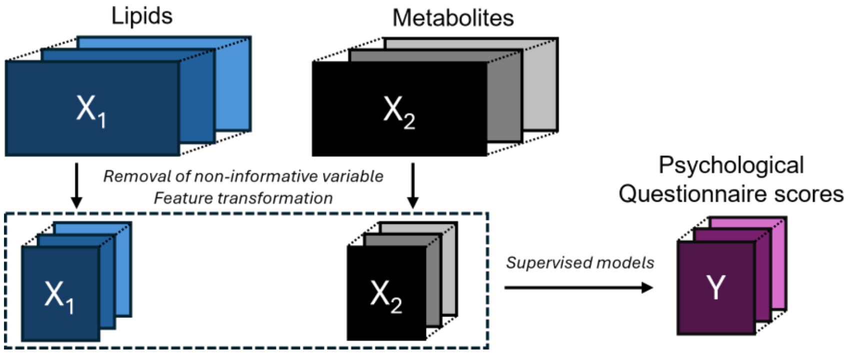


Figure 1 – Overview of Multiblock Chemometric of Lipidomic, Metabolomic, and Psychological Data

3 Conclusion

The preliminary results highlight the value of multiblock integration for generating biologically grounded hypotheses on MBIs. The joint interpretation of molecular and psychological data opens new perspectives for understanding the mechanisms underlying mindfulness practice and its impact on well-being in the studied population.

4 References

[1] Strauss, C. et al. Reducing stress and promoting well-being in healthcare workers using mindfulness-based cognitive therapy for life. Int. J. Clin. Health Psychol. IJCHP 21, 100227 (2021)

[2] Di Mario, S., Rollo, E., Gabellini, S. & Filomeno, L. How Stress and Burnout Impact the Quality of Life Amongst Healthcare Students: An Integrative Review of the Literature. Teach. Learn. Nurs. 19, 315–323 (2024).

A two-step strategy based on ATR-FTIR fingerprinting in tandem with chemometric tools for the authentication and quantification of artificial honey adulteration in premium Moroccan honeys

A. En-Najy[1] M. Kharbach[2] Y. Vander Heyden[3] A. Bouklouze[1]

[1] Biopharmaceutical and Toxicological Analysis Research Team, Laboratory of Pharmacology and Toxicology, Unit of Instrumental Analysis and Data Handling, Faculty of Medicine and Pharmacy, Mohammed V University in Rabat, 10100 Rabat, Morocco, anas.ennajy@um5r.ac.ma, a.bouklouze@um5r.ac.ma

[2] Circular Economy/Sustainable Solutions, LAB University of Applied Sciences, Mukkulankatu 19, 15101 Lahti, Finland, mourad.kharbach@lab.fi

[3] Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, 1090 Brussels, Belgium, yvan.vander.heyden@vub.be

Keywords: FTIR Spectroscopy, Chemometric analysis, Authentication, Adulteration, Moroccan honey, Artificial honey.

1 Introduction

Honey authentication faces increasing challenges due to the widespread global problem of adulteration [1]. Honey is a natural sweet substance produced mainly from floral nectar [2], yet it is frequently mixed with low-cost sugar syrups of similar composition, misleading consumers and compromising market integrity [3].

Although conventional techniques such as elemental analysis isotope ratio mass spectrometry [4], gas chromatography [5], high-performance liquid chromatography [6], and nuclear magnetic resonance spectroscopy [7] can detect adulteration, they remain costly and time-consuming. Vibrational spectroscopic techniques, including infrared and Raman spectroscopy, offer faster and simpler alternatives. Among them, FTIR combined with chemometric analysis provides rapid, non-destructive fingerprinting and has shown strong potential for honey authentication.

Given their high market value and vulnerability to fraud, applying FTIR-chemometrics is particularly relevant for ensuring the authenticity of high-priced Moroccan monofloral honeys.

2 Material and methods

A total of 16 Moroccan monofloral honey samples, representing four high-priced Moroccan floral origins, Daghmous, Thyme, Jujube, and Zakoum, were collected, with four samples per botanical type. Each honey was adulterated with artificial honey at five levels (5%, 10%, 20%, 40%, 50% w/w). FTIR spectra were recorded for all pure and adulterated samples. Spectral datasets (merged and unmerged) were preprocessed and analysed using Principal Component Analysis (PCA), Partial Least Squares–Discriminant Analysis (PLS-DA), and Partial Least Squares Regression (PLSR). Thus, all prepared samples and three replicate measurements resulted in two hundred and ninetyone spectra ($16 \times 5 \times 3 + 16 \times 3 + 3 = 291$), being two hundred and forty spectra from adulterated samples, forty-eight spectra from pure honey samples, and three from the adulterant.

3 Results and discussion

PCA successfully separated pure honey samples from adulterated ones, revealing well-defined clusters for each botanical origin. PLS-DA models achieved excellent classification performance, showing high sensitivity and specificity across all adulteration levels (Figure 1). PLSR yielded strong predictive ability, with coefficients of determination (R^2) exceeding 0.98 and error metrics (RMSEC, RMSECV, RMSEP) all below 2% for both merged and unmerged datasets.

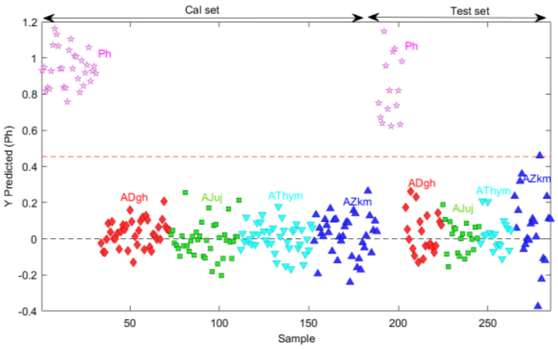


Figure 1 – PLS-DA classification results for the merged dataset, showing pure honey vs all other samples.

4 Conclusion

FTIR spectroscopy combined with chemometric analysis provides a reliable, fast, and non-destructive tool for detecting and quantifying adulteration in high-value Moroccan honeys. The proposed strategy involves first applying a global model to the merged dataset to identify adulteration, followed by individual models for precise quantification. These findings highlight the strong potential of FTIR-based chemometrics for routine quality control and market protection of Moroccan honey.

5 References

[1] X.-H. Zhang, H.-W. Gu, R.-J. Liu, X.-D. Qing, and J.-F. Nie, “A comprehensive review of the current trends and recent advancements on the authenticity of honey,” Food Chem X, vol. 19, p. 100850, Oct. 2023, doi: 10.1016/j.fochx.2023.100850.

[2] C. Kumaravelu and A. Gopal, “Detection and Quantification of Adulteration in Honey through Near Infrared Spectroscopy,” Int J Food Prop, vol. 18, no. 9, pp. 1930–1935, Sep. 2015, doi: 10.1080/10942912.2014.919320.

[3] A. Naila, S. H. Flint, A. Z. Sulaiman, A. Ajit, and Z. Weeds, “Classical and novel approaches to the analysis of honey and detection of adulterants,” Food Control, vol. 90, pp. 152–165, Aug. 2018, doi: 10.1016/j.foodcont.2018.02.027.

[4] L. Elflein and K.-P. Raezke, “Improved detection of honey adulteration by measuring differences between 13 C/ 12 C stable carbon isotope ratios of protein and sugar compounds with a combination of elemental analyzer - isotope ratio mass spectrometry and liquid chromatography - isotope ratio mass spectrometry (δ 13 C - EA/LC-IRMS),” Apidologie, vol. 39, no. 5, pp. 574–587, Sep. 2008, doi: 10.1051/apido:2008042.

[5] I. K. Karabagias, A. Badeka, and M. G. Kontominas, “A decisive strategy for monofloral honey authentication using analysis of volatile compounds and pattern recognition techniques,” Microchemical Journal, vol. 152, p. 104263, Jan. 2020, doi: 10.1016/j.microc.2019.104263.

[6] C. Egido, J. Saurina, S. Sentellas, and O. Núñez, “Honey fraud detection based on sugar syrup adulterations by HPLC-UV fingerprinting and chemometrics,” Food Chem, vol. 436, p. 137758, Mar. 2024, doi: 10.1016/J.FOODCHEM.2023.137758.

[7] C. He, Y. Liu, H. Liu, X. Zheng, G. Shen, and J. Feng, “Compositional identification and authentication of Chinese honeys by 1H NMR combined with multivariate analysis,” Food Research International, vol. 130, p. 108936, Apr. 2020, doi: 10.1016/j.foodres.2019.108936.

Use of Near Infrared Spectroscopy to predict chemical composition of excreta and nutriment digestibility in broiler chicken

N. Fumat[1], M. Traineau[2], M. Faure[1], M. Pareux[2], M. Vilariño[2]

[1]ARVALIS, Route de Malesherbes, 91720 Boigneville, France, n.fumat@arvalis.fr

[2] ARVALIS, 2 Pouline, 41100 Villerable, France, m.traineau@arvalis.fr

Keywords: NIRS, digestibility, broiler chicken, excreta, animal feed.

1 Introduction

At the ARVALIS experimental station, near-infrared spectroscopy (NIRS) coupled with chemometrics methods is routinely employed to characterize the chemical composition of chicken and rooster excreta, including starch, nitrogen, and gross energy contents. NIRS is also used to directly predict nutrient digestibility in broilers. Traditionally, both types of analyses have been performed on dried samples (freeze-dried excreta). In 2025, ARVALIS developed new prediction equations for estimating digestibility based on fresh excreta (without drying or grinding) using a portable NIRS device. This approach enables faster analyses and, in the future, could be applied directly to floor-reared broilers chicken.

2 Theory

Even today, the reference method for acquiring this data remains a digestive assessment over several consecutive days, monitoring consumption (with fasting) and excretion in chickens or roosters [1, 2]. NIRS has been used for several decades as an alternative method for predicting the chemical composition of nutrients and excreta [3]. Recent work (particularly at ARVALIS) has established relationships between infrared spectra and nutrient digestibility (AMEn digestibility) (kcal/kg DM), digestible energy (kcal/kg DM), nitrogen (%), starch (%)) in freeze-dried and crushed samples by combining infrared spectral information from feed with that from droppings [4, 5]. The miniaturization of NIRS has opened up new possibilities in this area, such as the direct measurement of fresh excreta.

3 Material and methods

Unlike the chemical composition, for which the spectral excitation bands are known, predicting digestibility based on freeze-dried and crushed excreta is an indirect measurement process. The chosen NIR instrument was a spectrometer with a wide spectral range (400–2,500 nm) and good resolution (2 nm): the NIRSystem XDS (FOSS, DNk). For analysing fresh excreta, we use the NIR instrument MicroNIR (VIAVI, USA). This miniature spectrometer has an infrared range of 900–1,700 nm with a resolution of 6 nm. Each spectrum is obtained from the average of six to eight measurements taken across the entire surface to account for the heterogeneity of the faeces. NIR digestibility models are created by combining the spectra of feed and faeces. Models for the chemical composition of freeze-dried excreta on the XDS are used routinely. These models have been developed for several chicken species, including fast-growing ROSS chickens aged 3–35 days, slow-growing JA chickens, and Isa Brown roosters. Currently, analyses of digestibility in freeze-dried and fresh excreta are only carried out on adult ROSS chickens. However, the database of fresh excreta is smaller, containing 513 samples compared to over 3,000 for freeze-dried faeces.

4 Results and discussion

Table 1 summarizes the performance of the models for the different digestibilities on freeze-dried (XDS) and fresh (MicroNIR) excreta. The database was divided into calibration and test sets.

Table 1 – Performance of chemical composition models for freeze-dried excreta and digestibility.

	Digestibility Freeze-Dried Excreta (XDS)				Digestibility Fresh Excreta (MicroNIR)			
	Starch	Nitrogen	Energy	AMEn	Starch	Nitrogen	Energy	AMEn
Min - Max	62 - 100	56 - 94	46 - 83	2268 - 3842	85 - 100	76 - 89	60 - 76	2762 - 3468
RMSEC	0.5	1.0	1.5	76	1.0	1.2	1.1	48
R² calibration	0.97	1.00	0.87	0.84	0.78	0.99	0.99	0.99
RMSEP	0.7	1.3	1.7	90	1.0	1.2	1.1	50
Biais	0.01	0.00	0.06	-0.70	-0.02	0.11	0.15	-0.07
Pente	0.96	0.99	0.95	0.97	1.05	1.01	1.00	1.09
R² validation	0.95	0.90	0.85	0.80	0.74	0.81	0.85	0.84

The models' performance is similar or even better for two digestibility criteria with fresh excreta due to a lack of variability in the database. These results, obtained using a miniature spectrometer on a less homogeneous matrix than freeze-dried excreta and containing more water, are encouraging for the future. The next step is to expand the database of fresh excreta while maintaining similar levels of error and performance to those achieved with freeze-dried excreta and a higher-resolution laboratory spectrometer.

5 Conclusion

Using NIRS on freeze-dried faeces enables the rapid and reliable estimation of feed composition and digestibility by broiler chickens. The innovation of using fresh excreta with a miniature NIRS instrument (currently under development) will enable the future estimation of digestibility in broiler chickens on litter under farming conditions.

6 References

[1] Bourdillon A, (a), Carré B., Conan L., Francesch M., Fuentes M. European reference method of in vivo determination of metabolisable energy in poultry: Reproducibility, effect of age, comparison with predicted values. *British Poultry Science, 1990,* volume 31, 567-576, doi : 10.1080/00071669008417288.

[2] Bourdillon A, (b), Carré B., Conan L., Duperray J., Huyghebaert G. European reference method for the in vivo determination of metabolisable energy with adult cockerels: Reproducibility, effect of food intake and comparison with individual laboratory methods. *British Poultry Science, 1990*, volume 31, 557-565, doi : 10.1080/00071669008417287.

[3] Bastianelli D., Bonnal L., Barre P. La spectrométrie dans le proche infrarouge pour la caractérisation des ressources alimentaires. *INRA Prod. Anim, 2018*, 31 (3), 237-254.

[4] Vilariño, M., Métayer J.P., Mahaut B., Bouvarel I. Caractériser la valeur nutritionnelle des aliments par des méthodes innovantes de mesure de la digestibilité pour une aviculture durable. *Innovations Agronomiques, 2016,* 49, 163-177.

[5] Traineau M., Pareux M., Danel J., Ammari F., Faure M., Vilariño M. Développement de nouvelles calibrations de spectroscopie dans le proche infrarouge pour la prediction de la digestibilité des nutriments chez le poulet de chair. *14èmes Jour Rech, Avicole et Palmipèdes à Foie Gras, 2022*, JRA22002-000088.

PASTIS-Net: Self-Aggregation of Chemometric Pre-processing for Spectral Deep Learning

D. Tanzilli[1,2] F. Abdelghafour[1,2] M.Metz[1,3,4]

[1] LabCom Aioly, Artificial Intelligence and Optics Laboratory, 34196, Montpellier, France, [2] ITAP, Univ. Montpellier, INRAE, Institut Agro, 34196, Montpellier, France, [3] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale Aix-Marseille Université, UMR CNRS IRD Avignon Université, Site de l'Etoile Marseille, France, [4] Pellenc ST, Applied Research Group, 84120, Pertuis, France

daniele.tanzilli@inrae.fr

Keywords: DeepChemometrics, Attention mechanism, Spectral pre-processing, Design of Experiments.

1 Introduction

In recent years, there has been a progressive convergence between chemometrics and deep learning, driven by the need to model increasingly complex, nonlinear analytical data. Spectroscopy is a particularly challenging application domain within this evolving framework, where data complexity is compounded by measurement-related variability. Raw spectral measurements are commonly affected by noise, instrumental artefacts and other sources of variability, which can obscure chemically relevant information and negatively impact model performance. In chemometrics, therefore, data pre-processing is recognised as a key step in improving data quality and isolating meaningful chemical information [1]. While deep learning models, such as convolutional neural networks (CNNs), can learn representations directly from raw data, the systematic integration of chemometric pre-processing knowledge into deep learning architectures is under-explored. While recent approaches combining multiple pre-processing by concatenating processed spectra have shown promising performance gains [2], this strategy substantially increases model complexity. This is particularly problematic in spectroscopic applications, where large, annotated datasets are rarely available, which limits the practical deployment of highly parameterised models.

2 Theory

Spectral pre-processing modifies the information content of the original measurements by emphasising relevant chemical features and reducing unwanted variability. Different pre-processing techniques emphasise different aspects of the spectral signal, so no single transformation can be considered optimal in all cases. Therefore, providing multiple pre-processed representations to a deep learning model can enrich the information available for learning. However, common strategies involve concatenating spectra that have been pre-processed in different ways along the wavelength dimension, which leads to a rapid increase in model complexity, in fact in fully connected layers, the number of trainable parameters grows with the size of the flattened input vector.

To address this issue, PASTIS-Net combines multiple pre-processed spectra along the channel dimension, preserving the original spectral length and enabling more compact architectures. The resulting multi-channel input is first embedded through a convolutional step and subsequently aggregated using a self-aggregation strategy based on an attention mechanism. This allows the network to adaptively weight the contribution of each pre-processed representation, retaining the most informative spectral features while controlling model complexity. Additionally, the classical hyperparameter optimisation method has been replaced by a Design of Experiments (DoE) approach to reduce the computational effort required for hyperparameter selection.

3 Material and methods

The proposed methodology was evaluated using the Open Soil Spectral Library (OSSL) [3], which is an open-access dataset comprising over 155,500 soil spectra samples from a variety of soil types and geographic regions. The database includes visible-near-infrared and mid-infrared spectra acquired using different spectrometers. It also provides laboratory-measured soil properties, such as organic carbon content. PASTIS-Net was trained on this dataset and compared with existing deep learning approaches.

4 Results and discussion

The results demonstrate that PASTIS-Net achieves superior predictive performance to state-of-the-art deep learning models while maintaining substantially reduced model complexity. The proposed aggregation strategy effectively exploits multiple pre-processing techniques without over-parameterisation. Furthermore, the use of the self-attention mechanism enables them to be combined into a more robust strategy in the case of incorrect pre-processing. Furthermore, applying DoE principles to hyperparameter optimisation is an efficient, systematic strategy that significantly reduces the computational cost and training time required to identify optimal model configuration.

5 Conclusion

This work demonstrates that chemometric principles can play a significant part in the development of deep learning models for spectroscopic analysis. The proposed PASTIS-Net architecture efficiently integrates multiple pre-processing strategies within a compact framework. These strategies provide the network with complementary information in the form of differently pre-processed spectral representations, thereby improving its predictive performance on real spectral data.

6 References

[1] Rinnan, Å. (2014). Pre-processing in vibrational spectroscopy–when, why and how. Analytical Methods, 6(18), 7124-7129.

[2] Mishra, P., & Passos, D. (2021). A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit. Chemometrics and Intelligent Laboratory Systems, 212, 104287.

[3] Zhou, L., Zhang, C., Taha, M. F., Wei, X., He, Y., Qiu, Z., & Liu, Y. (2020). Wheat kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method. Frontiers in plant science, 11, 575810.

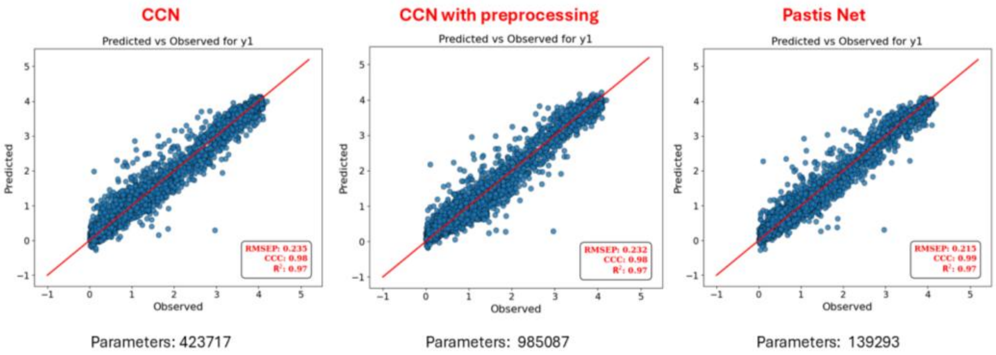


Figure 1 – Observed vs predicted values on the external test set for the considered models.

Analyse non-destructive du stress chez *Myriophyllum spicatum*

Yohann KUBLER[1] Orlane BABIN[2,3] Manuel PELLETIER[4] Elisabeth M. GROSS[5]

[1] Laboratoire Interdisciplinaire des Environnements Continentaux, yohann.kubler@univ-lorraine.fr

[2] Laboratoire Interdisciplinaire des Environnements Continentaux, orlane.babin@univ-lorraine.fr

[3] Nantes University, Master 2 Ecosystème et Bioproduction Marine

[4] Laboratoire Interdisciplinaire des Environnements Continentaux, manuel.pelletier@univ-lorraine.fr

[5]Laboratoire Interdisciplinaire des Environnements Continentaux, gross5@univ-lorraine.fr

Mots clés : plante aquatique ; biosurveillance active ; spectroscopie hyperspectrale.

1 Introduction

L'évaluation des risques environnementaux repose aujourd’hui essentiellement sur l’utilisation d'essais normalisés de laboratoire. Une limite forte de cette démarche est qu’elle ne reflète pas la réalité environnementale qui présente une contamination multiple (pesticides, métaux, etc…), avec des fréquences et des intensités variables. En parallèle, il existe une forte demande pour le développement d’une biosurveillance plus efficiente. De nombreux travaux ont été développés chez les poissons et les macroinvertébrés (amphipodes et bivalves) pour les utiliser comme organismes tests dans l’évaluation de la toxicité des milieux aquatiques. La biosurveillance active, via l’encagement de ces organismes animaux, a ainsi récemment montré sa pertinence pour une meilleure surveillance et évaluation de la qualité des milieux aquatiques.

Bien que les plantes aquatiques supérieures, spécialement les plantes immergées, aient un rôle essentiel dans le fonctionnement des milieux aquatiques, elles sont beaucoup moins utilisées dans l’évaluation écotoxicologique de ces écosystèmes. Si le groupe des plantes aquatiques immergées est inclus dans 2 tests normalisés pour *Myriophyllum spicatum* [1,2], il n’existe pas de méthodologie de biosurveillance active permettant d’étudier la réponse de ces organismes exposés directement dans les milieux. Or, l’utilisation d’analyses hyperspectrales à petite échelle, comme la feuille d’une plante, s’est développée les dernières années [3,4] et offre de nouvelles perspectives dans l’évaluation de l’impact de la pollution chimique des milieux aquatiques sur les plantes aquatiques.

Le *M. spicatum* est un bon modèle pour cette mise au point car elle est connue pour répondre de façon spécifique à certains polluants, avec une « empreinte chimique spécifique » qui reflète les changements physiologiques [5,6] mais aussi d’autres facteurs de stress environnementaux [7,8]. Il présente une variation de pigments lorsqu’il est soumis à un stress ; en lien avec une augmentation de la concentration en anthocyanine et/ou une baisse de la chlorophylle [5]. Cependant, la méthode classique de quantification par extraction et analyses est laborieuse et destructive. Des méthodes basées sur les propriétés spectrales peuvent permettre de détecter ces changements de façon rapide et non-destructive.

Dans ce contexte, le projet DISEPAM (Développement d'une détection non destructive *in situ* des effets des polluants sur une plante aquatique modèle ; projet inter CARNOT ICEEL + Eau&Environnement) vise à développer une technique de mesure non-destructive et in-situ.

2 Matériel et méthodes

L'induction du stress s'effectue *via* l'exposition à un fongicide et trois herbicides ayant des modes d'action différents (azoxystrobin, fluroxypyr, chlortoluron + mesosulfuron-methyl). Après deux semaines d'exposition, les concentrations en pigments ont été mesurées *via* trois méthodes : utilisation du Dualex®, une pince à feuille spectrale, de la spectroscopie spectrale (FLAME Vis-NIR avec une fibre optique QR400-7-VIS-NIR (400-2100 nm) et une lampe Krypton ecoVis (1.3 W) (Ocean Optics, Dunedin, FL, USA)), ainsi qu'avec une méthode classique d'extraction [6]. Ces méthodes ont été comparées par une analyse de corrélation.

3 Résultats

Les différents tests ont montré des résultats plus ou moins attendus sur la croissance et la pigmentation de la plante. De façon générale, l'analyse des résultats a montré une corrélation de modérée à forte entre les trois méthodes. Les relations relativement faibles entre la méthode optique et l'extraction peuvent s'expliquer par la difficulté d'extraire des biomasses inférieures à un milligramme.

4 Conclusion et ouverture

Ces premiers résultats suggèrent que la méthode optique présente un potentiel à détecter des stress induits par des polluants dans le myriophylle. Elle nécessite cependant des adaptations et améliorations pour prendre en compte la morphologie complexe de la plante. Dans ce contexte, un sujet de thèse à vu le jour et vise dans un premier temps à développer des approches de détection non-destructive, spectrales ou hyperspectrales, des effets de polluants chimiques sur le *M. spicatum*. Dans un second temps, l'objectif est de développer et valider des approches de biosurveillance active avec les plantes aquatiques immergées et enracinées afin de pouvoir tester une exposition *via* l'eau et le sédiment, directement sur le terrain.

5 Remerciements

Merci à Orlane pour la réalisation des expériences, le traitement et l'analyse des données. Merci à Manel, Bintou, Ellaura et Anaïs pour leur aide lors de la réalisation des expériences.

6 References

[1] OECD (2014). TG 238: Sediment-free Myriophyllum spicatum toxicity test. https://doi.org/10.1787/9789264224131-en.

[2] OECD (2014). TG 239: Water-sediment Myriophyllum spicatum toxicity test. https://doi.org/10.1787/9789264224155-en

[3] Dmitriev, P., B. Kozlovsky, T. Minkina, V. D. Rajput, T. Dudnikova, A. Barbashev, M. A. Ignatova, O. A. Kapralova, T. V. Varduni, V. K. Tokhtar, E. P. Tarik, İ. Akça and S. Sushkova (2023). Hyperspectral imaging for small-scale analysis of Hordeum vulgare L. leaves under the benzo[a]pyrene effect. Environmental Science and Pollution Research 30(55): 116449-116458. 10.1007/s11356-022-19257-0

[4] Zhai, Y., L. Zhou, H. Qi, P. Gao and C. Zhang (2023). Application of Visible/Near-Infrared Spectroscopy and Hyperspectral Imaging with Machine Learning for High-Throughput Plant Heavy Metal Stress Phenotyping: A Review. Plant Phenomics 5: 0124. https://doi.org/10.34133/plantphenomics.0124

[5] Gross, E. M., A. Nuttens, D. Paroshin and A. Hussner (2018). Sensitive response of sediment-grown Myriophyllum spicatum L. to arsenic pollution under different CO2 availability. Hydrobiologia 812(1): 177-191. https://doi.org/10.1007/s10750-016-2956-7

[6] Nuttens, A., S. Chatellier, S. Devin, C. Guignard, A. Lenouvel and E. M. Gross (2016). Does nitrate co-pollution affect biological responses of an aquatic plant to two common herbicides? Aquatic Toxicology 177:355-364. 10.1016/j.aquatox.2016.06.006

[7] Fornoff, F. and E. M. Gross (2014). Induced defense mechanisms in an aquatic angiosperm to insect herbivory. Oecologia 175(1): 173-185. https://doi.org/10.1007/s00442-013-2880-8

[8] Gross, E. M. (2022). Aquatic chemical ecology meets ecotoxicology. Aquatic Ecology 56(2): 493-511. https://doi.org/10.1007/s10452-021-09938-2

The R4multidata project:
Comparison of R tools for multidimensional data analysis.
Example with RGCCA and mixOmics for supervised methods

Marion BRANDOLINI-BUNLON[1] Elise MAIGNE[2] Sébastien THEIL[3] Isabelle SANCHEZ[4]
Virginie ROSSARD[5] Eric LATRILLE[6] Gwendal CUEFF[7] Marie TREMBLAY-FRANCO[8]
Nadia BESSOLTANE[9] Caroline PELTIER[10,11] Alyssa IMBERT[12]

[1] Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, 63000 Clermont-Ferrand, France, marion.brandolini-bunlon@inrae.fr

[2] Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet Tolosan, France, elise.maigne@inrae.fr

[3] Université Clermont Auvergne, INRAE, VetAgro Sup, UMR545 Fromage, 15000 Aurillac, France, sebastien.theil@inrae.fr

[4] INRAE, MISTEA, 2 place Pierre Viala, 34060 Montpellier, France, isabelle.sanchez@inrae.fr

[5] INRAE, Univ Montpellier, LBE, 102 Avenue des Etangs, F-11100 Narbonne, France, virginie.rossard@inrae.fr

[6] INRAE, Univ Montpellier, LBE, 102 Avenue des Etangs, F-11100 Narbonne, France, eric.latrille@inrae.fr

[7] Université Clermont Auvergne, INRAE, UNH, 63000 Clermont-Ferrand, France, gwendal.cueff@inrae.fr

[8] Toxalim, Université de Toulouse, UMR INRAE 1331, Metabohub-Metatoul-AXIOM, 31027 Toulouse cedex 3, France, marie.tremblay-franco@inrae.fr

[9] IJPB - Institut Jean-Pierre Bourgin - Sciences du végétal, nadia.bessoltane@inrae.fr

[10] Université Bourgogne Europe, Institut Agro, CNRS, INRAE, UMR CSGA, 21000 Dijon, France

[11] Probe Research Infrastructure, Chemosens facility, CNRS-INRAE, Dijon, France, caroline.peltier@inrae.fr

[12] INRAE, Université Clermont Auvergne, Vetagro Sup, UMRH, 63122 Saint-Genès-Champanelle, France, alyssa.imbert@inrae.fr

Keywords: multidimensional data analysis, R software, comparison.

1 Introduction

Multidimensional methods are essential for analyzing complex data (omics, spectral, etc.). Many R packages exist, but they have different philosophies. This leads users to question their differences in terms of functionality, maintenance, reproducibility, and results. The R4multidata project aims to create a standardized and collaborative environment for testing and comparing, with real and simulated data, the functions of these packages in terms of approach and application. In the event of algorithmic differences, the associated limitations are studied. The ultimate goal is to provide the necessary elements for making informed choices about tools.

2 Material and methods

In the statistical analysis and integration of complex heterogeneous data, multidimensional methods are mainly applied. Among the R packages, mixOmics [1] is one of the most widely used (2,500 to 3,000 downloads of the package per month). Functions in the initial mixOmics package were built based on methods developed by the authors of the RGCCA package [2], but in a way that simplifies their use by biologists. Besides, RGCCA was developed for including more methods in a unique and general framework [3]. These two packages therefore share a number of basic statistical methods, such as partial least squares (PLS) regression and its discriminant (“DA”) or variable selection (“sparse”) versions, Canonical Correlation Analysis (CCA), and their derived multiblock methods considering more than two blocks of data. However, the two packages then evolved independently, still with two different philosophies.

In the R4multidata project, eight methods from these two packages were considered: PLS regression and discriminant, sparse, and/or multiblock variants (PLS, PLS-DA, sPLS, sPLS-DA, mbPLS, mbPLS-DA, mbsPLS, mbsPLS-DA). Initially, these methods were studied from a theoretical point of view (optimization problem, initialization, deflation, block weighting, regularization, variable selection method, missing values handling, prediction methods). Their implementations in the two packages were compared (inputs and outputs of the main functions, functions to tune or evaluate the models, plots). At the same time, comparison criteria were determined, and datasets were prepared using real data from research projects or available in the packages, before the application of the functions and the comparison of the results.

3 Results and discussion

The conceptual differences that have been identified between the packages, are due to the fact that different parameters have been set by the developers, or left to the user's choice. For example, four deflation modes are offered in mixOmics, whereas a single mode is offered in RGCCA. Conversely, in multiblock methods, the sum of covariances is always maximized in mixOmics, whereas it is possible to maximize the sum of covariances, the squares of covariances, or the absolute values of covariances in RGCCA. Moreover, data blocks can be weighted in RGCCA and not in mixOmics. There are also differences in the strategy applied, particularly for prediction in multiblock methods, which is done by block in mixOmics before averaging, and which is done from concatenated components in RGCCA.

Application to the datasets shows that, among other things, with equivalent settings, the first components obtained with the two packages are often comparable, and differences appear in the subsequent components.

There are also clear differences in terms of purpose and target users: The mixOmics R package is intended for use in regression and discriminant analysis (several classification methods and performance indicators dedicated to regression or discrimination), and/or by novice users. Conversely, the RGCCA R package is more intended for experienced users, as it allows for a more rigorous approach (more refined parameterization) and as it allows, through a judicious choice of parameters, a greater number of methods to be applied.

4 Conclusion

Although based on the same framework, the algorithms and functions of the two packages for applying PLS or PLS-DA regression with their sparse and multiblock variants have many differences that should be clearly identified when performing analyses to make the better choices.

5 References

[1] Rohart F., Gautier, B, Singh, A and Lê Cao, K. A. (2017) mixOmics: an R package for ‘omics feature selection and multiple data integration. *PLoS Comput Biol* 13(11): e1005752. doi: 10.1371/journal.pcbi.1005752.

[2] Tenenhaus, A., Tenenhaus, M. (2011) Regularized Generalized Canonical Correlation Analysis. *Psychometrika* 76(2), 257–284. doi: 10.1007/s11336-011-9206-8

[3] Tenenhaus M., Tenenhaus A. and Groenen P. J. (2017). Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika*, 82(3), pp.737-777. doi: 10.1007/s11336-017-9573-x.

PAT Blend Uniformity and Content Uniformity Platform at Roche Basel

Mr. Rokitowski[1] Ms. Bâty[2]

[1] F-Hoffmann-La Roche Grenzacherstrasse 124 4070 Basel (Switzerland), joris.rokitowski@roche.com

[2] F-Hoffmann-La Roche Grenzacherstrasse 124 4070 Basel (Switzerland), chloe.baty@roche.com

Keywords: Blend Uniformity, Content Uniformity, PAT, Drug Product, Pharma, PLS modeling

1 Introduction

Blend uniformity (BU) and content uniformity (CU) are two standard measurements [1] for monitoring proper uniformity of the active principal ingredient (API) during drug product (DP) manufacturing, in the pharmaceutical industry. These measurements usually rely on classical wet analytics (i.e HPLC), requiring manual sampling and extensive lead time before results are available. Using near infrared (NIR) and Raman spectroscopy, Roche has developed a complete process analytical technology (PAT) platform. This enables fast and reliable BU and CU results, bringing analytics one step closer to the development and manufacturing process.

2 Theory

BU and CU requirements as well as the use of spectroscopy-based tools are well defined and described in corresponding Pharmacopeias [2] and Health Authorities guidance [3].

3 Material and methods

PAT BU relies on a NIR SentroPAT BU - Sentronic® spectrometer which is directly attached to the bin blender for in-line measurements during the blending step of the manufacturing process.

PAT CU relies on an NIR Multipurpose Analyzer (MPA) – Bruker® spectrometer and/or on a Transmission Raman Spectrometer (TRS) – Agilent® which is used to perform at-line measurements of tablets coming out of the tablet press or the coating process.

The calibration of PAT BU and CU is usually performed by manufacturing lab scale calibration batches, with a ranging concentration of API. This requires low amount of material, and a “sample free” calibration methodology is also being developed.

Partial least squares (PLS) models are elaborated and used to predict the content of API in the corresponding drug product batch.

4 Results and discussion

Blend uniformity data (Figure 1) indicates that API content reaches uniformity after 15 rotations, achieving a steady state that guarantees homogeneity. Subsequent content uniformity analysis (Figure 2) demonstrates that tablet potency remains consistent and aligned with target values. This lack of segregation [4] during the compression phase validates the process capability and ensures the final dosage form meets the required potency targets.

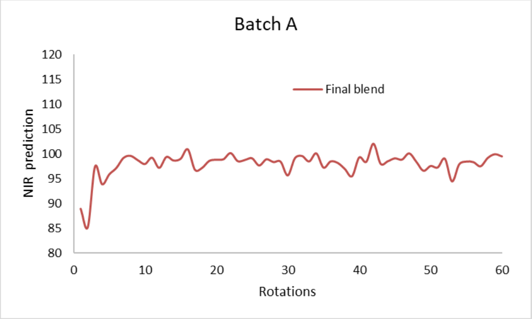


Figure 1: Example of BU results - API content predictions over blending rotations

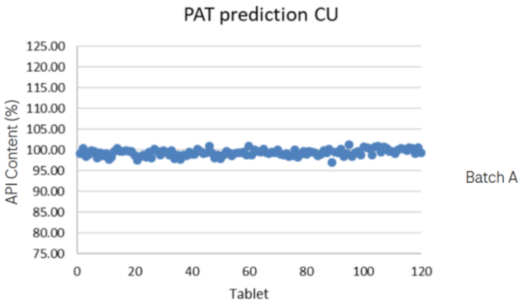


Figure 2: Example of CU results – API content predictions in tablet samples

5 Conclusion

The BU CU Platform adds significant value to numerous Roche projects by streamlining the development phase. By reducing analytical overhead and providing rapid access to data, the platform facilitates faster decision-making. In a recent application, the platform was key in the successful implementation of PAT-based release testing.

6 References

[1] Bautista M, Caille S, Corredor C, et al., Blend Uniformity and Content Uniformity in Oral Solid DosageManufacturing: an IQ Consortium Industry Position Paper, *AAPS J.*, 27(2):49, 2025.

[2] Council of Europe, 2.9.40 Uniformity of dosage units, *European Pharmacopeia 11th edition*, 2017.

[3] United States Food and Drug Administration Center for Drug Evaluation and Research, Development and Submission of Near Infrared Analytical Procedures Guidance for Industry, 2021

[4] Jakubowska E, Ciepluch N., Blend segregation in tablets manufacturing and its effect on drug content uniformity—a review, *Pharmaceutics*, 13(11):190, 2021.

Multiway decomposition of fluorescence spectral data for the prediction of lignocellulosic biomass chemical composition

O. Lehmam[1] A. Faraj[2] G. Paës[3]

[1] Université de Reims Champagne-Ardenne, INRAE, FARE, UMR A 614, Reims, France, oumaima.lehmam@inrae.fr

[2] Université de Reims Champagne-Ardenne, INRAE, FARE, UMR A 614, Reims, France, ali.faraj@inrae.fr

[3] Université de Reims Champagne-Ardenne, INRAE, FARE, UMR A 614, Reims, France, gabriel.paes@inrae.fr

Keywords: Lignocellulosic biomass, Fluorescence spectroscopy, Excitation emission matrices, Multiway spectral decomposition, Multivariate analysis, Machine learning

1. Introduction

Lignocellulosic biomass (LB) is a crucial renewable resource primarily composed of lignin, cellulose, and hemicelluloses [1]. Efficiently characterizing its chemical properties is vital for its industrial valorization. Traditional chemical analysis methods can be time-consuming and expensive.

The fluorescence spectra of steam-exploded LB samples, which correspond to Excitation Emission Matrices (EEMs) [2], were shown to be directly correlated to their saccharification potential [3]. For this reason, EEMs are good candidates as indicators of the chemical composition of lignocellulosic biomass. In the present work, we propose to apply the multiway spectral decomposition PARAFAC (Parallel Factor Analysis) on the EEMs. The scores resulting from the PARAFAC decomposition are then used to predict the sample sugars (including glucose) and lignin contents by the use of machine learning techniques. Although this approach is well developed for dissolved organic matter analyzed as EEMs [4], it is the first time that it is used for LB fluorescence spectra. Another novelty of the present work is that the PARAFAC analysis of EEMs, and the resulting predictions, were fully developed in Python (whereas Matlab or R are used in previous works).

2. Theory

The PARAFAC decomposition of three-way arrays of EEMs (sample × excitation wavelength × emission wavelength) correspond to the trilinear relationship [2]:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \ , \tag{1}$$

where x_{ijk} corresponds to the fluorescence intensity of the i^{th} sample, j^{th} excitation wavelength and k^{th} emission wavelength. Each r corresponds to a PARAFAC component.

3. Material and methods

The spectra of 42 LB samples were recorded (21 from beech wood + 21 from spruce wood). Samples underwent steam explosion pretreatment under various severity conditions: pre-soaking with water or acid, temperatures and process duration from 170°C to 210°C and from 5 min to 15 min, respectively. Fluorescence spectra were recorded at excitation wavelength from 450 nm to 500 nm and emission wavelength from 460 nm to 520 nm. Spectra were pre-processed through the development of an automated Python pipeline.

The Python package `TensorLy` was employed to perform the PARAFAC decomposition. The PARAFAC scores were used to train several machine learning models on 75% of the data (training set) for the prediction of glucose, sugars and lignin contents. The models accuracy was evaluated on the remaining 25% of the data (test set) by using the R^2, MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). In order to improve the relevance of the predictive performance evaluation on the test set, a 4-fold cross-validation iterator was applied to repeat the train-test split.

4. Results and discussion

The PARAFAC decomposition provided 4 spectral components such that Component 1 relates to glucose (through correlation with the corresponding scores), Component 2 to sugars, Component 3 to lignin and Composant 4 to glucose and lignin. The bar plot representation of the mean relative contributions of the 4 components by separating pretreatment experimental conditions allowed sample discrimination from spectral information. A Principal Component Analysis (PCA) applied to the PARAFAC scores confirmed the sample discrimination, since data expressed in the PCA axes clustered according to the water and acid pretreatment types.

Linear regression models were established from the PARAFAC scores of the whole dataset to predict chemical contents and provided R^2 metrics equal to 0.63 for lignin, 0.68 for glucose and 0.82 for the remaining sugars. The corresponding MAPE error for sugars exceeded 0.61%, which after outliers detection, was identified as coming from samples pretreated with acid. The separation of the dataset in two groups (acid pretreatment and water pretreatment) provided a significant improvement of the linear regression R^2 metrics for both datasets: more than 0.92 for lignin and sugars and more than 0.81 for glucose. The MAPE errors were also significantly reduced.

The 4-fold cross-validation allowed the evaluation of the predictive performances (on unseen data) of the models obtained by establishing machine learning regressors from the PARAFAC scores. The best prediction performances were obtained for the water pretreated samples with a mean MAE over the 4-fold test sets (expressed as percentage of target content mean) equal to 8.18% for lignin, 7.2% for glucose and 16.16% for remaining sugars, with narrow standard deviation. Similar, prediction performances were obtained on acid pretreated samples for lignin and glucose. For the remaining sugars of the acid pretraited samples, predictive performances were lower but could be improved by using a Partial Least Squares Regressor (PLS) for the machine learning step.

5. Conclusion

The results provide an automated end-to-end pipeline for pretreated LB polymer content characterization. While lignin, glucose and sugars concentrations could be predicted with high precision for water pretreated samples, predicting sugars for acid pretreated samples remains a challenge that may require deep learning approaches to capture more complex spectral structures. The present tools are slated for integration into the Galaxy platform for broader scientific use.

6. References

[1] M. Pauly & K. Keegstra: Plant cell wall polymers as precursors for biofuels. *Curr. Opin. Plant Biol.*, 13: 305-312, 2010.

[2] K.R. Murphy, C.A. Stedmon, D. Graeber & R. Bro: Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Anal. Methods*, 5(23): 6557-6566, 2013.

[3] T. Auxenfans, C. Terryn & G. Paës: Seeing biomass recalcitrance through fluorescence. *Sci. Rep.,* 7: 8838, 2017.

[4] Y. Yang, C. Shan & B. Pan: Machine learning modeling of fluorescence spectral data for prediction of trace organic contaminant removal during UV/H2O2 treatment of wastewater. *Water Res.*, 255: 121484, 2024.

All Together Now: Generating and Using Ensembles for Regression

Manuel A. Palacios and Barry M. Wise[1]

[1] Eigenvector Research, Inc. Manson, WA USA, bmw@eigenvector.com

Keywords: Calibration, regression, automation, machine learning, model building, ensembles

1 Introduction

In 2024 we introduced our semi-automated machine learning (semi-auto ML) method for accelerating the development of quality calibration models [1]. We refer to this approach as "Diviner" (Divine—to discover or locate something by intuition, insight or supernatural means.). As part of its search for the "best" calibration model, Diviner generates hundreds or even thousands of regression models with different numbers of factors (latent variables, LV), selected variables and preprocessing steps. Many of these models have similar, very good performance in spite of being constructed differently. Thus, we began to investigate if it was possible to achieve a "wisdom of the crowds" type advantage by leveraging these models with a fusion approach.

In this talk we present a method for selecting models based on the diversity or "ambiguity" of the ensemble along with its performance. Multiple methods for fusing the results are also considered. The performance of the ensembles is demonstrated using several data sets from NIR spectroscopy.

2 Theory

Diviner presents model performance on plots of overfit, as indicated by the ratio of the RMSECV to RMSEC), versus predictive ability, as indicated by RMSECV. Models that are low on both these scales can be further refined. After this refinement users can select a single model for use or select a group of models. Selections from this group can be used to form an ensemble: a collection of models that are used simultaneously with the results combined or fused to produce a single prediction. Ensembles can be further refined by selecting a subset that optimizes RMSECV and ambiguity. Ambiguity is a measure of an ensemble's diversity and is the mean squared variation in the predictions of the individual models comprising the ensemble.

Several methods for combining results from ensembles were tested. This includes median, weighted based on model RMSECV, and linear stacked, where a regression was performed on the model outputs to produce the final value.

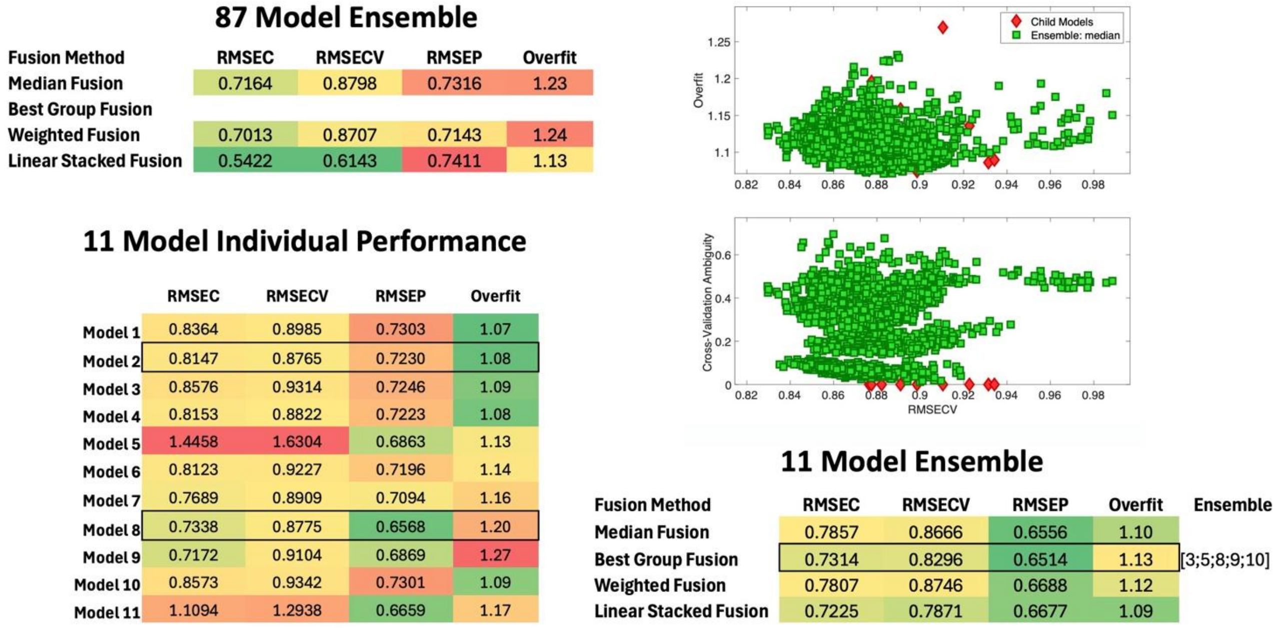
3 Material and methods

The ensemble methods were tested on five different NIR data sets. In this abstract results are shown for one of them, the SBR data set of Miller [2]. It includes 60 calibration samples and 10 test samples of NIR transmission spectra (1575-1850 nm) of styrene-butadiene copolymers in CCl_4 solutions of approximately 1% by mass different styrene, cis-, trans- and 1,2 butadiene contents (determined by NMR) all spectra obtained in 4mm cuvette.

All computations for this work were done with PLS_Toolbox 9.5.1 (Eigenvector Research, Inc., Manson, WA USA) running under MATLAB 2024b (The MathWorks, Natick, MA USA).

4 Results and discussion

The ensemble results for SBR data are shown in Figure 1 below. Initially 87 models were selected from the approximately 3000 Partial Least Squares (PLS) regression models generated by Diviner, each with different model parameters including preprocessing steps, wavelength selections and number of latent variables. This models were selected based on their RMSECV and overfit ratios. Results of fusion on the 87 models are shown in the upper left of Figure 1. A group of 11 models was selected from the 87 for use in the best ensemble search. The performance of the 11 models is shown in the lower left of Figure 1. All combinations of 3-11 models were tested in the search for the optimal ensemble. The overfit, ambiguity and RMSECV for these ensembles is shown in the upper right. Finally the performance of the ensemble of 11 and optimal 5 are shown in the lower right.



87 Model Ensemble

Fusion Method	RMSEC	RMSECV	RMSEP	Overfit
Median Fusion	0.7164	0.8798	0.7316	1.23
Best Group Fusion				
Weighted Fusion	0.7013	0.8707	0.7143	1.24
Linear Stacked Fusion	0.5422	0.6143	0.7411	1.13

11 Model Individual Performance

	RMSEC	RMSECV	RMSEP	Overfit
Model 1	0.8364	0.8985	0.7303	1.07
Model 2	0.8147	0.8765	0.7230	1.08
Model 3	0.8576	0.9314	0.7246	1.09
Model 4	0.8153	0.8822	0.7223	1.08
Model 5	1.4458	1.6304	0.6863	1.13
Model 6	0.8123	0.9227	0.7196	1.14
Model 7	0.7689	0.8909	0.7094	1.16
Model 8	0.7338	0.8775	0.6568	1.20
Model 9	0.7172	0.9104	0.6869	1.27
Model 10	0.8573	0.9342	0.7301	1.09
Model 11	1.1094	1.2938	0.6659	1.17

11 Model Ensemble

Fusion Method	RMSEC	RMSECV	RMSEP	Overfit	Ensemble
Median Fusion	0.7857	0.8666	0.6556	1.10	
Best Group Fusion	0.7314	0.8296	0.6514	1.13	[3;5;8;9;10]
Weighted Fusion	0.7807	0.8746	0.6688	1.12	
Linear Stacked Fusion	0.7225	0.7871	0.6677	1.09	

Figure 1 – Ensemble results for SBR data.

Note that the RMSEP values for the ensembles using the 11 models are all very close or slightly better than the best of the individual models.

5 Conclusion

Based on results across the data sets considered here, the performance of ensembles is generally nearly as good as the best individual model on independent test sets and often better. The methods for fusing the ensemble results (median, best ensemble, weighted, stacked) show no clear winner, in fact they all work.

6 References

[1] M. Palacios, S. Roginski and B.M. Wise, Diviner: A Semi-Automated Machine Learning Approach to Calibration Model Development. *Chimiométrie 2024, 2024.*

[2] C.E. Miller et al, Anal. Chem., 1990, 62, 1778.

Smartphone-Based Aspartame Analysis: Integrating Hydrophobic Barrier-Free LPAD with RGB Color-Ratio Modeling

M. Nakarin Noirahaeng[1] M. Jirawat Salungyu[2] M. Phoonthawee Saetear[1]

[1] Flow Innovation-Reseach for Science and Technology Laboratories (Firstlabs), Department of Chemistry and Center of Excellence in Innovation for Chemistry, Faculty of Science, Mahidol University, Rama 6 Road, Ratchatewi, Bangkok 10400 Thailand, nakarin.noa@gmail.com

[2] Department of Science Service (DSS), 75/7 Rama VI Road, Ratchathewi Bangkok 10400 Thailand

Keywords: Laminated paper-based analytical device; LPAD; ninhydrin; aspartame; heat-induced complex formation; Ruhemann's complex; color ratio

1 Introduction

This research presents a holistic approach to aspartame detection by integrating a heat-assisted, barrier-free LPAD design with digital quantification using smartphone-based RGB-colorimetry. By optimizing reagent distribution and reaction speed through heat, this system bridges the gap between low-cost paper fluidics and sophisticated digital analysis, establishing a reliable benchmark for rapid and robust agri-food safety testing.

2 Theory

Density Functional Theory (DFT) confirms the stability of the Ruhemann's purple complex, while Kinetic Modeling establishes the activation energy (E_a) justifying the 60°C thermal enhancement. For quantification, the study utilizes an Optical Model based on the Beer-Lambert law, where smartphone-captured RGB intensities are converted into normalized absorbance-equivalents (A) using the following relationship:

$$A = \text{color ratio} = \log (I_0/I_x) \qquad (1)$$

The model calculates the ratio between the intensity of a specific color channel (I_R, I_G, or I_B) and a reagent blank (I_0) for qualitative results. This color ratio is then normalized to a range of 0 to 1, ensuring data stability and consistent signal processing across different detection zones.

3 Material and methods

3.1 Analytical procedure for demonstrating in analysis of aspartame

The LPAD is fabricated by laminating filter paper between transparent sheets to create a stable, barrier-free platform [1-4]. Thermal energy is applied in two stages: (1) dry the ninhydrin reagent (18 g L^{-1}) and (2) accelerate the reaction with 4 µL aspartame samples at 60 °C for 7 minutes. The formation of a Ruhemann's purple complex was analyzed by the *Palette Cam app on* smartphone.

4 Results and discussion

4.1 Proof of concept of aspartame-ninhydrin complexation

The reaction between ninhydrin and aspartame triggers a color shift from pale-yellow to purple, characterized by a of 567 nm and a linear absorbance response across 0.4–5.2 g. To validate the formation of the **Ruhemann's purple complex**, density functional theory (DFT) and TD-DFT calculations were performed using the B3LYP functional [5]. The predicted absorption spectrum, modeled in water, showed excellent agreement with experimental data, theoretically confirming the successful complexation of ninhydrin and aspartame.

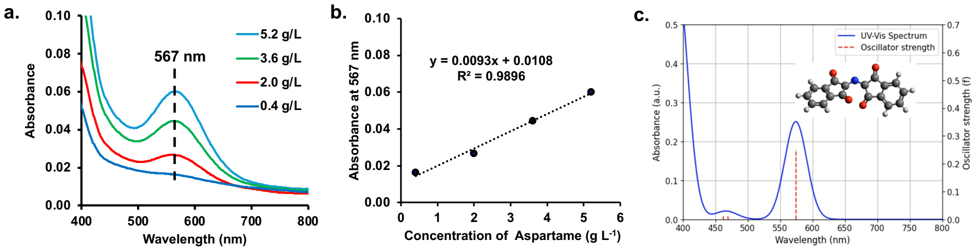


Figure 1 a) UV-Visible spectrum of aspartame-ninhydrin reaction of 0.4 – 5.2 g L^{-1} aspartame and 10 g L^{-1} of ninhydrin, b) Calibration plot of aspartame-ninhydrin at 565 nm and c) RGB-calibration curve from ImageJ of 0.1 – 0.5 g L^{-1} aspartame and 10 g L^{-1} of ninhydrin

4.2 Concentration of aspartame dependent on the color intensity on the LPAD

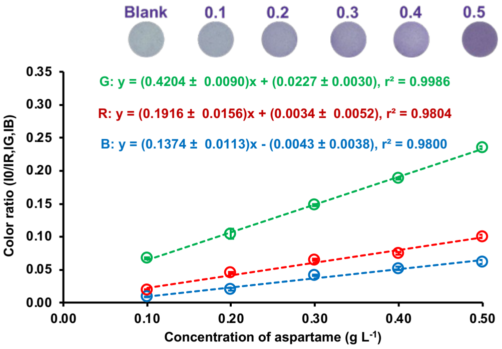


Figure 2 a) RGB- Calibration plots from 0.1 – 0.5 g L^{-1} aspartame with 10 g L^{-1} ninhydrin using the LPAD for the determination of aspartame and the corresponding images of the purple aspartame-ninhydrin complex.

Images of the LPAD are captured via smartphone in a light-controlled box and processed using the Palette Cam app with a 4-spot circumferential measurement strategy for consistency. The analytical signal is defined as the color ratio, where represents the DI water blank. Using 18 g of ninhydrin and thermal activation, the green channel (log (I_0/I_G)) was selected for its superior sensitivity and complementary color match to the purple complex. This method yielded a strong linear calibration across an aspartame concentration range of 0.1 to 0.5 g.

5 Conclusion

We demonstrate a seamless transition from innovative experimental design (heat-assisted fluidics using a barrier free LPAD) to sophisticated modeling (DFT and optical analysis), reducing detection to 7 minutes for aspartame detection. By coupling smartphone RGB-colorimetry with rigorous data quantification, this platform achieves 100–105% recovery, offering a low-cost, high-precision solution for modern agri-food quality control.

6 References

[1] Noirahaeng, N.; Uraisin, K.; Wattanasin, P.; Saetear, P.*Analytical Sciences* **2022**, *38* (3), 533-540.

[2] Noirahaeng, N.; Salungyu, J.; Teerasong, S.; Uraisin, K.; Saetear, P.*Talanta Open* **2024**, *9*, 100310.

[3] Jantasin, A.; Worakul, T.; Surawatanawong, P.; Saetear, P.; Ruangsupapichat, N.*Sensors and Actuators B: Chemical* **2024**, *418*, 136228.

[4] Nantapon, T.; Naweephattana, P.; Surawatanawong, P.; Saetear, P.; Chantarojsiri, T.; Ruangsupapichat, N.*Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2022**, *282*, 121662.

[5] Frisch, M.*Inc, Wallingford CT* **2009**, *201*.

Déploiement d'un système PAT : NIR–HPLC : modélisation PLS pour le suivi en ligne d'une réaction de synthèse

Belal Gaci[1] Cédric CAUFOURIER[2] Mélanie DELVAUX[3] Yoann GUT[4] Vincent GEMBUS[5]

[1]ORIL Industrie, Servier Bolbec, belal.gaci@servier.com

[2]ORIL Industrie, Servier Bolbec, cédric.caufourier@servier.com

[3]ORIL Industrie, Servier Bolbec, mélanie.delvaux@servier.com

[4]ORIL Industrie, Servier Bolbec, yoann.gut@servier.com

[5]ORIL Industrie, Servier Bolbec, vincent.gembus@servier.com

Mots-clés : PAT, NIR, HPLC, PLS, Suivi de réaction de synthèse.

1 Introduction

Les technologies d'analyse des procédés (PAT) constituent aujourd'hui un levier essentiel de la fabrication pharmaceutique moderne. Elles permettent une surveillance, un contrôle et une optimisation en temps réel des opérations, assurant ainsi une qualité de produit constante tout en répondant aux exigences réglementaires croissantes. Leur essor est porté par la recherche d'une plus grande efficacité industrielle, la réduction des coûts et la transition progressive de la production par lots vers des procédés continus [1].

ORIL Industrie, filiale du groupe Servier, s'inscrit pleinement dans cette dynamique d'innovation afin de relever les défis actuels du secteur. Dans le cadre du suivi en temps réel d'une réaction de synthèse, un système PAT a ainsi été déployé pour estimer en continu la concentration d'un produit d'intérêt et ajuster les paramètres opératoires, dans l'objectif d'optimiser le rendement et de fiabiliser le procédé.

2 Matériel et méthodes

Le dispositif expérimental est constitué d'un spectromètre NIR MPA II (Bruker) relié à une flowcell Indatech par des fibres optiques. Les spectres sont acquis toutes les 30 secondes et une mesure de référence HPLC est réalisée toutes les cinq minutes. Un modèle PLS a été élaboré à partir de 31 échantillons analysés simultanément par NIR et HPLC, puis enrichi avec des spectres représentatifs des phases initiales et finales de la réaction.

Le modèle appliqué repose sur une calibration PLS [2] construite à partir de 31 mesures NIR et HPLC, permettant d'extraire l'information pertinente du spectre et de prédire la concentration du produit suivi. Afin de réduire les effets additifs et multiplicatifs, un prétraitement SNV a été appliqué [3]. Le choix du nombre optimal de variables latentes a été réalisé au moyen d'une validation croisée k-fold, chaque fold étant constitué de 6 échantillons. Un ensemble indépendant de 6 échantillons a ensuite été utilisé pour valider le modèle.

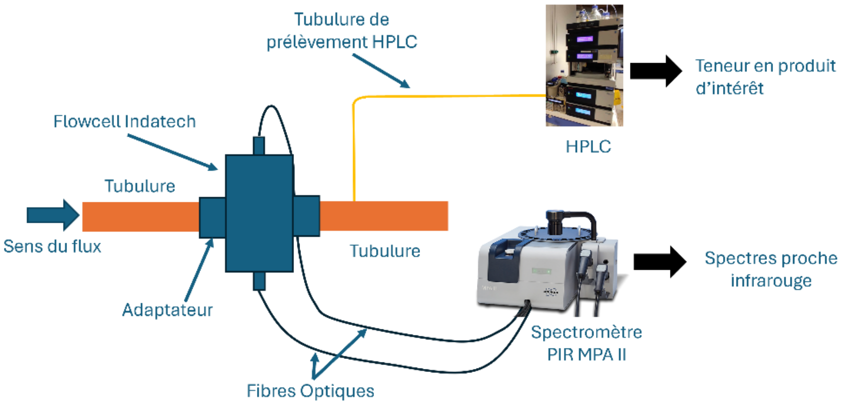


Figure 1 - Schéma illustrant le montage utilisé pour le suivi en ligne

3 Résultats et discussion

Le modèle développé affiche de bonnes performances avec une RMSEC relative de 1,2 %, une RMSECV relative de 1,2 % et une RMSEP relative de 1,4 %, obtenues en utilisant trois variables latentes. Ce modèle a été appliqué au suivi en ligne d'une synthèse sur une durée de quatre jours, permettant la collecte de 4 800 spectres NIR corrélés à 293 analyses de référence HPLC.

Grâce à une interface développée sous Matlab, les prédictions ont pu être suivies en temps réel. Les résultats montrent une forte corrélation entre les valeurs prédites par le modèle et les valeurs de référence. Toutefois, une légère surestimation est observée en début et en fin de réaction, attribuée à la présence d'un réactif en excès. Pour pallier ces effets, diverses stratégies d'amélioration de la robustesse sont actuellement à l'étude.

4 Conclusion

Le système PAT NIR-HPLC développé permet un suivi en ligne fiable de la réaction grâce à un modèle PLS performant. Cependant, les variations du procédé, telles que les changements de concentration des réactifs, les conditions physiques (température, pression) et la difficulté à maintenir un background spectrométrique stable sur plusieurs jours dans un système fermé, peuvent affecter les performances du modèle, nécessitant des stratégies pour renforcer la robustesse.

5 Références

[1] E. J. Kim, J. H. Kim, M.-S. Kim, S. H. Jeong, and D. H. Choi, "Process Analytical Technology Tools for Monitoring Pharmaceutical Unit Operations: A Control Strategy for Continuous Process Verification," *Pharmaceutics*, vol. 13, no. 6, p. 919, Jun. 2021, doi: 10.3390/pharmaceutics13060919.

[2] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986, doi: 10.1016/0003-2670(86)80028-9.

[3] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra," *Appl Spectrosc*, vol. 43, no. 5, pp. 772–777, Jul. 1989, doi: 10.1366/0003702894202201.

CHIMIOMÉTRIE XXVÈME
ENSIC LIEC SFdS UNIVERSITÉ DE LORRAINE cnrs
17-19 FÉVRIER 2026
CAMPUS GRANVILLE, NANCY

Decoding holistic aging perception using pairwise comparison of AI-generated faces

Olivia Dufour[1], Damien Brémaud[2], Philippe Courcoux[3], Anissa Azouaoui[1], Sileye Ba[1], Marie Thomas[1]

[1]L'Oréal R&I, Clichy, olivia.dufour@loreal.com; [2]Damien Brémaud Consulting, Nantes; [3]Statistics consultant, Nantes

Keywords: Apparent Age, Pairwise Comparison, Design of Experiments, Random Forest, Generative Adversarial Network.

1 Introduction

The facial anti-aging field is inherently holistic and complex, requiring precise understanding of holistic concepts such as consumer perception of age to develop targeted cosmetic products.

Today, decoding holistic beauty concepts at L'Oréal requires standardized photographs from clinical studies, manual scoring by experts, and pairwise comparison task - creating bottlenecks through high costs, extended timelines, and GDPR regulatory constraints.

We present an innovative digital methodology that addresses these challenges by providing a controlled experience while leveraging objective clinical signs. As a proof of concept, this approach was applied to decode apparent age using specific skin clinical signs assessed through pairwise image comparison test, selected for its simplicity and intuitiveness. This methodology can be readily adapted to other studies.

2 Material and methods

Subject's perception of age is assessed through a pairwise comparison test: subjects are presented with a series of pairs of face pictures and asked to identify which face appears older in each pair. Thirty (30) faces were sourced from https://generated.photos to ensure the use of synthetic (non-real) faces, thereby avoiding ethical concerns. Thirteen clinical signs of facial aging, each defined with 5 levels according to L'Oréal Atlas [1] are investigated (e.g.: crow's feet wrinkles or glabellar wrinkles). Image generation relies on a GAN-based (Generative Adversarial Network) [2] deep generative model that morphs initial faces by modulating aging signs intensity according to combinations determined by Federov's algorithm [3]. Each pair of images presented to subjects derives from the same initial images but displays distinct clinical sign variations. Pair presentation is balanced using Kraitchik's technique [4]. Figure 1 illustrates this complete sequential workflow.

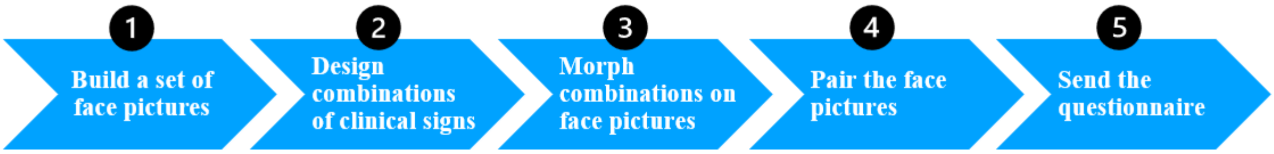

1
Build a set of face pictures
2
Design combinations of clinical signs
3
Morph combinations on face pictures
4
Pair the face pictures
5
Send the questionnaire

Figure 1 – Sequential approach to decode apparent age from clinical signs

Using an optimal design of experiments on 13 clinical signs with 5 intensity levels each, 220 unique combinations were generated. 219 subjects were required and each subject viewed 10 faces (out of 30) and evaluated 110 image pairs (220 images total) within a 10-minute session, producing 24,090 total observations for all participants.

3 Results

The analyses are performed on 219 French women from the Dijon region (mean age=43.8 ± 14.4 years, range=18-70). Analyses are based on the Bradley-Terry-Luce model [5] and reveal that: (a) ANOVA shows significant effects for all signs, confirming their influence on apparent age; (b) Response Surface Methodology (RSM) achieves high predictive performance with predominantly linear effects; (c) RSM predictions demonstrate perfect linearity for Eyebags front sign with significant inter-level differences (Figure 2) and clear perceptibility of all intensity levels for other signs; and (d) Random Forest analysis (100 forests of 500 trees) establishes a hierarchy of signs with significant differences in their importance.

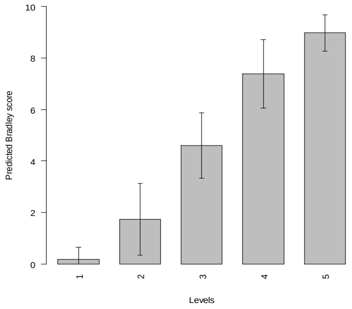


Figure 2 – Prediction of Bradley scores for the 5 levels of eyebags front sign.

4 Conclusion

All 13 clinical signs demonstrate highly significant effects on apparent age perception. However, a clear hierarchy emerges, validated across multiple statistical methods. The most influential signs include eyebags, upper lip wrinkles, ptosis, and underneath eye wrinkles; while forehead wrinkles, cheek pores, and interocular wrinkles exhibit less influence.

Linear effects are predominant, with significant interactions observed between several signs, whereas quadratic effects are weaker.

Both Response Surface and Random Forest models demonstrate good predictive quality. Importantly, all intensity levels across signs are clearly perceptible to consumers, confirming their relevance.

These results confirm the robustness of the methodology. To validate the operational results, future work should extend this approach to other holistic consumer-related concepts and age-stratified analyses, given the differences in perception across age groups described in the literature [6].

5 References

[1] Bazin R., Doublet E., Skin Aging Atlas. Volume 2 Asian Type. *Paris: Editions Med'Com,* 2010.

[2] Despois J., Flament F., Perrot M., AgingMapGAN (AMGAN): High-Resolution Controllable Face Aging with Spatially-Aware Conditional GANs. *European Conference on Computer Vision*:613-628. Springer, Cham, 2020.

[3] Fedorov V.V., Theory of Optimal Experiments. *Academic Press,* 1972.

[4] David, H.A., The Method of Paired Comparisons. *Oxford University Press*, 1988.

[5] Bradley R.A. and Terry M.E., Rank analysis of incomplete block designs. *Biometrika, 39,* pp. 324-345, 1952.

[6] Merinville E., Grennan G.Z., Gillbro J.M., Mathieu J. and Mavon A., Influence of facial skin ageing characteristics on the perceived age in a Russian female population. *International Journal of Cosmetic Science*, 37, 2015.

DOE–Based optimization of TD/Py-DART MS parameters
for analysis of fluoropolymers

L. Cadona[1,2] S. Schramm[1] G. Gaiffe[2] F. Progent[2] F. Aubriet[1]

[1] LCP-A2MC Université de Lorraine F-57000 Metz France, louise.cadona@univ-lorraine.fr
sebastien.schramm@univ-lorraine.fr frederic.aubriet@univ-lorraine.fr

[2] CEA DAM DIF F-91297 Arpajon France, gabriel.gaiffe@cea.fr frederic.progent@cea.fr

Keywords: Design of Experiment, Mass spectrometry, Fluoropolymers.

1 Introduction

Detecting polymers in complex environmental matrices, such as soils contaminated by industrial activities, remains a major analytical challenge. Fluoropolymers are of particular concern due to their high persistence in the environment, resulting from their specific physicochemical properties, including high thermal stability, chemical inertness and excellent resistance to ageing [1]. Improving analytical sensitivity for environmental samples is therefore essential to detect trace-level polymers, to better understand the industrial processes responsible for their production or release, and to identify potential sources of contamination.

This study presents the optimization of analytical parameters of a thermal desorption/pyrolysis (TD/Py) device coupled with a Direct Analysis in Real Time (DART) ion source to enhance mass spectrometry (MS) detection sensitivity using a Design of Experiments (DOE) approach.

2 Theory

TD/Py-DART MS enables rapid analysis of solid or liquid samples through controlled heating from ambient temperature up to 600 °C, allowing the emission, ionization, and detection of volatile compounds and pyrolysis products [2]. When coupled with high-resolution mass spectrometry (HRMS), this technique provides valuable information into the chemical structure of polymers, including repeat units, end groups, and additives, while requiring little to none sample preparation.

This approach has previously demonstrated its ability to differentiate fluoropolymers, identify polymer synthetized from multiple co-monomers, and distinguish fluoropolymers within a polymer blend [3].

3 Material and methods

Poly(vinylidene fluoride-*co*-hexafluoropropylene) (P(VDF-*co*-HFP)) was used as a model polymer. The sample was dissolved in acetone and deposited in a copper cup placed on the TD/Py heating plate. The released compounds were ionized using a DART source and analyzed with an Orbitrap mass spectrometer.

The effects of eight parameters related to both DART ion source and TD/Py device were evaluated. The sum of the signal-to-noise ratios (S/N) of the three most abundant polymer-related ions was used as the response variable, as it provides a robust and selective indicator of analytical sensitivity, by maximizing the signal relative to background noise.

A Plackett-Burman screening design was first employed to identify the most influential parameters. Subsequently, a Doehlert experimental design was applied to optimize these parameters in order to maximize analytical sensitivity. Data analysis was performed using Statistica software.

4 Results and discussion

TD/Py-DART HRMS analysis of P(VDF-*co*-HFP) revealed two distinct steps the along temperature program: thermal desorption (35–300 °C) and pyrolysis (400–600 °C). As these two steps generated distinct and characteristic ion patterns, the responses were evaluated separately.

The Plackett-Burman screening design indicated that the response in the thermal-desorption region was mainly influenced by gas flow dynamics and sample positioning. In pyrolysis region, only the temperature ramp rate was significant.

Based on these results, two separate four-variable Doehlert designs were implemented to account for a Boolean factor corresponding to ceramic cap aperture of 0.5 and 2.5 mm. In the pyrolysis region, the temperature ramp rate was the only parameter significantly affecting the response, with an optimal value of 50 °C.min^{-1}. Under these conditions, sensitivity gains of 430% and 126% for ceramic cap apertures of 0.5 and 2.5 mm, respectively, compared to the initial instrumental settings.

5 Conclusion

The DOE-based approach enabled improved understanding and optimization of TD/Py-DART operating parameters. These results demonstrate the strong potential of this methodology for the sensitive analysis of trace-level fluoropolymers in contaminated environmental matrices such as soils.

6 References

[1] B Ameduri. The promising Future of Fluoropolymers. *Macromolecular Chemistry and Physics, 2020, 221 (8)*, pp. 1900573

[2] R.B. Cody. Thermal desorption and pyrolysis direct analysis in real time mass spectrometry for qualitative characterization of polymers and polymer additives. *Rapid Communications in Mass Spectrometry, 2020, 34*, p. e8687

[3] P. Pacholski. Capability of thermodesorption/pyrolysis DART FT-ICR MS to distinguish fluoropolymers and identify blend composition. *Journal of Analytical and Applied Pyrolysis, 2025*, p. 107361

Développement d'une stratégie multi-analytique et chimiométrique pour la traçabilité et l'évaluation de la sécurité des plastiques recyclés dans des applications sensibles

L. Senanou[1] P-M. Nguyen[2] D. Goujot[3] A. Tonda[4] P. Cardinael[5] S. Domenek[6]

[1] Laboratoire National de Métrologie et d'Essais, 78190 Trappes, leane.senanou@lne.fr

[2] Laboratoire National de Métrologie et d'Essais, 78190 Trappes, phuong-mai.nguyen@lne.fr

[3] UMR SayFood, AgroParisTech INRAE, 91120 Palaiseau, daniel.goujot@agroparistech.fr

[4] UMR MIA Paris-Saclay, AgroParisTech INRAE, 91120 Palaiseau, alberto.tonda@inrae.fr

[5] Laboratoire SMS, Université de Rouen, 76000 Rouen, pascal.cardinael@univ-rouen.fr

[6] UMR SayFood, AgroParisTech INRAE, 91120 Palaiseau, sandra.domenek@agroparistech.fr

Mots clés : plastiques recyclés, traçabilité, prétraitement du signal, analyse multi-blocs, machine learning.

1 Introduction

La transition vers une économie circulaire, pilotée par des réglementations telles que la loi française AGEC et le règlement européen PPWR, impose une intégration massive de plastiques recyclés dans des applications sensibles (emballage alimentaire, cosmétique). Cependant, garantir la sécurité et l'authenticité de ces matériaux reste un défi majeur en raison de la variabilité des gisements de déchets et des procédés de recyclage (mécanique, chimique, hybride) [1]. Les méthodes actuelles de contrôle qualité manquent souvent de robustesse et de transférabilité. Ce projet de thèse vise à développer une approche analytique intégrée pour sécuriser la traçabilité des matériaux recyclés.

2 Matériels et méthodes

Un jeu d'échantillons diversifié de rPET, représentatif de diverses origines (post-consommation, post-industriel) et procédés de recyclage, est collecté via le projet collaboratif Twinloop. Les échantillons font l'objet d'une caractérisation multi-analytique : GC-MS (pour les composés organiques volatils/semi-volatils et contaminants potentiels) et spectroscopie vibratoire (FTIR, Raman) pour l'identification de la matrice polymère et des marqueurs de vieillissement. Pour garantir la robustesse, des échantillons de contrôle qualité sont analysés régulièrement. Pour le profil volatil, une stratégie d'harmonisation avancée est envisagée. Elle repose sur un prétraitement des données par des algorthimes pour les rendre propres [2,3], suivi d'une correction des effets de lots par des algorithmes dédiés (WaveICA [4], PARSEC [5]). Ces flux de données, une fois stabilisés individuellement, sont ensuite intégrés dans des modèles de fusion multi-blocs et de classification supervisée (PLS-DA, Random Forest) pour corréler les empreintes chimiques avec les métadonnées des procédés [6,7].

3 Résultats et discussion

Les travaux actuels se concentrent sur la construction d'une base de données d'empreintes chimiques standardisée. L'enjeu majeur réside dans la validation de la méthodologie d’acquisition et de traitement des données pour garantir que les données chromatographiques soient propres et comparables entre différents instruments/laboratoires et dans le temps. Une fois ces données stabilisées, des stratégies d'analyse multi-blocs ou de fusion de données seront appliquées pour corréler les signatures chimiques (marqueurs de vieillissement, contaminants) avec l’authenticité du matériau. Cette étape est prérequis indispensable pour l'établissement de modèles prédictifs fiables de l'innocuité.

4 Conclusion

Cette thèse vise à fournir un outil chimiométrique polyvalent capable d'authentifier les plastiques recyclés et de prédire leur profil de sécurité transférable pour l’industrie. En combinant des données multi-analytiques avec des stratégies avancées d'apprentissage automatique, cette thèse contribuera au développement de méthodes de diagnostic standardisées et automatisées, facilitant l'usage sécurisé des matériaux recyclés dans les filières réglementées.

5 Références

[1] C. Saldaña-Pierard, P-M. Nguyen, F. Debeaufort, O. Vitrac, R. Auras. Impact of Emerging Packaging Regulations on International Trade and Product Safety with Emphasis on Plastic Reuse and Recycling in Europe and North America. *Journal of Industrial Ecology* 29, n^{o} 5, 2025, pp. 1473-504.

[2] X. Fan, Z. Xu, H. Zhang, et al. Fully automatic resolution of untargeted GC-MS data with deep learning assistance. *Talanta,* 244, 2022.

[3] Y. Fan, C. Yu, H. Lu, et al. Deep learning-based method for automatic resolution of gas chromatography-mass spectrometry data from complex samples. *Journal of Chromatography A,* 1690, 2023.

[4] K. Deng, F. Zhang, Q. Tan, et al. WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. *Analytica Chimica Acta,* 1061, 2019, pp.60-69.

[5] E. Salanon, B. Comte, D. Centeno, S. Durand, J. Boccard, E. Pujos-Guillot. Improving metabolomics data comparability without long term quality controls using a post-acquisition correction strategy. *Analytica Chimica Acta*, 1380, 2025.

[6] R. Peñalver, C. Marín, N. Arroyo-Manzanares, N. Campillo, P. Viñas. Authentication of recycled plastic content in water bottles using volatile fingerprint and chemometrics. *Chemosphere,* 297, 2022.

[7] A. Zappi, A. Biancolillo, N. Kassouf, et al. Quantification of Recycled PET in Commercial Bottles by IR Spectroscopy and Chemometrics. *Analytica* 5, 2, 2024, pp. 219-32.

CHIMIOMÉTRIE XXVème
17-19 FÉVRIER 2026
CAMPUS GRANVILLE, NANCY

Automated histological segmentation of hair follicles via hierarchical PLS-DA and FTIR hyperspectral imaging

J. Poulain[1], J. Avila[2], C. Sandt[3], S. Roussel[1], T. Bornschlögl[2]

[1] Ondalys, 8, av de l'Europe, 34830 CLAPIERS, jpoulain@ondalys.fr

[2] L'Oréal campus Aulnay : 1 Avenue Eugène Schueller, 93600 Aulnay -sous-Bois, jocasta.avila@loreal.com

[3] Synchrotron Soleil, Ligne SMIS, synchrotron SOLEIL, RD128, Saint Aubin, France, sandt@synchrotron-soleil.fr

Keywords: Hyperspectral imaging, FTIR, hierarchical PLSDA, image segmentation, follicles

1 Introduction

The hair follicle is a complex, multi-compartment mini-organ responsible for the production of hair fibers. It can be divided into anatomically and biologically distinct regions: the dermal papilla (DP), the hair cortex that arises due to cell differentiation of keratinocytes situated in the matrix, along the transition zone (TZ) and the pre-cortex, the outer root sheath (ORS), inner root sheath (IRS), and the conjunctive tissue sheath (CTS). Recent advancements in infrared (FTIR) microscopy have established this label-free technique as an important tool for characterizing *ex-vivo* follicles, evolving from structural identification to the detection of biochemical markers and energy-storing carbohydrates [1]. This study aims to automate the identification of these tissues using hyperspectral imaging to better assess the impact of cosmetic active ingredients on the follicle health.

2 Material and methods

Transmission-mode FTIR microspectroscopy was performed on 49 hair follicles using an Agilent Cary 620 microscope coupled to a Cary 670 spectrometer (3900–800 cm^{-1}), located at the SOLEIL Synchrotron Facility. To capture the full morphological extent, spectral images were acquired as concatenated mosaics (up to 384 × 128 pixels), enabling chemical mapping across the longitudinal structure. Data were baseline-corrected using Air-PLS [4], and follicles were segmented from the resin background via Particle Size Analysis (MIA Toolbox, Eigenvector Research Inc., USA).

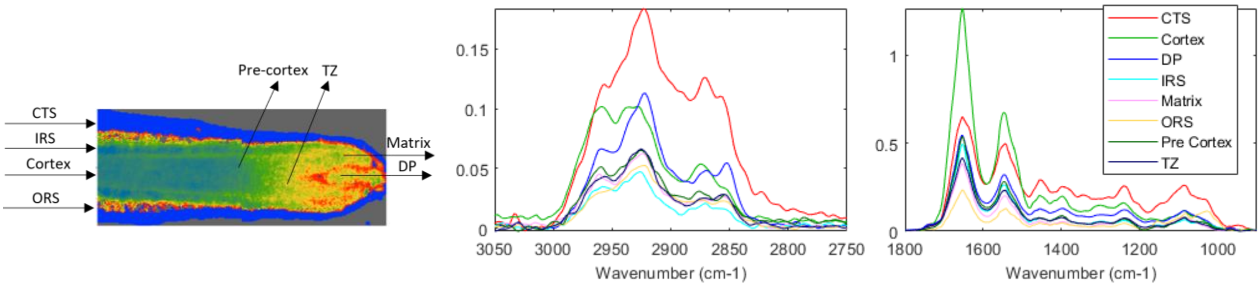


Figure 1 – Theoretical segmentation of hair follicles into 8 tissues and corresponding spectral profiles.

For model construction, reference pixels were manually selected based on the theoretical anatomical location of the 8 target tissues (Figure 1). Nine control samples were randomly selected for the calibration set, and validation was performed using block cross-validation per follicle. A **PLS-DA** [3] **double hierarchical approach** was implemented: the first level discriminates k classes versus all others, followed by a second model for the remaining classes. The double hierarchy corresponds to the fact that for each PLSDA model, the class is attributed hierarchically by first examining the probability of belonging to a class A, then to class B, and so on, until all classes have been checked in a specific order. This process is particularly useful for pixels with a high probability of belonging to multiple classes. Pixels belonging to none of the classes are excluded.

3 Results and discussion

Although eight tissues were initially targeted, the biological continuum of cell differentiation [4] and significant spectral overlap led to the merging of two adjacent tissues, resulting in a 7-class model with high coherence across different samples (Figure 2).

Occasional pixel misassignments and class spreading were observed, which can be attributed to several factors: the inherent difficulty of obtaining perfectly aligned longitudinal sections, the lack of a true reference method for pixel selection, and the high chemical similarity between transitional tissues. These factors contribute to a localized model error, particularly at tissue boundaries. To further elucidate the biochemical drivers of this segmentation, the interpretation of discriminant coefficients is currently underway.

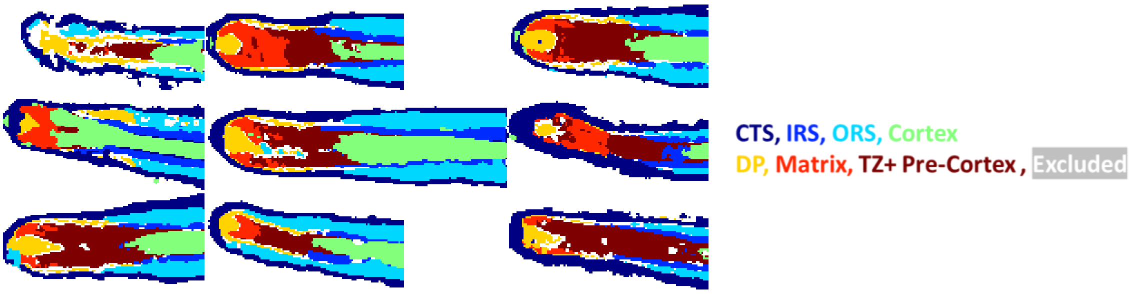


Figure 2 – Identification of 7 tissues in some follicles.

4 Conclusion

Despite the variability between hair follicles (biological samples, donors, sectioning heterogeneity, measurement campaigns) and the lack of true label-free references, multivariate data analysis based on hierarchical discrimination models has demonstrated its great suitability for the automatic segmentation of the different hair follicle tissues.

This methodology provides a robust framework for future studies evaluating the molecular impact of cosmetic active ingredients on hair growth.

5 References

[1] Sandt, C. and Bildstein, L. and Bornschlögl, T. and Baghdadli, N. and Thibaut, S. and Fazzino, P. and Borondics, F. Spectral histology of hair and hair follicle using infrared microspectroscopy. Internation Journal of Cosmetic Science 2024, *46*, 949-961.

[2] Z Zhang, S Chen & Liang Y. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, issue 5, 2010.

[3] M Barker & Rayens W. Partial least squares for discrimination. *Journal of the Chemometrics Society*, volume 17, issue 3: 166–173, 2003.

[4] Tissot N., Genty G., Santoprete R., Baltenneck F., Thibaut S., Michelet J., Sequeira I., Bornschlögl T. Mapping cell dynamics in human ex vivo hair follicles suggests pulling mechanism of hair growth. *Nature Communications* 2025, *16*, 10267.

Machine Learning interpretability methods applied to calibration models developed on Near Infrared spectroscopic data

A. Malechaux[1], J. Poulain[1], S. Roussel[1]

[1] Ondalys, 8, av de l'Europe, 34830 CLAPIERS, amalechaux@ondalys.fr

Keywords: Machine Learning, Explainable AI, Interpretability, SVM, ANN

1 Introduction

In the past decades, Machine Learning (ML) models have become more and more complex, leading to improvements in their predictive performance. However, these models can often be described as "black boxes", in the sense that it is very difficult to explain how results are obtained by a model from the input data. As complex Machine Learning models are increasingly used to make decisions, for instance in industrial or medical applications, there is a growing need to improve their interpretability in order to have greater confidence in their results [1], providing the so-called Explainable AI (Artificial Intelligence).

2 Material and methods

This study is focused on the interpretability of Machine Learning models after model calibration on near-infrared spectroscopic data, a.k.a. the "post-hoc explanation" of ML models.

Several regression models such as Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Extreme Gradient Boosting (XGBoost), have been trained and compared to the classical PLS calibration models.

The study is applied to the freely available Benchmark dataset consisting of Near Infrared spectra of minced meat samples recorded with the FOSS Tecator Infratec Food and Feed Analyzer in the 850-1050nm wavelength range [2].

Various ML model interpretability algorithms currently applied to non-spectroscopic data have been tested and compared: Local Interpretable Model-agnostic Explanations [3], Shapley Additive Explanations [4] and pseudo-samples prediction [5]. They have also been compared to the regression coefficients of an intrinsically interpretable Partial Least Squares model. While the applicability to spectroscopic data of pseudo-samples prediction has been demonstrated [5], Local Interpretable Model-agnostic Explanations and Shapley Additive Explanations are theoretically better suited to uncorrelated variables [6].

3 Results and discussion

The results indicate that the outputs of the interpretability methods tested appear to be in good agreement with the Partial Least Squares regression coefficients. Thus, they allow the identification of wavelengths ranges used by "black box" Machine Learning models. Moreover, they could be used as indicators of the risk of overfitting for complex models developed on spectroscopic data.

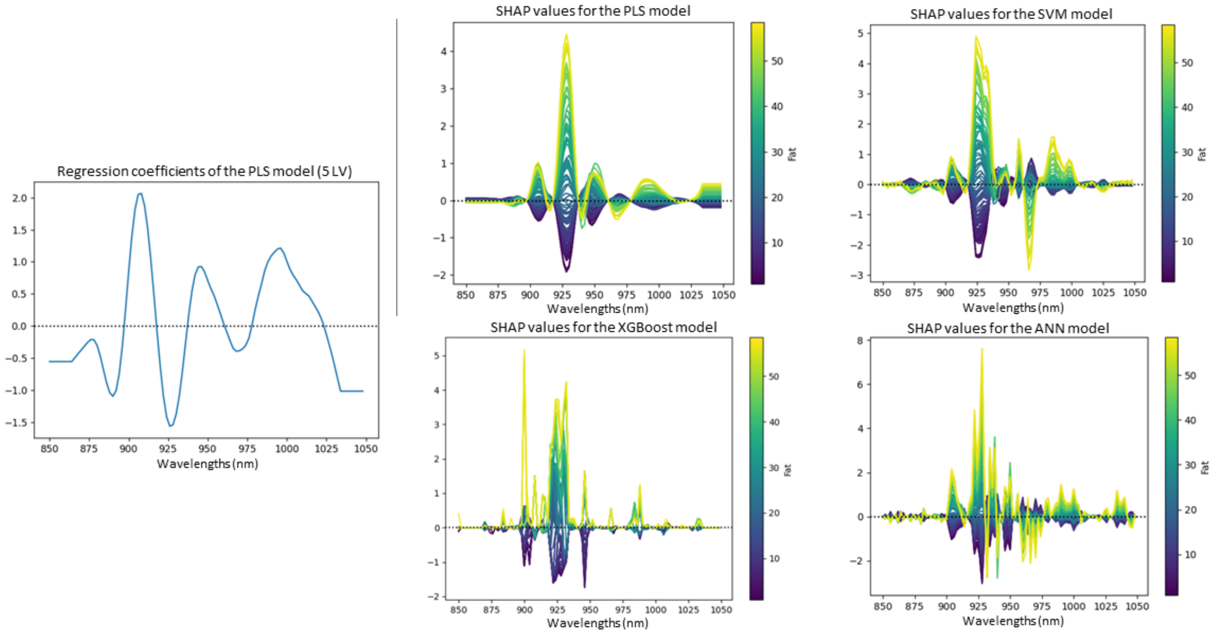


Figure 1: Example of results obtained on the calibration spectra with the Shapley Additive Explanations method for different Machine Learning models

4 Conclusion

In conclusion, this study shows the potential for applying different Machine Learning model interpretability methods to Near Infrared spectroscopic data. Further work on different NIR datasets will be conducted to confirm these results.

5 References

[1] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A Review of Machine Learning Interpretability Methods, Entropy, 2021, 23, 18.

[2] H. H. Thodberg, Tecator meat sample dataset, StatLib Datasets Archive, 1995.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, 1135-1144.

[4] S. M. Lundberg, S. Lee, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, 2017, 30, 4765-4774.

[5] G. J. Postma, P. W. T. Krooshof, L. M. C. Buydens, Opening the kernel of kernel partial least squares and support vector machines, Analytica Chimica Acta, 2011, 705(1-2), 123-134.

[6] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, G. Menegaz, A perspective on explainable artificial intelligence methods: SHAP and LIME, Advanced Intelligent Systems, 2024, 2400304.

Analysis of temporal metabolomics data by combining both ASCA and tensor decomposition

M. Bonsens[1] M. Moyon[2] M. Mahieu[1] M. Galharret[1] Mme. Alexandre-Gouabau[2] M. Hanafi[1]

[1] ONIRIS, INRAE, StatSC, 44300 – Nantes (matthieu.bonsens@oniris-nantes.fr, benjamin.mahieu@oniris-nantes.fr, jean-michel.galharret@oniris-nantes.fr) , [2] Nantes Université, INRAE, UMR1280 PhAN, 44000 – Nantes (Thomas.Moyon@inrae.fr, Marie-Cecile.Alexandre-Gouabau@univ-nantes.fr)

Keywords: ASCA, PARAFAC, PARALIND, interaction, time-series, metabolomics, tensor.

1 Introduction, motivations and objectives

As in many other fields, the use of time-series data is becoming increasingly common in metabolomics [5], particularly for describing complex metabolic responses to a treatment factor, but also for describing the temporal evolution of these metabolic responses. In the context of analysing metabolomics data with a temporal component, one of the characteristics is that a few time points are available, ASCA [1] is considered as one of the most frequently used methods, particularly because it takes into account the experimental design by decomposing the data by effect and analysing these effects separately using PCA [6]. The starting point for this paper is to highlight the limitations of ASCA, not in terms of its principle of decomposing data by effect, but particularly in terms of its use of different separate PCAs to analyses effect matrices and the interaction between main effects. These limitations will be highlighted by reformulating PCA as constrained tensor decomposition strategies applied to effect matrices reorganised as tensors. The value of adapting and exploiting alternative tensor decomposition methods then becomes apparent, especially as they are rarely used in this context.

This paper discuss updating the various PCAs considered in ASCA with alternative tensor decomposition methods, particularly PARALIND [2] and PARAFAC [3] (PARAFASCA [4]) which combine ASCA and PARAFAC. The contributions in terms of both methodology and interpretation of results will be illustrated using two data sets from the ANR GDM_MILK project.

2 Material, methods and results

Two datasets from the ANR GDM_MILK project will be used. The first (MILK) represents the composition of rat breast milk (35 rats) over time (4 time points) according to different (3 groups) dietary groups. The second dataset (Descendance) represents the composition of the blood plasma of rat offspring (48 rats) over time (4 time points) according to different dietary groups (4 groups). Analytical acquisition techniques were used: for the MILK dataset, SM and GC-FID measured metabolites (GC-FID was used for fatty acids only), and for the Descendance dataset, metabolites were measured by LC-MS/MS and FIA-MS/MS.

ASCA divide the total variation into factors contribution: 4,4% for the factor A (Diet), 28,9% for the factor B (Period) and 8,5% for the interaction between those factors. Since the interaction matrix is the matrix that allows describing the temporal trajectory of groups discrimination, among others it is necessary to focus on the analysis of the interaction matrix.

The principle variation of the interaction represented in ASCA (on 2 components) show 60% of the total variation interaction matrix (Figure 1). So, tensor models as PARALIND and PARAFAC applied to interaction fits 2 components well: the variations of the interactions represented by PARAFAC and PARALIND shows respectively 84,4% and 82,9% of the total interaction variation (Figure 2). This show that tensor decomposition methods are optimal, in term of explain variance, for the analysis of the interaction tensor.

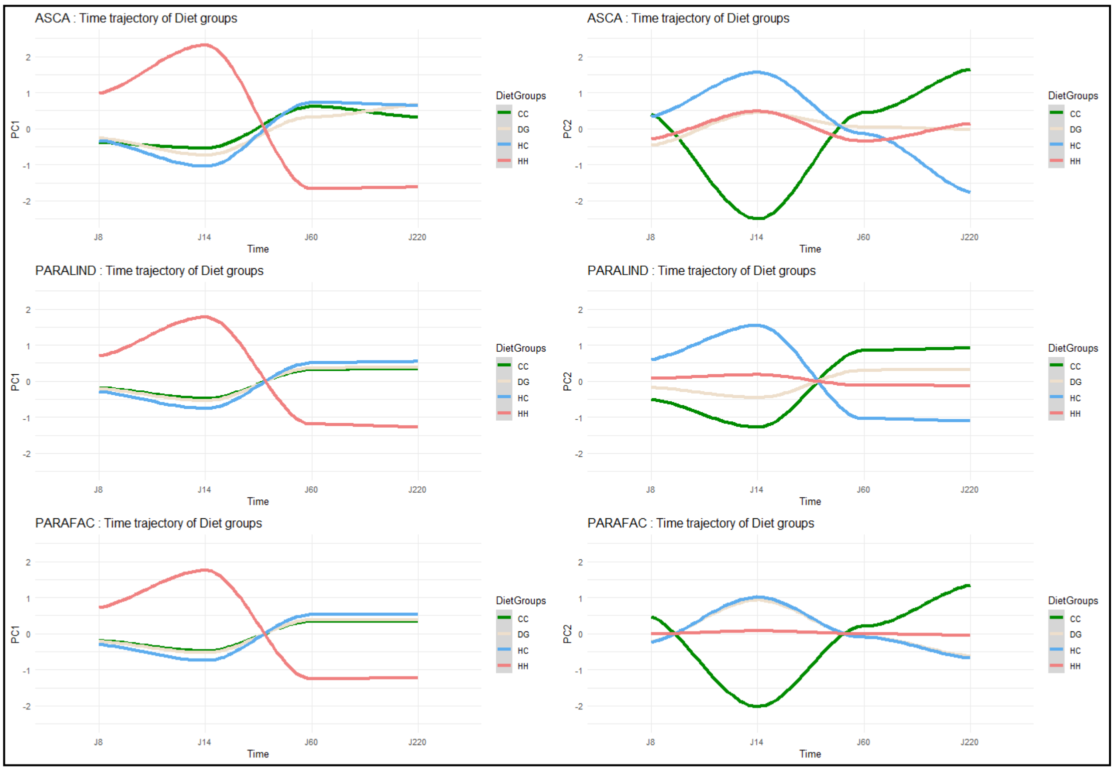


Figure 1 – ASCA, PARAFAC and PARALIND trajectory results on Descendance dataset

On the first component, PARAFAC and PARALIND shows the same time discrimination between the different groups, but on second component, the discrimination shows a different dynamic between those three methods. Moreover, PARALIND might be the better method to show the trajectory discrimination between the groups (particularly the DG’s group).

3 References

[1] Bertinetto Carlo, Engel Jasper, Jansen Jeroen, 6 octobre 2020, ANOVA simultaneous component analysis : A tutorial review, Analytica Chimica Acta : X n°6

[2] Bro Rasmus, Harshman Richard A., Sidiropoulos Nicholas D., Lundy Maragaret E., 12 janvier 2009, Modeling multi-way data with linearly dependent loadings, Journal of Chemometrics

[3] Bro Rasmus, 8 mars 1997, PARAFAC. Tutorial and application., Chemometrics and Intelligent Laboratory Systems n°38, p.149-171

[4] Jansen Jeroen J., Bro Rasmus, Hoefsloot Huub C. J., van den Berg Frans W. J., Westerhuis Johan A., Smilde Age K., 10 janvier 2008, PARAFASCA : ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data, Journal of Chemometrics n°22, p.114-121

[5] Smilde Age, Bro Rasmus, Geladi Paul, 2004, Multi-Way Analysis – Applications in the Chemical sciences, Chichester: Wiley

[6] Wold Svante, Geladi Paul, Esbensen Kim, Öhman Jerker, 1987, Multi-way principal components-and PLS-analysis, Journal of Chemometrics vol.1, p.41-56

Mid-level data fusion of hyperspectral images

B. Jaillais

INRAE, TRANSFORM, Nantes, France. Benoit.Jaillais@inrae.fr

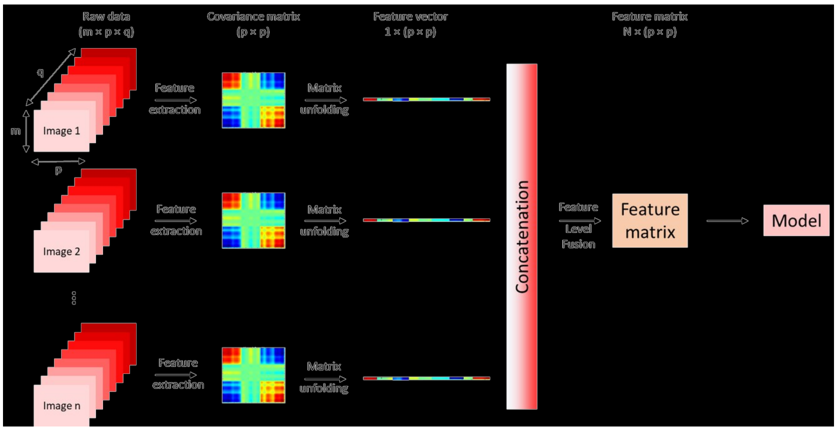
Keywords: data fusion, hyperspectral imaging, .

1 Introduction

Processing a batch of hyperspectral images is relatively complex due to the large number of spectral pixels. To process these images in the same factorial space, the images must be unfolded and the spectral pixels concatenated, resulting in a matrix with a very large number of rows. Furthermore, integrating this data in the form of a horizontal multiblock is impossible due to the mismatch between pixels from one image to another.

Principal component analysis (PCA) on a collection of images can be performed in two ways: either a representative number of pixels are extracted from each image and concatenated to create a PCA model, onto which all pixels are projected, or the variance-covariance matrix of each image is calculated, which, after accumulation and diagonalisation, leads to the obtaining of scores.

Here, a new way of obtaining scores based on the fusion of intermediate-level data is introduced, involving a step of creating synthetic variables or ‘features’. In order not to lose any information contained in the hyperspectral image, the variance-covariance matrix is calculated for each image and unfolded into a vector. Thus, a hyperspectral image corresponds to a new variable or feature that completely characterises the sample. The features of all the images are concatenated into a matrix, on which PCA is performed. (Fig 1).



2 Material and methods

Samples have been described elsewhere [1].

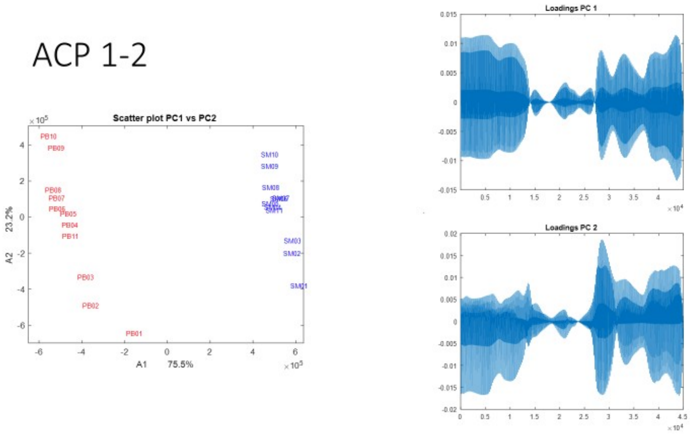
Two types of commercial biscuits (coded B1 and B2) were selected for this study and ten biscuits from each brand from the same pack were conditioned in 10 desiccators, each of them containing a different saturated salt solution, leading to different water activities (Aw) ranging from 0.114 to 0.907 (coded from 01 to 10). Two biscuits (one for each brand) were chosen as references and placed in a plastic bag in the lab where the hygrometry level was measured between 0.4 and 0.5. They were coded 11.

Biscuits were analyzed by NIR-HIS, consisting of a pushbroom hyperspectral SWIR camera (Burger Metrics, SIA, Riga, Latvia). The integration time was set to 1921 ms and spectral range of this system is between 950 and 2500 nm, and the spectral resolution is 7 nm.

PCA was performed on the feature matrix and scores were plotted in figure 2.

3 Results and discussion

The PC1-PC2 factorial plot shows a clear separation of cookies on PC1 based on cookie type, i.e., brand. In addition, axis 2 is characteristic of the water content of foods (cookies placed in desiccators with high water content are located on the positive side of the axis). Loadings may seem complicated, but their interpretation will be explained in the presentation.



4 Conclusion

This method is particularly effective because the separation between the two types of cookies was significantly more accurate than in published studies [1]. In addition, this method is particularly intuitive for the user and innovative in terms of data fusion. The next step will be to compare the prediction of content based on these new features.

5 References

[1] E. Lancelot, P. Courcoux, S. Chevallier, A Le-Bail, B. Jaillais Prediction of water content in biscuits using near infrared hyperspectral imaging spectroscopy and chemometrics. Journal of Near Infrared Spectroscopy. 2020;28(3):140-147. doi:10.1177/0967033520902538

Rapid identification of bacterial spores by Raman and OPTIR infrared spectroscopy using chemometric and machine learning

E. SARKEES[1,2] P. WINCKLER[1,2] J.M. Perrier-Cornet[1,2]

[1] Université Bourgogne Europe, Institut Agro Dijon, INRAE, UMR PAM, Dijon, France.

[2] Dimacell Imaging Facility, Institut Agro Dijon, INRAE, INSERM, Université Bourgogne Europe, Université Marie & Louis Pasteur, Dijon, France

Keywords: Bacterial spores, Raman spectroscopy, Infrared spectroscopy, OPTIR, Chemometrics, Food safety, Bacillus species, PCA, Machine learning classification

1 Introduction

The shift toward minimally processed foods with reduced preservative content increases susceptibility to microbial contamination and can shorten shelf life. Bacterial spores are particularly problematic because of their high resistance to heat, desiccation, and chemical stress, making them critical targets for rapid food-quality control. Conventional detection relies on slow culture-based methods (24–72 h), while molecular assays (e.g., qPCR) require predefined targets, sample preparation and are not inherently label-free. Here, we investigate O-PTIR microspectroscopy enabling simultaneous infrared and Raman measurements, combined with PCA and machine-learning classification, for rapid, non-destructive identification of six *Bacillus* spore strains.

2 Theory

Optical photothermal infrared (O-PTIR) microscopy uses a tunable mid-IR pump to excite molecular vibrations; absorption generates localized a refractive-index change that modulates a co-aligned visible probe, enabling IR spectra/chemical maps with visible-light spatial resolution (~0.3–0.4 µm). A Raman spectrum can be collected through the same objective using the visible laser probe and co-registered with the IR signal from the same pixel [1,2]. Bacillus endospores consist of a dehydrated core surrounded by an inner membrane, cortex (peptidoglycan), and multilayer protein coats. The core contains Ca^{2+}-dipicolinic acid (Ca-DPA; typically, 5–15% of spore dry weight) and small acid-soluble proteins (SASPs) that bind/protect DNA and contribute to resistance. While Ca-DPA is conserved, variations in coat composition, proteins, lipids, and surface polysaccharides can drive discriminative IR/Raman signatures across strains [3–5].

3 Material and methods

Six spore-forming strains were analyzed, spanning two major phylogenetic clusters: the *Bacillus subtilis* group (PS533, *B. licheniformis*, *B. pumilus*) and the *Bacillus cereus* group (*B. cereus* ATCC and KBAB4, classified as *B. weihenstephanensis*), with *Heyndrickxia coagulans* included as an additional spore-former. Purified spores were suspended in water and sparsely deposited on CaF_2. Co-localized IR and Raman spectra were simultaneously acquired from single spores using an O-PTIR microscope (≈0.3 µm lateral resolution), yielding ~200 spectra per strain collected over four days (~50/day). Spectra were baseline-corrected and vector-normalized. The **same 1D-CNN** (same architecture/training strategy) was applied three times, IR only, Raman only and IR+Raman ensemble. Robustness was assessed using an independent-day test (models trained on previous days and tested on a new acquisition day). Trained models were integrated into a GUI that loads the post-acquisition O-PTIR output file, runs preprocessing automatically, and returns predicted classes per spectrum/spore with an overlay visualization.

4 Results and discussion

Raman and IR regions can be broadly assigned to proteins, lipids, carbohydrates, and CaDPA. On the independent new-day test, IR-only achieved good but lower performance (≈ 85% accuracy) with more confusion. Raman-only achieved higher performance (≈ 90%) and the best class balance, while the IR+Raman ensemble provided the best overall performance (≈ 90%), confirming modality complementarity. Errors were dominated by phylogenetically close pairs: ATCC ↔ KBAB4 and Licheniformis ↔ Pumilus. In contrast, PS533 and H. coagulans were consistently well recognised (especially with Raman and the ensemble). The GUI enables direct post-acquisition prediction from O-PTIR files, supporting rapid end-to-end identification (**Figure 1**).

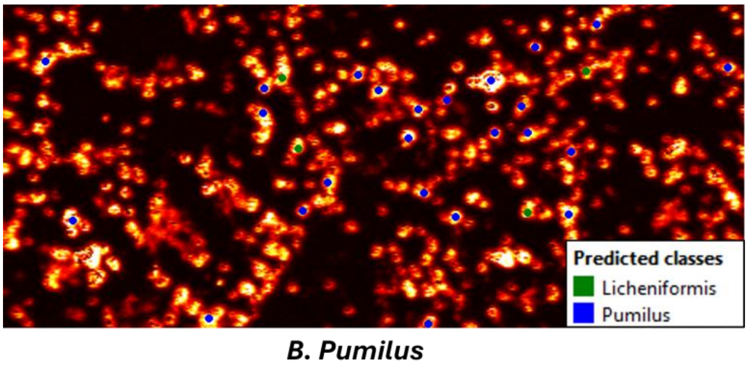


Figure 1. User interface for rapid identification (example: *B. pumilus*). Most detected spores are correctly labeled Pumilus (blue), with a few misclassified as Licheniformis (green).

5 Conclusion

Overall, O-PTIR microspectroscopy combined with a 1D-CNN enables rapid, label-free identification of spores, achieving ~85% accuracy with IR alone and ~90% with Raman and the IR+Raman ensemble on an independent-day test. Remaining errors mainly come from closely related strains (ATCC↔KBAB4, Licheniformis↔Pumilus). Increasing the number and day-to-day diversity of single-spore spectra should further improve robustness and strain-level discrimination.

6 References

[1] M. Kansiz et al. : Optical Photothermal Infrared Microspectroscopy with Simultaneous Raman – A New Non-Contact Failure Analysis Technique for Identification of <10 μm Organic Contamination in the Hard Drive and other Electronics Industries. Microsc, 28(3):26-36, 2020.

[2] C. B. Prater, M. Kansiz & J.-X. Cheng : A tutorial on optical photothermal infrared (O-PTIR) microscopy. APL Photonics, 9(9):091101, 2024.

[3] V. Bhandari, N. Z. Ahluwalia, S. H. Bajaj & R. S. Gupta : Molecular signatures for Bacillus species: demarcation of the Bacillus subtilis and Bacillus cereus clades in molecular terms and proposal to limit the placement of new species into the genus Bacillus. Antonie Van Leeuwenhoek, 103(6):1219-1235, 2013.

[4] G. C. Stewart : The Exosporium Layer of Bacterial Spores: a Connection to the Environment and the Infected Host. Microbiol. Spectr., 3(3), 2015.

[5] P. Scheldeman, L. Herman, S. Foster & M. Heyndrickx : Bacillus sporothermodurans and other highly heat-resistant spore formers in milk. J. Appl. Microbiol., 101(3):542-555, 2006.

[6] B. Setlow, S. Atluri, R. Kitchel, K.-D. Koziol-Dube & P. Setlow : Role of dipicolinic acid in resistance and stability of spores of Bacillus subtilis with or without DNA-protective alpha/beta-type small acid-soluble proteins. J. Appl. Microbiol., 101(3):549-558, 2006.

Deepchemometrics for food authentication, developing one-class modelling tools based on VAE and SIMCA framework for DOP cheese certification

F. Abdelghafour[1,2] D. Tanzilli[1,2] J.-M. Roger[1,2] M. Metz[1,3,4] A. Biancolillo[5]

[1] LabCom Aioly, Artificial Intelligence and Optics Laboratory, 34196, Montpellier, France, [2] ITAP, Univ. Montpellier, INRAE, Institut Agro, 34196, Montpellier, France, [3] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale Aix-Marseille Université, UMR CNRS IRD Avignon Université, Site de l'Etoile Marseille, France, [4] Pellenc ST, Applied Research Group, 84120, Pertuis, France

[5]Department of physical and chemical sciences, University of L'Aquila, Via Vetoio 67100, L'Aquila, Italy

florent.abdelghafour@inrae.fr

Keywords: Deepchemometrics, food authentication, SIMCA, VAE, One Class Modelling

1 Introduction

Product authentication is essential to ensure food safety and quality, and to protect high-value products such as DOP cheeses. Infrared spectroscopy combined with chemometrics enables rapid and non-destructive compositional fingerprinting. Classical classification approaches, such as PLS-DA[1], achieve high accuracy when all classes are known. However, this assumption rarely holds in operational contexts, where unknown sources of adulteration may arise. In such cases, one-class modeling is more appropriate, focusing on characterising the authentic product and rejecting non-conforming products [2]. SIMCA[3] is the standard method used in chemometrics, however it can face limitations for complex and nonlinear data distributions. Recent developments in deep learning offer potential to model spectral variability with greater flexibility. In this work, linear and deep approaches are compared, exploring CNN classification for baseline performance and one-class extensions using VAE-SIMCA[4] models.

2 Theory

VAE-SIMCA is designed to combine the statistical foundations of classical SIMCA with the representational flexibility of variational autoencoders (VAEs). In SIMCA, anomaly detection relies on distances defined in a PCA subspace: the score distance $\mathbf{T^2}$ measures variation within the modelled subspace, while the orthogonal distance $\mathbf{Q}$ quantifies deviations from it. Assuming approximately normally distributed scores and residuals, these statistics admit well-defined confidence limits for one-class modelling. A VAE is a generative neural network that enforces Gaussian prior through the Kullback–Leibler divergence, regularising the latent space such that $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. This is conceptually analogous to the orthonormalisation in PCA, leading to decorrelated and standardised latent coordinates. Consequently, distances in the latent space admit a Mahalanobis-like interpretation, theoretically compatible with SIMCA. The essential difference is the non-linear projection and latent representation of the VAE that captures complex data manifolds while preserving the framework.

3 Material and methods

The dataset consists of 2000 mid-infrared (MIR) spectra acquired at the University of L'Aquila (Italy) from three classes of cheese. Class 0 corresponds to "*Canestrato di Castel del Monte*", a traditional sheep cheese and the target class of interest. Class 2 corresponds to sheep cheeses produced in Tuscany, while Class 3 corresponds to soft sheep cheese produced in the same factory as Class 1. Each sample is described by 3500 spectral variables covering the range 3000–10 000 nm, with a spectral resolution of 2 nm. This work compares conventional linear chemometric methods with deep chemometric approaches for two tasks: (a) closed-set multiclass classification and (b) one-class modelling. In case (a), Partial Least Squares Regression (PLSR) is compared with two custom-designed Convolutional Neural Network (CNN) architectures. In case (b), the classical SIMCA method is benchmarked against variant approaches of VAE-SIMCA.

4 Results and discussion

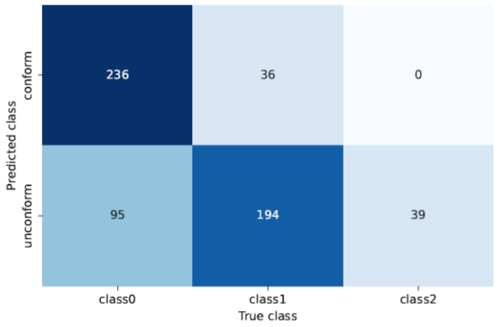


Figure 1: Confusion matrix obtained with optimal ALT-SIMCA

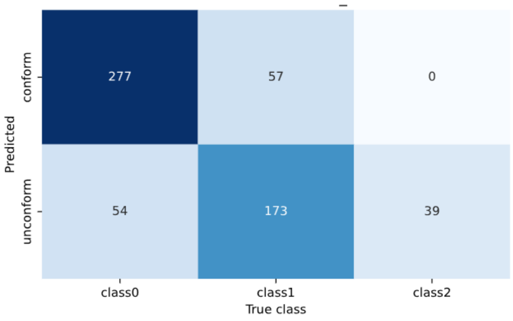


Figure 2: Confusion matrix obtained with the standard VAE-SIMCA framework

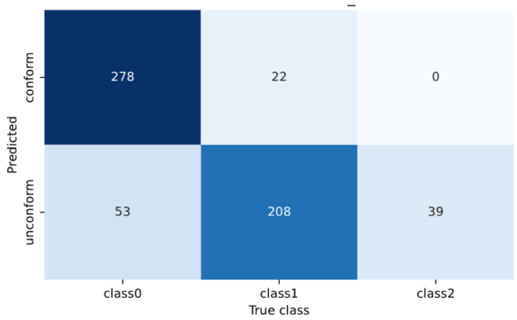


Figure 3 : Confusion matrix obtained with cosine variant of VAE-SIMCA

5 Conclusion

This work explores the potential and limitations of deep-learning–based one-class modelling, showing that complex nonlinear latent spaces can be leveraged without abandoning the rigorous statistical foundations that underpin classical chemometric methods.

6 References

[1] Indahl, U.G., Martens, H. and Næs, T. (2007). From dummy regression to prior probabilities in PLS-DA. J. Chemometrics, 21: 529-536. https://doi.org/10.1002/cem.1061

[2] Strani L., Cocchi M., Tanzilli D., Biancolillo A., Marini F., Vitale R. (2025). One class classification (class modelling): State of the art and perspectives, TrAC Trends in Analytical Chemistry, Volume 183, 118117, ISSN 0165-9936, https://doi.org/10.1016/j.trac.2024.118117.

[3] Wold S., Sjöström M. (1977). SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy, Chemometrics: Theory and Application. June 1, 243-282, DOI:10.1021/bk-1977-0052.ch012

[4] Petersen A., Kucheryavskiy S. (2025). VAE-SIMCA: Data-driven method for building one class classifiers with variational autoencoders, Chemometrics and Intelligent Laboratory Systems, Volume 256, 105276, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2024.105276.

Décomposition PARATUCK pour le démélange de spectres de fluorescence

Amir Ayadi[1], Xavier Luciani[1], Roland Redon[2]
[1] Université de Toulon, LIS, équipe SIIM, France, amir-ayadi@etud.univ-tln.fr, xavier.luciani@univ-tln.fr
[2] Aix Marseille Université, MIO, France, roland.redon@univ-tln.fr

Mots-clés : Spectroscopie de fluorescence, décomposition tensorielle, filtrage, TF, Paratuck.

1 Introduction

La spectroscopie de fluorescence d'excitation-émission (EEM) est une technique largement utilisée pour l'analyse de mélanges chimiques complexes en environnement, biologie et chimie analytique. Les données issues de cette technique sont organisées sous forme de tenseurs tridimensionnels (excitation × émission × échantillons), ce qui rend les méthodes de décomposition tensorielle particulièrement adaptées à leur analyse. La décomposition PARAFAC (CP) [1] est couramment employée en chimiométrie pour son unicité et son interprétabilité chimique. Toutefois, en pratique, la présence de décalages spectraux, et d'effets de filtrage d'un échantillon à l'autre viole souvent les hypothèses de multilinéarité stricte, dégradant l'estimation des spectres purs et des profils de concentration.

Afin de pallier ces limitations, plusieurs extensions du modèle CP ont été proposées, telles que PARAFAC2 [2]ou les modèles Shift-Invariant Soft Trilinearity (SIST [3]). Néanmoins, ces approches présentent parfois des limites en termes de flexibilité, de stabilité numérique ou de complexité algorithmique. L'objectif de ce travail est de proposer un cadre unifié, fondé sur une approche algébrique de type PARATUCK2 [4], permettant la prise en compte explicite des décalages et des effets de filtrage dans le domaine de Fourier.

2 Théorie

Un principe fondamental du traitement du signal est que la translation d'un signal dans le domaine direct correspond à une modulation de phase linéaire dans le domaine de Fourier, tandis qu'un filtrage linéaire se traduit par une multiplication fréquentielle. Ces propriétés permettent de reformuler le modèle dans le domaine de Fourier, où les décalages et les filtres sont représentés par des opérateurs diagonaux complexes.

Dans ce cadre, nous montrons que les données peuvent être décomposées par un modèle de type PARATUCK2, considéré ici dans un cas particulier structuré, où les interactions sont limitées à un seul mode, pouvant être celui des émissions ou des excitations. Ainsi, en supposant que le même effet de filtrage s'applique aux différents facteurs du mode considéré, le modèle dans le domaine de Fourier s'écrit pour une tranche du tenseur (échantillon k):

$$\tilde{X}_k = A D_k \tilde{B}^T \Lambda_k \qquad (1)$$

où D_k et Λ_k sont des matrices diagonales. A représente les spectres d'excitation (ou d'émission) D_k les concentrations relatives de l'échantillon k, $\tilde{B}$ la transformée de Fourier des spectres d'émission (ou d'excitation) et Λ_k le filtre linéaire appliqué à l'échantillon k.

3 Matériels et méthodes

La décomposition PARATUCK2 peut être calculée à l'aide d'un algorithme de type ALS que nous avons adapté à notre cas particulier. Les paramètres de décalage et de filtrage sont estimés à partir de régressions à coefficients complexes élément par élément.

Les performances de cette approche, nommée ici PARAfilt sont évaluées sur des données de fluorescence simulées, générées à partir de profils spectraux synthétiques, avec introduction de décalages spectraux, d'effets de filtrage différents pour chaque échantillon et d'un bruit Gaussien additif. Des simulations de type Monte-Carlo sont réalisées afin d'évaluer la robustesse et la précision des estimations.

4 Résultats et discussion

Le tableau ci-dessous résume les erreurs de reconstruction et d'estimation des facteurs pour différentes valeurs de SNR, L'approche proposée présente de meilleures performances, en particulier pour des niveaux de bruit moyens à élevés.

Tableau 1 - Erreurs de reconstruction et d'estimation des facteurs pour les méthodes PARAfilt, SIST et PARAFAC2 en fonction du SNR

	PARAfilt				SIST				PARAFAC2			
SNR	errX	errC	errA	errBk	errX	errC	errA	errBk	errX	errC	errA	errBk
20	9.55 10^{-2}	5.13 10^{-3}	6.31 10^{-3}	8.52 10^{-2}	9.60 10^{-2}	5.55 10^{-3}	3.68 10^{-3}	1.18 10^{-1}	9.17 10^{-2}	7.44 10^{-3}	8.44 10^{-3}	3.08 10^{-1}
40	9.59 10^{-3}	4.98 10^{-4}	6.23 10^{-4}	7.40 10^{-3}	9.70 10^{-3}	3.62 10^{-3}	3.89 10^{-4}	7.33 10^{-3}	9.21 10^{-3}	3.64 10^{-3}	8.60 10^{-4}	2.36 10^{-2}
80	9.61 10^{-5}	5.05 10^{-6}	6.07 10^{-6}	7.85 10^{-5}	1.05 10^{-3}	3.60 10^{-3}	4.46 10^{-6}	3.45 10^{-3}	1.11 10^{-4}	3.60 10^{-3}	1.69 10^{-4}	3.11 10^{-3}

5 Conclusion

Ce travail propose une approche algébrique de type PARATUCK2 en domaine de Fourier pour la prise en compte des décalages et du filtrage en spectroscopie de fluorescence. La méthode PARAfilt montre, sur données simulées, une amélioration de la robustesse par rapport aux approches existantes, ouvrant la voie à des applications sur des données réelles.

6 Références

[1] BRO, Rasmus. PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, 1997, vol. 38, no 2, p. 149-171.

[2] BRO, Rasmus, ANDERSSON, Claus A., et KIERS, Henk AL. PARAFAC2—Part II. Modeling chromatographic data with retention time shifts. Journal of Chemometrics: A Journal of the Chemometrics Society, 1999, vol. 13, no 3-4, p. 295-309.

[3] SCHNEIDE, Paul-Albert, GALLAGHER, Neal B., et BRO, Rasmus. Shift invariant soft trilinearity: Modelling shifts and shape changes in gas-chromatography coupled mass spectrometry. *Chemometrics and Intelligent Laboratory Systems*, 2024, vol. 251, p. 105155.

[4] HARSHMAN, Richard A. et LUNDY, Margaret E. Uniqueness proof for a family of models sharing features of Tucker's three-mode factor analysis and PARAFAC/CANDECOMP. *Psychometrika*, 1996, vol. 61, no 1, p. 133-154.

Extending the concept of essential information selection to three- and multi-way data analysis

R. Vitale[1], N. Omidikia[2], C. Ruckebusch[1]

[1] Univ. Lille, CNRS, LASIRE (UMR8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France, raffaele.vitale@univ-lille.fr, cyril.ruckebusch@univ-lille.fr

[2] University of Sistan and Baluchestan, Department of Chemistry, Faculty of Science, P.O. Box 98135-674, Zahedan, Iran, nematkia@gmail.com

Keywords: essential rows (ERs), essential columns (ECs), essential tubes (ETs), three-mode factor analysis, Parallel Factor Analysis-Alternating Least Squares (PARAFAC-ALS), convex polytopes.

1 Introduction

The identification and extraction of essential information (EI) from sets of multivariate measurements have recently garnered significant attention in the domain of bilinear curve resolution [1]. In this presentation, the idea of EI-based data reduction is extended to trilinear and multilinear datasets through the description of an original algorithmic procedure leveraging the principles of Higher Order Singular Value Decomposition [2, 3].

2 Material and methods

For the sake of simplicity, let $\underline{\mathbf{D}}$ be a generic three-way data array (tensor) of dimensions N rows × J columns × K tubes and effective rank F (*i.e.*, F is the actual amount of components/factors underlying $\underline{\mathbf{D}}$). The EI-based compression of $\underline{\mathbf{D}}$ basically boils down to recognising and isolating the rows, columns and tubes of this tensor that allow by themselves to reconstruct the entire space of the original measurements. The number of these so-called essential rows (ERs), essential columns (ECs) and essential tubes (ETs) is obviously lower compared to N, J and K, respectively, but the reduced version of $\underline{\mathbf{D}}$ - say $\underline{\mathbf{D}}_\mathrm{r}$ of size $N_\mathrm{r} \times J_\mathrm{r} \times K_\mathrm{r}$ - shares with it the same trilinear properties. Subsequently, the trilinear factorisations of $\underline{\mathbf{D}}$ and $\underline{\mathbf{D}}_\mathrm{r}$ (carried out by means of, *e.g.*, PARAFAC-ALS) yield, in principle, indistinguishable results from the perspective of self-modelling curve resolution. Operationally speaking, the ERs, ECs and ETs of $\underline{\mathbf{D}}$ can be retrieved through the following 5-step computational procedure:

1. $\underline{\mathbf{D}}$ is subjected to a Higher Order Singular Value Decomposition (HOSVD) that can be written as:

$$\underline{\mathbf{D}} = \underline{\mathbf{S}} \times_1 \boldsymbol{\mathcal{U}} \times_2 \boldsymbol{\mathcal{V}} \times_3 \boldsymbol{\mathcal{W}} \quad (1)$$

   where $\underline{\mathbf{S}}$ is a $(N \times J \times K)$-dimensional core tensor, $\boldsymbol{\mathcal{U}}$ $(N \times N)$, $\boldsymbol{\mathcal{V}}$ $(J \times J)$ and $\boldsymbol{\mathcal{W}}$ $(K \times K)$ carry the row, column and tube space factors of $\underline{\mathbf{D}}$, and $\times_1$, $\times_2$ and $\times_3$ connote the mode-1, mode-2 and mode-3 product, respectively;

2. $\boldsymbol{\mathcal{U}}$, $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{W}}$ are then truncated retaining (at least) their first F columns;

3. row, column and tube space HOSVD scores are calculated as:

$$\mathbf{X} = \boldsymbol{\mathcal{U}}^{*}\boldsymbol{\Sigma}_1 \tag{2}$$

$$\mathbf{Y} = \boldsymbol{\mathcal{V}}^{*}\boldsymbol{\Sigma}_2 \tag{3}$$

$$\mathbf{Z} = \boldsymbol{\mathcal{W}}^{*}\boldsymbol{\Sigma}_3 \tag{4}$$

   with $\boldsymbol{\mathcal{U}}^{*}$ ($N \times F$), $\boldsymbol{\mathcal{V}}^{*}$ ($J \times F$) and $\boldsymbol{\mathcal{W}}^{*}$ ($K \times F$) deriving from the truncation of $\boldsymbol{\mathcal{U}}$, $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ being the $F \times F$ diagonal matrices containing the first F singular values of the two-way arrays resulting from the row-wise, column-wise and tube-wise unfolding of $\underline{\mathbf{D}}$;

4. $\mathbf{X}$ ($N \times F$), $\mathbf{Y}$ ($J \times F$) and $\mathbf{Z}$ ($K \times F$) are normalised as:

$$\tilde{\mathbf{X}} = \mathbf{X} \oslash \mathbf{x}_1\mathbf{1}^{\mathrm{T}} \tag{5}$$

$$\tilde{\mathbf{Y}} = \mathbf{Y} \oslash \mathbf{y}_1\mathbf{1}^{\mathrm{T}} \tag{6}$$

$$\tilde{\mathbf{Z}} = \mathbf{Z} \oslash \mathbf{z}_1\mathbf{1}^{\mathrm{T}} \tag{7}$$

   where $\mathbf{x}_1$, $\mathbf{y}_1$ and $\mathbf{z}_1$ denote the first column of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, respectively, $\oslash$ represents the element-wise (Hadamard) division operator and $\mathbf{1}$ is a vector of ones of dimensions $F \times 1$. After normalisation, the first column of $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Z}}$ features only unit entries and is therefore removed before further processing. For this reason, the number of columns of $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Z}}$ becomes from now on equal to F-1.

5. ERs, ECs and ETs are finally obtained by identifying the subset of rows of $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Z}}$ supporting their corresponding (F-1)-dimensional convex hulls.

3 Results and discussion

The performance of this novel algorithm was evaluated in both real-world and simulated case-studies which permitted to highlight the benefits it can bring in domains like multiway fluorescence spectroscopy and imaging.

4 Conclusion

The inspection of the results obtained in the aforementioned case-studies led to three key conclusions:

- PARAFAC-ALS decompositions of full-size and reduced size-data are virtually indistinguishable;
- PARAFAC-ALS model quality and adequacy is preserved;
- PARAFAC-ALS decompositions of reduced-size data are considerably faster than PARAFAC-ALS decompositions of full-size data.

Notice that HOSVD can be exploited to factorise also arrays with more than three modes. Therefore, the developed reduction procedure can readily be extended to handle multiway datasets featuring multilinear structures.

5 References

[1] Ruckebusch, C., Vitale, R., Ghaffari, M., Hugelier, S. & Omidikia, N. Perspective on essential information in multivariate curve resolution. *Trend. Anal. Chem.* 132, 116044, 2020.

[2] Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*. 31, 279-311, 1966.

[3] Vitale, R., Azizi, A., Ghaffari, M., Omidikia, N. & Ruckebusch, C. Three-way data reduction based on essential information. *J. Chemometr*. 38, e3617, 2024.

Exploring the potential of non-linear chemometric approaches for olive oil quality assessment using NIR spectroscopy

M. Garrido-Cuevas[1] M. Lesnoff[2] M.T. Sánchez[3] D. Pérez-Marín[1]

[1] Departamento de Producción Animal, Grupo de Investigación ISAG, Unidad de Sensores NIR, ETSIAM, Universidad de Córdoba (Spain), g52gacum@uco.es, dcperez@uco.es

[2] CIRAD, UMR SELMET, Montpellier (France), matthieu.lesnoff@cirad.fr

[3] Departamento de Bromatología y Tecnología de los Alimentos, ETSIAM, Universidad de Córdoba (Spain), bt1sapim@uco.es

Keywords: Virgin olive oil quality, near infrared spectroscopy, non-linear modelling, open-source tools.

1 Introduction

Despite considerable efforts devoted to physico-chemical and sensory methods for evaluating the quality, purity, and authenticity of virgin olive oils (VOOs), the adulteration of VOOs with lower-quality oils remains a persistent international concern. One key reason is the limited number of samples officially analysed, restricted by reduced national budgets and the high cost of conventional methods endorsed by the International Olive Council (IOC) and European standards. Near Infrared Spectroscopy (NIRS) has shown strong potential for predicting quality parameters in olive oil [1, 2]. However, most chemometric approaches rely on linear models such as Partial Least Squares Regression (PLSR), which may not capture the complexity of real-world datasets. This opens the way for alternative modelling strategies, particularly non-linear techniques, which may improve predictive accuracy and robustness. The present work aims to explore and compare linear methods, known for their simplicity and interpretability, with non-linear methods, which are more flexible and better suited for modelling complex relationships.

2 Material and methods

In this study, a total of 1252 olive oil samples were analysed in transflectance mode (log (1/R)) using a NIRS^{TM} DS2500 monochromator (FOSS Analytical, Hillerød, Denmark). The instrument provides absorbance readings between 400 and 2500 nm, in 2 nm steps.

All samples were assessed by an accredited laboratory of Fundacion Citoliva (Jaen, Spain), following the official methods established in the Regulation (EU) 2022/2104 [3]. The physicochemical analysis included the evaluation of acidity (% oleic acid), peroxide index (meqO_2/kg), the extinction coefficients K_{232} and K_{270} (AU), and ethyl esters (mg/kg).

Prior to model development, samples were randomly split into training (70%) and test (30%) sets.

Linear methods such as PLSR were evaluated alongside several non-linear approaches, including Kernel PLSR (KPLSR), Kernel Ridge Regression (KRR), Support Vector Machine (SVM), k-Nearest Neighbours Locally Weighted PLSR (kNN-LWPLSR), kNN-LWPLSR with averaging, and Random Forest (RFR). These methods were compared to investigate the potential presence of non-linearity in the data and, consequently, to determine whether the use of non-linear methods improves the prediction of key physicochemical parameters. Model hyperparameters were optimized using grid search combined with 5-fold cross-validation to ensure a fair and robust comparison. Different data pretreatments were applied to the raw spectral data, including first and second derivatives, both with and without scatter correction. All models were trained independently for each parameter, and their performance was evaluated based on prediction errors calculated on the test sets. Statistical significance of the differences between model performances was assessed using Fisher's test. All data pre-processing and chemometric treatments were implemented using the Jchemo package in the Julia programming language [4].

3 Results and discussion

Table 1 shows the comparison of Root Mean Squared Error of Prediction (RMSEP) values obtained by both linear and non-linear models for each physico-chemical parameter.

Table 1 – RMSEP values comparison between PLSR and non-linear models for olive oil quality parameters.

#Parameter	PLSR	KPLSR	KRR	RFR	kNN-LWPLSR	kNN-LWPLSRavg	SVM
Acidity	0.32	0.32	0.32	0.46*	0.33	0.32	0.29
Peroxide index	3.79	3.79	3.57	3.52	3.73	3.48	3.59
K_{232}	0.34	0.40*	0.37	0.35	0.34	0.45*	0.33
K_{270}	0.03	0.03	0.03	0.03	0.03	0.03	0.05*
Ethyl esters	180.82	194.43	152.04*	144.19*	184.11	184.11	198.06*

* statistically significant differences compared to PLSR

Non-linear models showed performance comparable to PLSR for most physicochemical parameters, with only minor variations in RMSEP values. For acidity, peroxide index, K232, and K270, no statistically significant differences were observed. By contrast, for ethyl esters, RMSEP values obtained with KRR and RFR differed significantly from those of PLSR, as confirmed by Fisher's test. These findings suggest that non-linear modelling may be particularly beneficial for this parameter, likely due to its greater chemical complexity and non-linear spectral response.

4 Conclusion

PLSR confirmed its role as a robust baseline method for the prediction of olive oil physicochemical parameters, providing reliable and consistent performance. Non-linear approaches offered additional benefits only in specific cases, indicating that their effectiveness is strongly dependent on the parameter under study and the underlying complexity of its relationship with the spectral data.

5 References

[1] Arroyo-Cerezo, A., Yang, X., Jiménez-Carvelo, A.M., Pellegrino, M., Savino, A.F., Berzaghi, P. Assessment of extra virgin olive oil quality by miniaturized near infrared instruments in a rapid and non-destructive procedure. *Food chemistry, 430*, 2024, 137043.

[2] Garrido-Cuevas, M.M., Garrido-Varo, A.M., Sánchez, M.T., Pérez-Marín, D. Assessment of a portable near-infrared spectral sensor for enhancing virgin olive oil quality control. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 339, 2025, 126288.

[3] Commission Delegated Regulation (EU) 2022/2104 of 29 July 2022 Supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as Regards Marketing Standards for Olive Oil, and Repealing Commission Regulation (EEC) No 2568/91 and Commission Implementing Regulation (EU) No 29/2012. L 284, 1-22.

[4] Lesnoff, M. 2021. Jchemo: Chemometrics and machine learning for high-dimensional data with Julia. https://github.com/mlesnoff/Jchemo.jl. UMR SELMET, Univ Montpellier, CIRAD, INRA, Institut Agro, Montpellier, France.

Topological Data Analysis of OPTIR Infrared and Raman spectra for the study of UV-induced structural changes in bacterial spores

Romain Demelle[1], Laurence Dujourdy[2], Pascale Winckler[1,3], Jean-Marie Perrier-Cornet[1,3], Pierre-Yves Louis[1,4]

1 Univ. Bourgogne Europe, Institut Agro, INRAE, UMR PAM, 21000 Dijon, France, romain.demelle@institut-agro.fr

2 Institut Agro, LIB, Equipe Science des Données, 21000 Dijon, France, laurence.dujourdy@institut-agro.fr

3 Dimacell Imaging Facility, Institut Agro Dijon, INRAE, INSERM, Université Bourgogne Europe, Université Marie & Louis Pasteur, Dijon France, Pascale.winckler@institut-agro.fr, jean-marie.perrier-cornet@institut-agro.fr

4 Univ. Bourgogne Europe, CNRS, IMB, UMR 5584, 21000 Dijon, France, pierre-yves.louis@institut-agro.fr

Keywords: alternative proteins, infrared spectroscopy, functionality, bibliometrics, chemometrics, gelation; sustainability.

1 Introduction

Infrared (IR) and Raman spectroscopy are widely used to characterize biological materials, but the high dimensionality and non-linear structure of spectral data often limit the efficiency of classical chemometric approaches. This limitation is particularly pronounced for bacterial spores, whose biochemical organization is highly robust and whose response to external stressors, such as ultraviolet (UV) irradiation, can be subtle and heterogeneous.

Topological Data Analysis (TDA) [1,2] provides an alternative framework by focusing on the shape and structure of data rather than on pointwise spectral variations. Through persistent homology, TDA extracts multiscale topological descriptors that are robust to noise and continuous deformations [3]. In this work, TDA is evaluated as a chemometric tool for the analysis of IR and Raman spectra of bacterial spores, with the aim of discriminating spore strains and quantifying UV-induced structural modifications.

2 Theory

Persistent homology [4-6] is a central tool of Topological Data Analysis that characterizes the evolution of topological features, such as connected components (H_0) and loops (H_1), across a filtration parameter. Applied to spectral data, persistent homology captures multiscale structural patterns that are robust to noise and continuous deformations.

The output of persistent homology is a persistence diagram, which summarizes the birth and death of topological features along the filtration. Quantitative comparisons between spectra are performed by computing distances between persistence diagrams. In this study, two complementary metrics are used: the Bottleneck distance, which emphasizes the largest topological discrepancies and highlights dominant local changes, and the Wasserstein distance, which integrates contributions from all features and reflects more global structural differences. These distances provide a compact and interpretable representation of topological dissimilarities between spectra [7,8].

3 Material and methods

Three Bacillus spore strains were investigated: the wild-type strain PS533 and two mutant strains, PS4150 and FB122. For each strain, non-treated (NT) spores and UV-treated spores were analyzed using infrared (OPTIR) and Raman spectroscopy. UV treatments included two irradiation wavelengths (UVB and UVC) applied under dry and aqueous conditions.

Each dataset consisted of multiple spectral measurements per replicate. Aberrant spectra were detected and removed at the level of individual measurements using robust distance-based criteria prior to averaging. Persistence diagrams were computed using the Ripser algorithm [8] from raw spectra and from min–max normalized spectra. For each strain, Bottleneck and Wasserstein distance matrices were constructed. Intra-condition and inter-condition distances were compared, and the effect of UV treatment was quantified using Cohen's d.

4 Results and discussion

For the wild-type strain PS533, TDA revealed clear topological differences between non-treated and UV-treated spores. Moderate effect sizes were obtained with the Bottleneck distance, while stronger discrimination was observed with the Wasserstein distance. A more detailed analysis showed that UVC irradiation under dry conditions induced particularly strong topological changes, whereas UVB treatment in aqueous conditions led to much weaker effects.

In contrast, the mutant strain PS4150 exhibited Cohen's d values close to zero, indicating that inter-condition distances were not larger than intra-condition variability. For FB122, however, elevated Bottleneck effect sizes were observed under UVC treatments, suggesting the presence of localized topological changes, while global differences remained limited. This behavior is consistent with a more localized structural response to UV irradiation.

5 Conclusion

This study demonstrates the relevance of Topological Data Analysis for the chemometric analysis of IR and Raman spectra of bacterial spores. TDA enables robust discrimination of UV-induced structural changes in the wild-type strain PS533, while highlighting the limited sensitivity of mutant strains PS4150 and FB122 to UV irradiation. The complementary use of Bottleneck and Wasserstein distances provides insights into both local and global spectral changes, opening new perspectives for the analysis of complex biological systems using topological methods.

6 References

[1] E. J. Amézquita, M. Y. Quigley, T. Ophelders, E. Munch, and D. H. Chitwood. The shape of things to come: Topological data analysis and biology, from molecules to organisms. Developmental Dynamics, 249 (7): 816–833, July 2020.

[2] M. Offroy and L. Duponchel. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. Analytica Chimica Acta, 910: 1–11, 2016.

[3] P. Malbos. Raconte-moi... la persistance topologique. La Gazette des mathématiciens, 169: 8, July 2021.

[4] G. Carlsson. Topology and data. Bulletin of the American Mathematical Society, 46 (2): 255–308, 2009.

[5] F. Chazal, B. Michel. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. Frontiers in Artificial Intelligence, 4: 667963, September 2021.

[6] H. Edelsbrunner and J. Harer. Computational topology: an introduction. American Mathematical Society, Providence, R.I, 2010.

[7] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. Scientific Reports, 3: 1236, 2013.

[8] U. Bauer. Ripser: Efficient computation of Vietoris–Rips persistence barcodes. Journal of Applied and Computational Topology, 5 (3): 391–423, 2021.

Author Index

Chimiométrie 2026

17-19 Feb 2026
Nancy
France