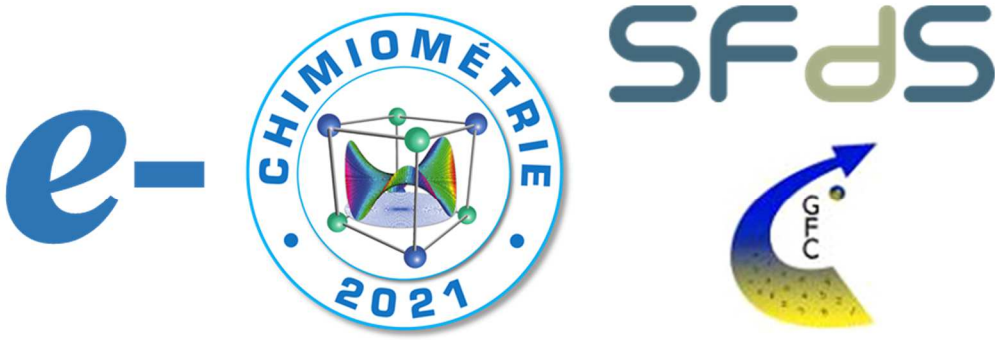


e-CHIMIOMETRIE 2021

2-3 février 2021

https://chemom2021.sciencesconf.org/

_____

Program / Book of Abstracts

Program

2 Février 2021

08:45 - 9:00	Accueil - Mot du Président du GFC - Organisation du congrès / *Welcome*			
	Conférence plénière / *Keynote* ***Chairmen : S. RUDAZ and G. SAPORTA***			
09:00 - 9:40	Achim KOHLER	Norwegian University of Life Sciences, Ås, Norway	Model-based pre-processing and deep learning for correcting scatter effects in highly scatter-distorted infrared spectra of cells and tissues	**p 5**
	Session "Jeunes Chimiométriciens" Part. 1 / *"Young Chemometricians" Part 1.* ***Chairmen : S. RUDAZ and G. SAPORTA***			
09:40 - 10:00	Mohamad AHMAD	Università di Modena e Reggio Emilia, Italy	Isolation of distinct spatial components in convoluted hyperspectral images	**p 6**
10:00 - 10:20	Oumauma BOUKRIA	Université Sidi Mohamed Ben Abdellah, Maroc	Monitoring of molecular structure modifications during coagulation of mixed camel and cow milk by MIR and PARAFAC	**p 8**
10:20 - 10:40	Maxime METZ	Irstea, ITAP Montpellier, France	RoBoost-PLSR : Robust PLS regression method inspired from boosting principles	**p 11**
10:40 - 11:00	Louna ALSOUKI	Université Claude Bernard Lyon, France	Interpretable Dual Sparse Partial Least Squares (DS-PLS) regression; Application to NMR/NIR petroleum data sets	**p 13**
11:00 - 11:20	**Pause / *Break***			
11:20 - 11:40	Marie CHION	Université de Strasbourg, France	Using monotone spline smoothing to combine label-free and label-based accurate quantifications with DIA-MS: application to bovine muscle samples	**p 15**
11:40 - 12:00	Puneet MISHRA	Wageningen University and Research, The Netherlands	Challenging deep learning with simple chemometrics for predicting leaf nitrogen using visible and near-infrared spectroscopy	**p 18**
12:00 - 12:20	Benjamin BRUNEL	Université de Reims, France	Toward automated machine learning in vibrational spectroscopy: contribution of genetic algorithm for optimal pre-processing and regression	**p 20**
12:20 - 14:00	**Déjeuner / *Lunch***			
	Session "Jeunes Chimiométriciens" Part. 2 / *"Young Chemometricians" Part 2.* ***Chairman : L. DUPONCHEL and M. EL RAKWE***			
14:00 - 14:20	Michel BAQUETA	University of Campinas, Brasil	Classification of Brazilian Coffea canephora cultivated by natives in the Amazon rainforest using portable near-infrared spectroscopy	**p 22**
14:20 - 14:40	Laureen COIC	Université de Liège CIRM, Belgium	Pixel-based identification of Raman hyperspectral data: Application to pharmaceutical tablets impurities detection	**p 24**
	Prix de thèse du GFC / GFC Award ***Chairman : L. DUPONCHEL***			
14:40 - 15:00	Daniel PALACI-LOPEZ	Universitat Politècnica de València, Spain	Product design based on Latent Variable Model Inversion: new tools for process exploration and optimization	**p 26**

Conférence plénière / *Keynote* *Chairmen : D. RUTLEDGE and J-M. ROGER*				
15:00 - 15:40	Oxana RODIONOVA	Semenov Federal Research Center, Russian Academy of Sciences	Beneficial Features of Procrustes Cross Validation	**p 28**
Session "Méthodes" Part. 1 / *Methods Part. 1* *Chairmen : D. RUTLEDGE and J-M. ROGER*				
15:40 - 16:00	Andrea CAPPOZZO	University of Milano-Bicocca, Italy	Robust variable selection in the framework of classification with label noise and outliers: applications to spectroscopic data in agri-food	**p 29**
16:00 - 16:20	**Pause / *Break***			
16:20 - 16:40	Ludovic DUPONCHEL	Université de Lille, France	Saturated signals in spectroscopic imaging: Why and How should we deal this regurlaly observed phenomenon?	**p 31**
16:40 - 17:00	Marina COCCHI	University of Modena and Reggio Emilia, Italy	Assessing different facets of SIMCA modelling: decision rule, parameter optimization and their interplay	**p 33**
17:00 - 17:20	Raffaele VITALE	Université de Lille, France	p-SIMCA: a non-parametric probabilistic version of the SIMCA classifier	**p 35**

3 Février 2021

Conférence plénière / *Keynote* *Chairmen : C. RUCKEBUSCH and B. JAILLAIS*				
09:00 - 9:40	José Manuel AMIGO	University of the Basque country, Spain	The Needle in the Haystack, or Microplastics in Natural Samples. What is More Complex to Find? Analytical Strategies using Raman and Mid-Infrared Imaging	**p 37**
Session "Méthodes" Part. 2 / *Methods Part. 2* *Chairmen : C. RUCKEBUSCH and B. JAILLAIS*				
09:40 - 10:00	Alessandro NARDECCHIA	Université de Lille, France	Spectral and spatial fusion: An interesting approach for classification in hyperspectral imaging	**p 38**
10:00 - 10:20	Elaheh TALEBANPOUR BAYAT	University of Shiraz, Iran	Automated chemical rank determination by hybridizing dependency concept and permutation testing	**p 40**
10:20 - 10:40	Olivier DEVOS	Universié de Lille, France	Multi-exponential analysis with MCR slicing	**p 42**
10:40 - 11:00	**Pause / *Break***			
Session "Applications" Part. 1 / Applications Part. 1 *Chairmen : E. ZIEMONS and P. LANTERI*				
11:00 - 11:20	Jhon BUENDIA	Irstea ITAP Montpellier, France	Diesel cetane number prediction by data fusion of near-infrared and nuclear magnetic resonance spectroscopy	**p 44**
11:20 - 11:40	Yves ROGGO	Novartis, Suisse	Industry 4.0 enablers for pharmaceutical manufacturing	**p 46**

11:40 - 12:00	Morandise RUBINI	Université de Pau et des Pays de l'Adour, France	Bois imprégnés avec des produits de préservation commerciaux : Utilisation de stratégies de fusion de données pour améliorer leur discrimination	**p 48**
Session "Data challenge" ***Chairmen : P. DARDENNE***				
12:00 - 12:40	Présentation du Data Challenge 2021 / Exposés des meilleures solutions			
12:40 - 14:00	**Déjeuner / *Lunch***			
Conférence plénière / *Keynote* ***Chairmen : C. CORDELLA and P. BASTIEN***				
14:00 - 14:40	Pierre LEBRUN	Pharmalex, Belgium	QbD tools to inscribe PAT control measurement into the process validation lifecycle	**p 50**
Session "Applications" Part. 2 / Applications Part. 2 ***Chairmen : C. CORDELLA and P. BASTIEN***				
14:40 - 15:00	Priyanka KUMARI Thomas VAN LAETHEM	Université de Liège, CIRM, Belgium	Quantitative structure-retention relationship modelling of small pharmaceutical compounds in reverse phase liquid chromatography	**p 51**
15:20 - 15:40	Sébatien PREYS	Ondalys, France	From complex real-world data to process understanding and monitoring, a use case in the chemical industry	**p 53**
15:40 - 16:00	T. Hermane AVOHOU	Université de Liège, CIRM, Belgium	Using prediction bands for near-infrared spectra for authentication and verification of drug products	**p 55**
16:20 - 16:40	Patricia VALDERRAMA	Universidade Tecnologica Federal do Parana, Brasil	Pseudo-univariate calibration by UV spectroscopy in the determination of resveratrol in grape juice	**p 57**
Remise des prix / clôture du congrès ***Chairmen : S. RUDAZ and L. DUPONCHEL***				

Book of Abstracts

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Model-based pre-processing and deep learning for correcting scatter effects in highly scatter-distorted infrared spectra of cells and tissues

A Kohler*, J. Solheim, E. A. Magnussen, U. Blazhko, V. Tafintseva, B. Zimmermann, M.A. Brandsrud, S. Dzurendova, V. Shapaval

Faculty of Science and Technology, Norwegian University of Life Sciences, PO Box 5003, 1432 Ås, Norway

*Email: achim.kohler@nmbu.no

Keywords: Extended multiplicative signal correction, Descatter Autoencoder

Abstract

Scatter effects can severely distort infrared spectra. Especially, in infrared microscopy of cells and tissues strong scatter effects have been encountered [1]. Extended multiplicative signal correction (EMSC) has been introduced in the 90ies as a model-based chemometric method for correcting scatter effects such as baseline variations and multiplicative effects.

During the recent years we have developed a framework for retrieving pure absorbance spectra from highly scatter-distorted infrared spectra of cells and tissues by incorporating Mie theory and other electromagnetic models in the EMSC framework. The Mie Extinction Extended Multiplicative Signal Correction (ME-EMSC) algorithm is the state-of-the-art pre-processing technique [2] which can recover pure absorbance spectra from highly scatter distorted spectra [3].

The ME-EMSC algorithm is computationally expensive, and the correction of large infrared images often requires hours of computations. In order to address this problem, we have trained a deep convolutional Descatter Autoencoder on ME-EMSC corrected spectra for correction of hyperspectral infrared images of cells and tissues [4]. In terms of speed, robustness and noise-levels, the Descattering Autoencoder outperformed the ME-EMSC algorithm which makes the Descattering Autoencoder particularly appropriate for correcting hyperspectral maps. The speed advantage of the Descattering Autoencoder could allow to pre-process hyperspectral images in near real-time, and thereby making it feasible to use such images in medical applications.

References

[1] Mohlenhoff, B., Romeo, M., Diem, M., & Wood, B. R. Mie-type scattering and non-Beer-Lambert absorption behavior of human cells in infrared microspectroscopy. *Biophysical journal*, 88(5), 3635-3640, 2005.

[2] Solheim, J. H., Gunko, E., Petersen, D., Großerüschkamp, F., Gerwert, K., & Kohler, A. An open-source code for Mie extinction extended multiplicative signal correction for infrared microscopy spectra of cells and tissues. *Journal of biophotonics*, 12(8), e201800415, 2019.

[3] Sirovica S., Solheim J.H., Skoda M.W., Hirschmugl C.J., Mattson E.C., Aboualizadeh E., Guo Y., Chen X., Kohler A., Romanyk D.L. Origin of micro-scale heterogeneity in polymerisation of photo-activated resin composites. *J Nature Communications* 11(1):1-10, 2020.

[4] Magnussen, E.A., Solheim J.H., Blazhko U., Tafintseva V., Tøndel K., Liland K.H., Dzurendova S., Shapaval V., Sandt C., Borondics F., & Kohler A. Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells. *Journal of Biophotonics* 2020:e202000204, 2020.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Isolation of distinct spatial components in convoluted hyperspectral images

M. Ahmad[1] R. Vitale[2] C. Ruckebusch[3] M. Cocchi[4]

[1]Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Modena, Italy, m.ahmad@live.nl.

[2]Université de Lille, LASIR - Laboratoire de Spectrochimie Infrarouge et Raman, Lille, France, rvitale86@gmail.com.

[3]Université de Lille, LASIR - Laboratoire de Spectrochimie Infrarouge et Raman, Lille, France, cyril.ruckebusch@univ-lille.fr.

[4]Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Modena, Italy, marina.cocchi@unimore.it

Keywords: Hyperspectral images, spatial features, wavelet transform, grey-level co-occurrence matrix, descriptors, multivariate image analysis.

1 Introduction

Hyperspectral imaging (HSI) is a very powerful tool for the analysis of complex systems such as food commodities, biological specimens, natural or synthetic materials, etc. In fact, it permits to simultaneously explore spatial and spectral (chemical) domains for a more comprehensive characterization of the samples under study. Unfortunately, interpreting HSI data may not be always straightforward because the information of interest is usually shaded by effects due to different factors, mostly linked to the physical properties of these samples. These effects are usually corrected by spectral preprocessing. Here we present a methodology [1] to isolate the distinct spatial features within HSI data, giving the user the possibility of inspecting and interpreting the distinct spatial contributions of the signal.

2 Theory

The method utilizes 2D-wavelet transform (2D-WT), grey-level co-occurrence matrices (GLCM), descriptors and principal component analysis (PCA). 2D-WT highlights distinct spatial features by decomposing an image, at a given spectral channel, into disjoint sub-images which capture the different frequencies content across the different spatial directions at varying resolution scales. Grey-level co-occurrence matrices applied to these sub-images allow for the retrieval of descriptors encoding the local spatial pattern in terms of e.g. homogeneity, contrast, variance etc. [2]. PCA is an extremely versatile method, as it can highlight extreme variability, reduce dimensionality, and model data, by decomposing the data into eigenvalues and eigenvectors.

3 Material and methods

The method is applied on a near-infrared hyperspectral image (NIR HSI) of a semen droplet on cotton fabric [3], where the cotton background shows significant interference which hampers the recognition of the semen droplet.

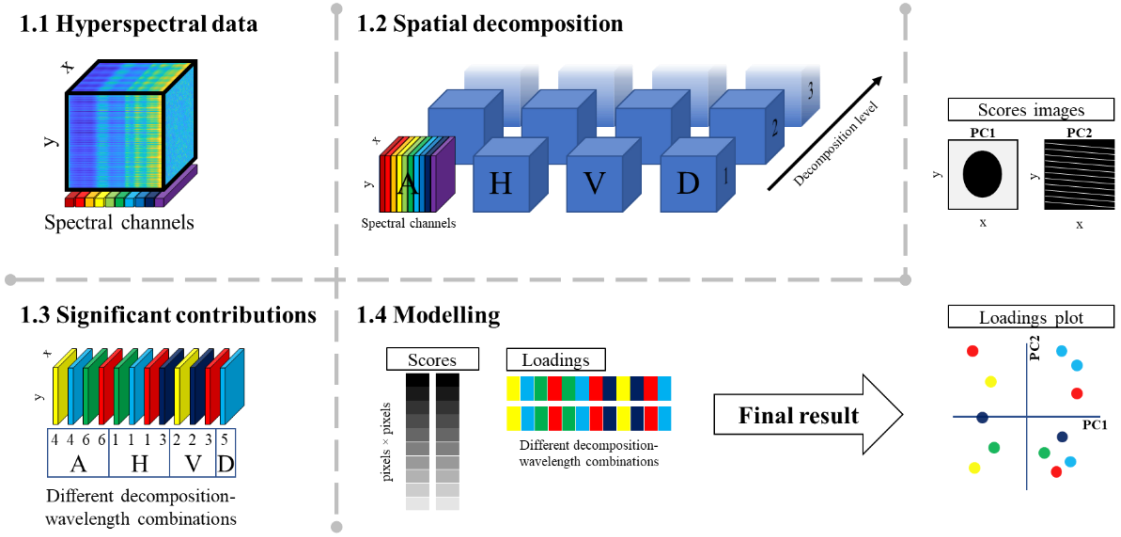


Figure 1 – Method flowchart, the coloring represents the spectral domain.

The flowchart is shown in Figure 1: 1.1) HSI data; 1.2) 2D-WT decomposition of each spectral channel; 1.3) the WT sub-images, capturing distinct and significant spatial contributions, are retrieved by PCA analysis of GLCM descriptors; 1.4) PCA is applied to these (unfolded) selected sub-images. The scores images and loadings plot depict the different components and their corresponding distinctive spectral channels.

4 Results and discussion

Figure 2 shows the most significant scores images: PC1 seems clearly linked to the semen droplet, PC2 to the cotton fabric (only observable pattern in the raw data), while PC3 is showing an intense spot which can be connected to semen considering the relatively low signal also seen in the same position (lower left corner) in PC1.

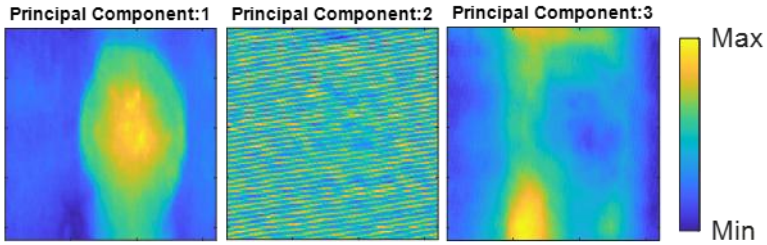


Figure 2 – First three (from left to right) principal component (PC) scores images are shown.

5 Conclusion

A clear separation of the semen and cotton background is observed in figure 2, while simultaneously highlighting a third component, which is connected to the semen. This connection is observed in the loadings (not shown).

6 References

[1] M. Ahmad, R. Vitale, C. Silva, C. Ruckebusch, M. Cocchi. Exploring local spatial features in Hyperspectral images. Journal of Chemometrics 2020, Volume 34, Issue 10.

[2] RM. Haralick, L. Shanmugan, I. Dinstein. Textural features for image classification. IEEE Trans Syst Man Cybern 1973, 3(6):610-621.

[3] C. S. Silva, M.F. Pimentel, J.M. Amigo, R.S. Honorato, C. Pasquini. Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models. Trends in Analytical Chemistry 2017, 95 p: 23-35.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Monitoring of molecular structure modifications during coagulation of mixed camel and cow milk by MIR and PARAFAC

O. Boukria[1] A. Aït-Kaddour[2] E. M. El Hadrami[1]

[1] Applied Organic Chemistry Laboratory, Sciences and Techniques Faculty, Sidi Mohamed Ben Abedallah University, BP 2202 route d'Immouzer, Fès, Morocco. oumayma.boukria@usmba.ac.ma elmestafa.elhadrami@usmba.ac.ma

[2] Université Clermont-Auvergne, INRAE, VetAgro Sup, 63370 Lempdes, France. abderrahmane.aitkaddour@vetagro-sup.fr

Keywords: Coagulation, Milk, MIR Spectroscopy, Particle size, PARAFAC analysis.

1 Introduction

Mixing milks from different species can be a strategy to increase the consumption of non-bovine milks and enable consumers and dairy companies to benefit from their nutritional and technological advantages [1]. It is particularly admitted that camel milk (CaM) presents a high nutritional quality; e.g. it has 3 times more vitamin C, more minerals, more essential and polyunsaturated fatty acids than CM [2], [3]. Regarding this opportunity, it is important to characterize quality features of products from mixing milks of different animal species in order to manage products with proper characteristics and satisfactory acceptance by consumers.

2 Material and methods

The CaM and CM mixture were prepared after warming each milk samples to 40 °C by using a water bath, gently mixing each milk by hand at least during 1 min. The volume fractions (%) of CaM in the different formulations were 100%, 75%, 50%, 25%, and 0%. Milk coagulations were performed at 40 °C (±1°C) by using 2.5 µL/mL of CHY-MAX® M (Chr. Hansen). MIR spectra were recorded between 3800 and 900 cm−1. All the spectra were recorded on the same sample formulation each 5 min during coagulation 115 min at a temperature of 40 °C. Parallel Factor Analysis (PARAFAC) was used in combination with particle size measurements, to evaluate the structure evolution at a molecular level during milk rennet coagulation of CaM (Camel milk) and CM (Cow milk) mixtures.

3 Results and discussion

Regarding MIR spectroscopy (figure 1), the regions located between 1700 and 1500 cm−1 (amide I and II bands), 1500–900 cm−1 (fingerprint region), and 3000–2800 cm–1 (fatty acids) were considered for the characterization of milk coagulation kinetics. When considering the fingerprint

region, it was possible to identify the gelation point of the milk mixtures. The PARAFAC analysis was performed separately on those three MIR regions.

PARAFAC results applied to amide I/amide II range are presented in Figure. 2A-C. The similarity map revealed a discrimination of the scores on the first component (Figure. 2a) according to the coagulation time. Positive scores were observed from 0 to 75 min, while negative ones were observed between 80 and 115 min. Considering the component 2, the scores exhibited positive values from 0 to 65 min, and at 115 min, while negative ones were observed from 70 to 110 min. Moreover, a minimum value was noted at 80 min followed by a subsequent increase until 115 min.

Regarding the component one, the spectral pattern (Figure. 2B) exhibited positive peaks at 1628, 1566, 1549, 1530 and 1516 cm-1 and a negative band between 1700 and 1643 cm-1 presenting two minima at 1668 and 1556 cm-1. This highlighted that the most important information is provided by the first component.

The information related to the samples mode (Figure. 2C) showed that samples formulated only with pure milks (camel or cow) presented positive values for component 1, while samples containing a mixture of CaM and CM milks (i.e. 1CaM:1CM, 1CaM:3CM and 3CaM:1CM) showed negative values.

Considering the analysis of the fingerprint region, PARAFAC results are presented in Figure. 2D-F. Figure. 2D showed components 1 and 2 similarity map presenting an inversed bell-shaped curve. The bottom of the inversed bell-shaped curve corresponded to 80 min, and could be assigned to the gelation point as previously reported by Boubellouta et al.[4]. This time was also identified when considering the second component obtained after PARAFAC analysis of the protein region (Figure. 2B).

Regarding the components 1 and 2, the spectral patterns (Figure. 2E) exhibited different positive peaks at 1460, 1419, 1233, 1159, 1115, 1095, 963 cm-1 and negative ones at 1396, 1019, 988 cm-1 for the component 1. The information related to the samples mode (Figure. 2F) showed that samples formulated only with pure milks (camel and cow) presented negative values for the component 1, while samples containing a mixture of CaM and CM (i.e. 1CaM:1CM, 1CaM:3CM and 3CaM:1CM) presented positive values. Concerning the lipids band region, Figure. 2G showed components 1 and 2 similarity map. The scores on the map presented a bell-shaped curve as previously reported for the fingerprint region. Regarding the components 1 and 2, the spectral patterns (Figure. 2H) exhibited two prominent positive peaks at 2930 and 2850 cm-1 and a shoulder at 2960 nm. The information related to the samples mode (Figure. 2i) showed that the first component discriminated the pure milks from the other formulations. Concerning the second component, it can be observed that the formulations were almost separated depending on the concentration of the CM content.

4 Conclusion

MIR spectroscopy was performed to investigate modifications affecting molecular structure during milk mixture coagulation. The results provided different information during coagulation related to milk components which complementarity makes it possible to obtain information associated to changes of their molecular structure and interactions in the different milk formulations studied. MIR spectroscopy permitted to identify a variation of the coagulation time depending on the initial milk formulation. It is concluded that spectroscopic methods such as MIR spectroscopy combined with chemometric tools such as PARAFAC have the potential to characterize structural changes at the molecular level in milk coagulation

5 References

[1] O. Boukria, E. M. El Hadrami, S. Boudalia, J. Safarov, F. Leriche, and A. Aït-Kaddour, "The effect of mixing milk of different species on chemical, physicochemical, and sensory features of cheeses: a review," Foods, vol. 9, no. 9, pp. 1–23, 2020, doi: 10.3390/foods9091309.

[2] Z. Farah, R. Rettenmaier, and D. Atkins, "Vitamin content of camel milk," Int. J. Vitam. Nutr. Res., vol. 62, pp. 30–33, 1992.

[3] W. N. Sawaya, J. K. Khalil, A. Al- Shalhat, and H. Al- Mohammad, "Chemical composition and nutritional quality of camel milk," J. Food Sci., vol. 49, no. 3, pp. 744–747, 1984.

[4] T. Boubellouta, V. Galtier, and É. Dufour, "Structural changes of milk components during acid-induced coagulation kinetics as studied by synchronous fluorescence and mid-infrared spectroscopy," Appl. Spectrosc., vol. 65, no. 3, pp. 284–292, 2011.

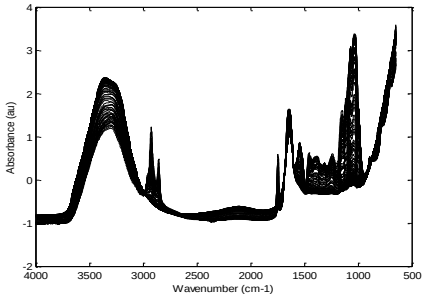


Figure 1 – MIR spectra recorded during enzymatic coagulation of the different mixtures of camel milk and cow milk.

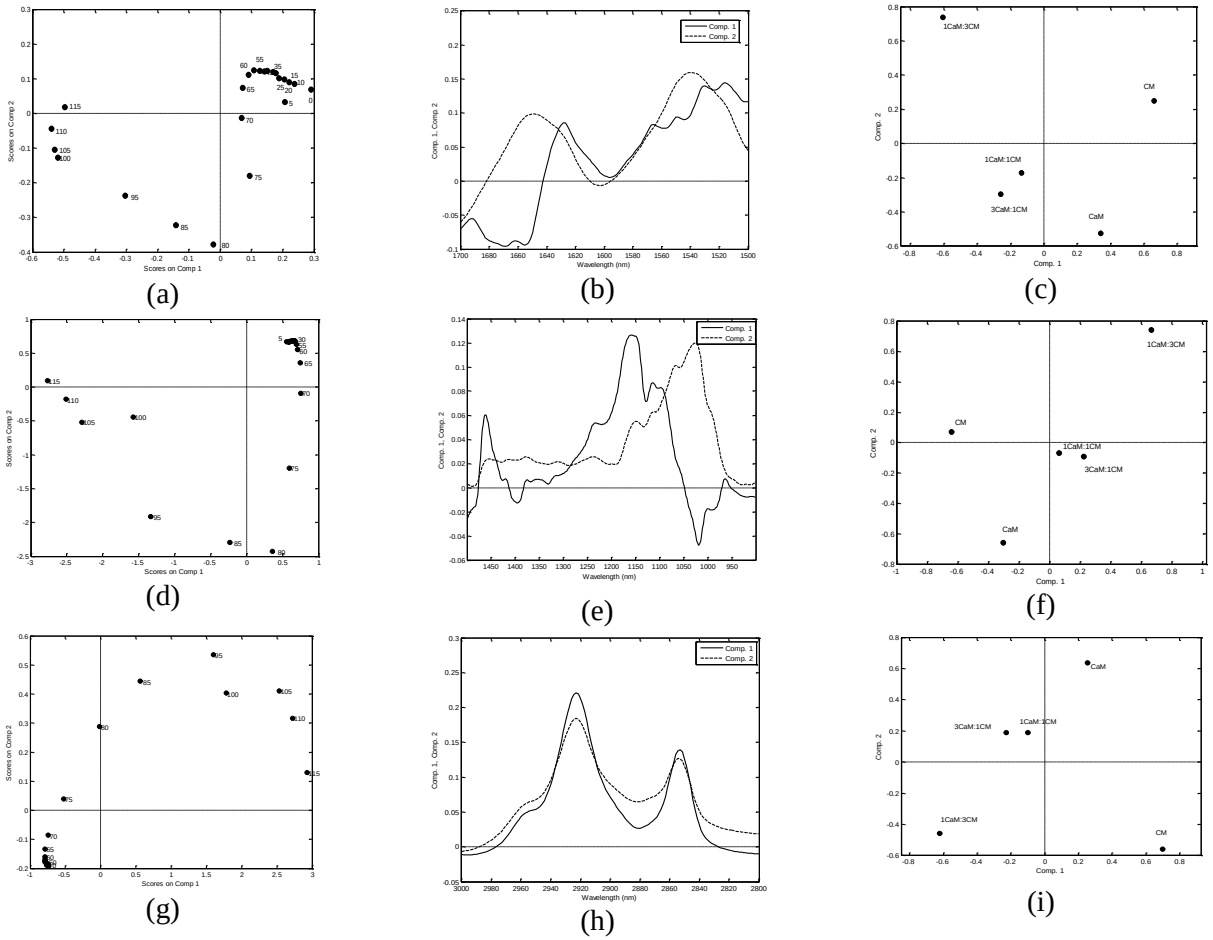


Figure 2 –Two-component PARAFAC model derived from the MIR data (a-c: amide I/amide II; d-f: fingerprint and g-i: fat range) recorded during milk coagulation of the different milk formulations containing a mixture of camel milk (CaM) and cow milk (CM). (a, d, and g) Kinetic mode, fingerprint mode (b, e, and h) and formulation mode (c, f, and i).

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

RoBoost-PLSR : robust PLS regression method inspired from boosting principles

M. Metz[1,2], F. Abdelghafour[1,2], JM. Roger[1,2], M. Lesnoff[2,3]

1 ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France
2 ChemHouse Research Group, Montpellier, France
3 CIRAD, UMR SELMET, Montpellier, France
4 SELMET, Univ Montpellier, CIRAD,INRA, Institut Agro, Montpellier,France

Keywords: Robust-PLSR, outliers, calibration

1 Introduction

The calibration of Partial Least Square regression (PLSR) models can be disturbed by outlying samples in the data. In these cases, the models can be unstable and their predictive potential can be depreciated. To address this issue, a new method and algorithm to better apprehend the downweighting of outliers in a context of high dimensional data processing is proposed. This novel robust PLSR algorithm is inspired from the principles of boosting and is called RoBoost-PLSR.

2 Theory

RoBoost-PLSR consists in achieving a series of *K* unidimensional (1 LV) iteratively reweighted PLSR[1] models. The weigthed PLSR algorithm used is the weighted NIPALS[2].

The model (*k*+1) is calibrated with the X and Y residuals of the previous *k* models. Within each model, weights are computed according to a combination of X-residuals, Y-residuals and leverages. The more the samples deviate from the model, the lower the weights. Iteratively, the model is updated according to the weights previously attributed until convergence to a stable solution.

3 Material and methods

RoBoost-PLSR was compared with the PLSR algorithm calibrated with and without outliers (i.e. the reference) and with Partial Robust M-regression (PRM), a reference robust method. This evaluation was conducted on the basis of a simulated dataset and a real dataset.

The simulated dataset was generated with the framework proposed in [4]. The simulation objective is to reproduce a contamination in the samples leading to inconsistent spectral measurements.

The real dataset is an example of one animal nutrition application: the prediction of the protein content of feed materials and of the presence of incorrectly categorised samples. In this database the samples resulting from animal bonemeal (noted ANF) represent the outliers polluting the regular soyabean cakes (noted TTS).

4 Results and discussion

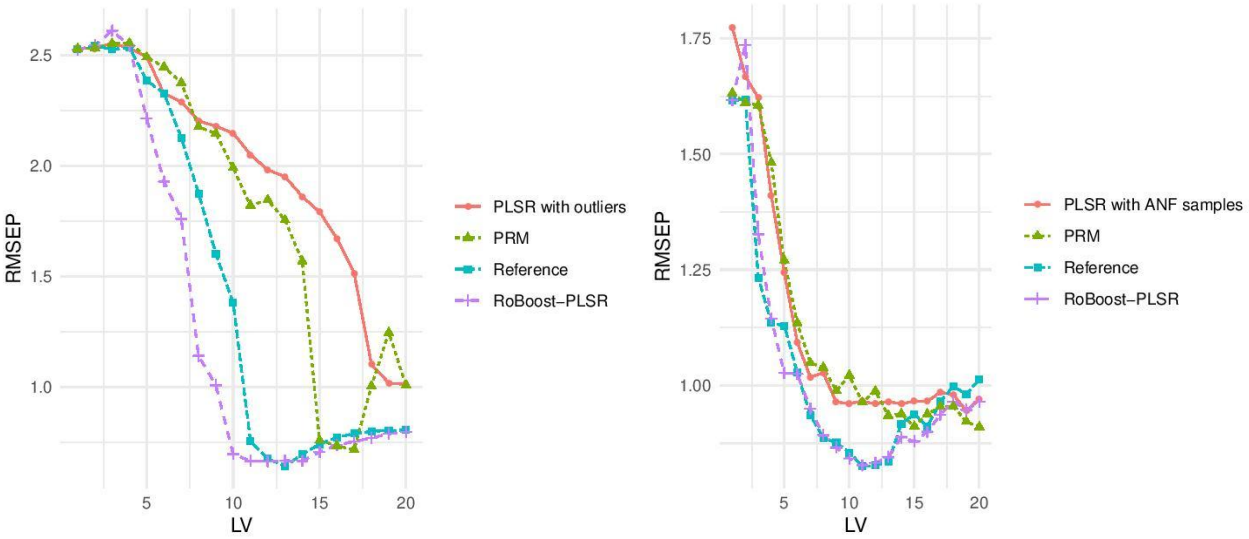


Figure 1: Evolution of the RMSEP as a function of latent variables, for the reference, PLSR with outliers, PRM and RoBoost-PLSR for the dataset simulated dataset (left) and the real dataset (right).

Figure 1: shows that for this type of outliers the RoBoost-PLSR method performs well and can reach the prediction performance of the PLSR method without outliers.

5 Conclusion

Roboost-PLSR proves to be resilient to the tested outliers, and can achieve the performances of the reference PLSR calibrated without any of these outliers.

6 References

[1] Cummins, David J., et C. Webster Andrews. « Iteratively Reweighted Partial Least Squares: A Performance Analysis by Monte Carlo Simulation ». Journal of Chemometrics 9, nº 6 (1995): 489‑507. https://doi.org/10.1002/cem.1180090607. [2] D. Tirambic : *The book*. The editor, the edition, 1929.

[2] Schaal, Stefan, Christopher G. Atkeson, et Sethu Vijayakumar. « Scalable Techniques from Nonparametric Statistics for Real Time Robot Learning ». Applied Intelligence 17, nº 1 (1 juillet 2002): 49‑60. https://doi.org/10.1023/A:1015727715131.

[3] Serneels, Sven, Christophe Croux, Peter Filzmoser, et Pierre J. Van Espen. « Partial Robust M-Regression ». Chemometrics and Intelligent Laboratory Systems 79, nº 1 (28 octobre 2005): 55‑64. https://doi.org/10.1016/j.chemolab.2005.04.007.

[4] Metz, Maxime, Alessandra Biancolillo, Matthieu Lesnoff, et Jean-Michel Roger. « A Note on Spectral Data Simulation ». Chemometrics and Intelligent Laboratory Systems 200 (15 mai 2020): 103979. https://doi.org/10.1016/j.chemolab.2020.103979.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFKL

Interpretable Dual Sparse Partial Least Squares (DS-PLS) regression; Application to NMR/NIR petroleum data sets

Louna Alsouki[1,2,3] François Wahl[1,2] Laurent Duval[2] Clément Marteau[1] Rami El-Haddad[3]

[1]Université Claude-Bernard Lyon 1, 43 boulevard du 11 Novembre 1918, 69100 Villeurbanne, France
[2]IFP Energies nouvelles, 1-4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France
[3]Université Saint-Joseph de Beyrouth, Mar Roukoz – Dekwaneh, B.P. 1514, Liban
louna.al-souki@univ-lyon1.fr, francois.wahl@univ-lyon1.fr, marteau@univ-lyon1.fr, laurent.duval@ifpen.fr, rami.haddad@usj.edu.lb

Keywords: Partial Least Squares, sparsity, regression, dual norm.

1 Introduction

Regression analysis helps in inferring relationships between data sets, with the additional objective of extracting interpretable information. Partial Least Squares [1] (PLS) is often used when dealing with NMR or NIR spectra to predict properties of petroleum samples. In spite of its ability to operate with high-dimensional data, and its efficiency in predicting responses, PLS lacks in considering the functional nature of the data and shows weaknesses in result interpretation. In order to improve these two aspects, we developed a new strategy called Dual Sparse Partial Least Squares (DS-PLS) that gives equivalent prediction accuracy along with facilitated interpretation of regression coefficients, due to the sparsity of the representation.

2 Theory

The proposed method was devised from noticing the similarity between finding the PLS components (with the PLS1 methodology) and expressing the dual L_2 norm of a vector.

Let be a norm. Its dual [2] has the following form:

Meanwhile, the optimization problem solved by the PLS method for the first component writes:

Comparing (1) and (2), one notices that optimizing the PLS function amounts to finding the vector that goes with the conjugate of the L_2-norm of z, where
Therefore, we propose to evaluate different norm expressions, notably adding adaptive penalization. An example is the norm $\Omega(w) = \lambda\|w\|_1 + \|w\|_2$. Interestingly, this formulation leads to closed-form expressions as in [3] and requires only slight modifications of the standard PLS1 algorithm: the solution is known as the soft thresholding operator in the lasso [4] literature: ∀

— where μ is tuned to guarantee that 1.

Moreover, this framework allows to vary the form of the norm. Another possibility is for example ‖ like in fused lasso, where N is a penalty matrix. These constraints would introduce a functional aspect in the treatment.

3 Material and methods

We apply the DS-PLS to 208 samples of NIR spectra represented by 2594 variables and to 243 samples of NMR spectra represented by 20998 variables. After dividing the data sets in two similar sets (calibration and validation) and using the R programming platform, we evaluate the prediction using both Root Mean Squares and Mean Absolute Errors. We also compare the coefficients for each model with the raw data.

4 Results and discussion

Comparing methods, the proposed strategy matches the prediction accuracy of the PLS, and additionally provides a good interpretation of the coefficients (Figure 1) due to the sparsity of its results. In the following figure, we require 99 % of null coefficients while applying DS-PLS and use 6 components for both standard and DS-PLS regressions. Note that x-axis units are not represented due to data preprocessing.

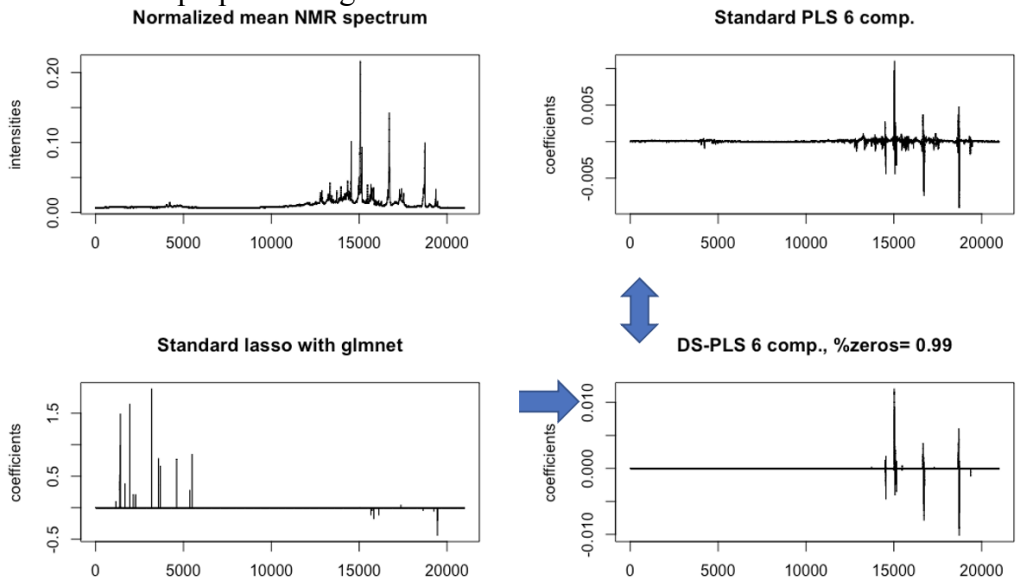


Figure 1 – Comparing coefficients of regression with the normalized mean MNR spectrum.

5 Conclusion

DS-PLS is a novel family of regression methods that provides a general framework: it encompasses the standard PLS method, and gives us the possibility to use other norm shapes. At this point, it preserves the accuracy of prediction of the PLS method and adds on sparsity in the coefficients for interpretation. The next challenge is to evaluate norms like ones for in fused or grouped lasso.

6 References

[1] Tenenhaus M. La régression PLS: théorie et pratique. Paris: Éditions Technip, 1998.

[2] Bach F., Jenatton R., Mairal J., and Obozinski G. Optimization with Sparsity-Inducing Penalties. *Found. Trends Mach. Learn.*, 2012, 4(1), 1–106.

[3] Durif G., Modolo L., Michaelsson J., Mold J., Lambert-Lacroix S., Picard F. High Dimensional Classification with combined Adaptive Sparse PLS and Logistic Regression. *Bioinformatics, Oxford University Press,* 2018, 34 (3), pp.485-493.

[4] Tibshirani R. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological). Wiley*, 1996, 58 (1): 267–88.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Using monotone spline smoothing to combine label-free and label-based accurate quantifications with DIA-MS: application to bovine muscle samples

M. Chion[1] J. Bons[2] M. Bonnet[3] M. Maumy-Bertrand[4] C. Carapito[5] F. Bertrand[6]

[1] Institut de Recherche Mathématique Appliquée, CNRS - Université de Strasbourg, UMR 7501, et Laboratoire de Spectrométrie de Masse Bio-Organique, Institut Pluridisciplinaire Hubert Curien, UMR 7178 CNRS – Université de Strasbourg, Strasbourg, France. chion@math.unistra.fr

[2] The Buck Institute for Research on Aging, Novato, California, USA. JBons@buckinstitute.org

[3] Université Clermont Auvergne, INRA, VetAgro Sup, UMR Herbivores, Saint-Genès-Champanelle, France. muriel.bonnet@inrae.fr

[4] Laboratoire de Modélisation et Sûreté des Systèmes, Institut Charles Delaunay, Université de Technologie de Troyes, Troyes, France. myriam.maumy@utt.fr

[5] Laboratoire de Spectrométrie de Masse Bio-Organique, Institut Pluridisciplinaire Hubert Curien, UMR 7178, CNRS – Université de Strasbourg, Strasbourg, France. ccarapito@unistra.fr

[6] Laboratoire de Modélisation et Sûreté des Systèmes, Institut Charles Delaunay, Université de Technologie de Troyes, Troyes, France. frederic.bertrand@utt.fr

Keywords: Beef meat quality, DIA-MS, Monotone spline smoothing, Quantitative mass spectrometry, Targeted proteomics.

1 Introduction

Proteomic analysis consists in studying proteins from a given biological system, at a given time and under given conditions. Global quantification methods make it possible to compare thousands of proteins expression levels across the different biological samples that are considered. Targeted quantification methods allow, by introducing labelled synthetic standards corresponding to previously selected peptides of interest, to know precisely the quantity of specific in the biological sample considered. A recent approach, called Data-Independent Acquisition [1], enables to combine these two methods in a single analysis. In quantitative proteomics, the strong hypothesis is made of a relationship of proportionality between the quantity of a peptide and its intensity through the response factor, specific to each peptide in each sample. From the intensity and quantity data obtained in targeted quantification, we propose to fit monotone spline models explaining the quantity of a peptide by its intensity in the considered sample. These models then allow us to estimate the amounts of all detected peptides thanks to the use of internal labelled standards for a subset of peptides.

2 Material and methods

2.1 DIA SWATH-MS analysis

The work described here relies on the experiment from Bonnet et al. (2020) [2]. 64 samples of bovine muscles for which 20 peptides corresponding to the 10 potential biomarker proteins for beef tenderness and marbling have been analyzed using a DIA SWATH-MS approach [3]. A first step of targeted quantification enabled, based on the intensity measured by liquid chromatography coupled

with mass spectrometry, to precisely determine the quantity of the 20 peptides of interest within each of the 64 samples considered. In order to do this, the following relationship was used:

$$\frac{\text{peptide quantity}}{\text{synthetic peptide quantity}} = \frac{\text{peptide intensity}}{\text{synthetic peptide intensity}} \quad (1)$$

A second step of data extraction from the same analyses enabled to measure the intensity of approximately 5500 peptides per sample.

2.2 Monotone spline analysis

Monotone spline smoothing combines I-spline regression analysis [4] and non-negative least squares estimation of parameters to ensure monotonicity. The parameters of the monotone spline models were estimated using the Lawson-Hanson algorithm for non-negative least square estimation. I-spline analysis was conducted using *iSpline* function from the "splines2" package [5] and non-negative least squares models were fitted using the *nnls* function from the "nnls" package [6] in R 4.0.2 software.

3 Results and discussion

An I-spline regression model was fitted for each of the 64 bovine samples considered, using the data obtained from the label-based quantification step. An example of these fits is given in Figure 1a. The log-transformed peptide intensity was chosen as the predictor and the log-transformed peptide quantity as the dependent variable.

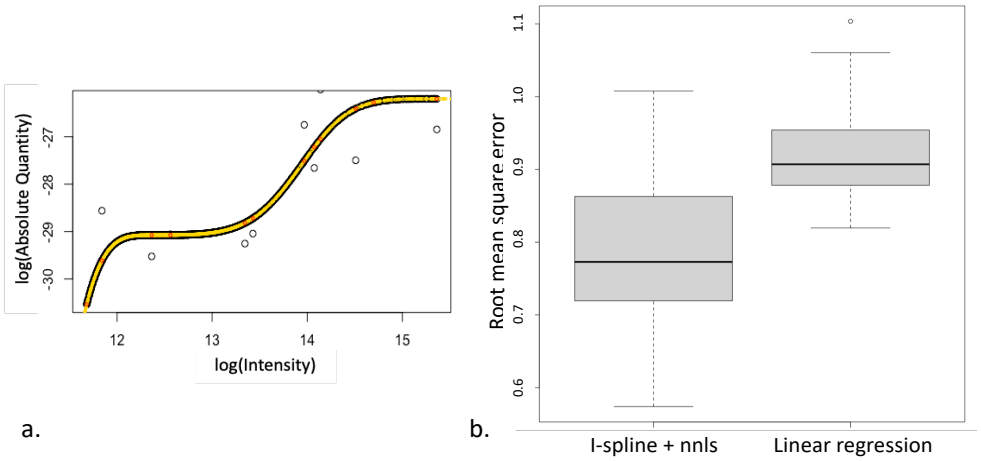


Figure 1 – Results of the monotone spline regression fit. 1a. Example of a monotone spline regression fit. 1b. Comparison of root mean square errors between monotone spline regression and simple linear regression.

As illustrated in the example on figure 1a, the linear relation hypothesis between the intensity and the quantity cannot be retained. Over the 64 samples considered, monotone spline regression offers a satisfying fit. Moreover, it outperforms the usual linear regression in terms of root mean square error, as represented in figure 1b.

4 Conclusion

From the label-based quantification data we used monotone spline smoothing to explain absolute amounts of targeted proteins by their intensities. Our approach led to a better fit than the simple linear regression. Then from the intensities of proteins quantified in the label-free quantification step, we estimated their absolute quantity. A further biological analysis of the predicted absolute protein quantities showed that our results were consistent with the literature on bovine muscles.

5 References

[1] Gillet, L.C. et al. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics*, 11(6), 2012.

[2] Bonnet, M., Soulat, J., Bons, J., Léger, S., De Koning, L., Carapito, C. & Piccard, B. Quantification of biomarkers for beef meat qualities using a combination of Parallel Reaction Monitoring- and antibody-based proteomics. *Food Chemistry*, 317, 2020.

[3] Ludwig, C., L. Gillet, L., Rosenberger, G., Amon, S., Collins, B.C. & Aebersold, R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 14(8), 2018.

[4] J.O. Ramsay. Monotone Regression Splines in Action, *Statistical Science*, 3(4), 425-461, 1988.

[5] Wang, W. & Yan, J.: splines2: Regression Spline Functions and Classes. R package version 0.3.1, <URL: https://CRAN.R-project.org/package=splines2>. 2020.

[6] Mullen, K. M. & Van Stokkum, I. H. M. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). R package version 1.4. <URL: https://CRAN.R-project.org/package=nnls>. 2012.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Challenging deep learning with simple chemometrics for predicting leaf nitrogen using visible and near-infrared spectroscopy

Puneet Mishra[1], Douglas N. Rutledge[2], Jean-Michel Roger[3]

[1]Wageningen Food and Biobased Research, The Netherlands (Puneet.mishra@wur.nl)
[2]Université Paris-Saclay, Paris, France (rutledge@agroparistech.fr)
[3]ITAP, University Montpellier, France (jean-michel.roger@inrae.fr)

Keywords: deep learning; sequential learning; complementary

1 Introduction

Deep learning is emerging as a key tool in spectral data modelling [1, 2]. It is often hyped that deep learning can outperform standard chemometric approaches in terms of prediction accuracies [2]. That may be true in some cases but deep learning lacks the explanatory aspect of models obtained with classical chemometric techniques such as principal component analysis and partial least-squares regression. Hence, in this work our aim is to demonstrate how classical chemometric methods can attain very similar performance as deep learning, while giving the possibility to explore model characteristics such as the regression coefficient etc. In order to do this, we analyze exactly the same data that has been used to develop deep learning models to predict nitrogen in oilseed rape [2]. We analyze the data with a new chemometric approach called sequential pre-processing through orthogonalization (SPORT) [3] which allows to extract complementary information from the data following different pre-processings.

2 Material and methods

Data set

The data set used in this study was first presented in [2] and consisted of 192 mean Vis-NIR (380-1030 nm) hyperspectral signals of rapeseeds (*Brassica napus, Zheyou51*). The extracted mean spectra and the reference N values used for the data modelling were supplied as the supplementary material to the article [2]. Out of the 192 samples, 128 were used for model calibration and tuning, and 64 were used as the independent test set. The calibration and the test sets were exactly the same as those used in the original article [2], hence, the results from this study will be directly comparable to those obtained previously with several machine learning and deep learning techniques [2].

SPORT modelling

The SPORT approach [3] allows to extract complementary information from a differently pre-processed data set. For the simplest case of only two differently preprocessed-data blocks ($\mathbf{X}_1$ and $\mathbf{X}_2$), the SPORT algorithm functions as follows :

1. The **Y** responses are fitted to the $\mathbf{X}_1$ by PLS regression;

2. $\mathbf{X}_2$ is orthogonalized with respect to the scores obtained from the first PLS regression;
3. The orthogonalized $\mathbf{X}_2$ is used to predict the $\mathbf{Y}$ residuals;
4. The overall predictive model is obtained by combining the sub-models (concatenating the scores) calculated in steps 1 and 3

In the present study, 4 pre-treatments were in fact used. The learning order was raw reflectance, reflectance corrected by SNV [REFERENCE], reflectance corrected by VSN [REFERENCE] and 2nd derivative [REFERENCE] of reflectance. The number of LVs was optimized by examining all possible combinations of LVs and the optimal model was the one giving the lowest RMSECV.

3 Results and discussion

The optimal SPORT model gave a prediction $R^2_P = 0.91$ and an RMSEP = 0.31 %. Not only are these results better than those obtained in the original study [2] on exactly the same calibration and test sets by PLS regression ($R^2_P = 0.85$ and RMSEP = 0.38 %), and by LS-SVM regression (R^2_P = 0.88 and RMSEP = 0.35 %), they are equivalent to those obtained in that article by the deep learning approach ($R^2_P = 0.90$ and RMSEP = 0.31 %) [2].

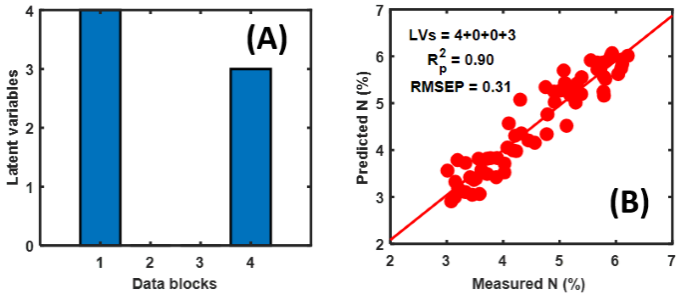


Figure 1: *Summary of the sequential pre-processing through orthogonalization (SPORT) model. (A) Number of latent vectors extracted from raw reflectance (1st data block), SNV (2nd data block), VSN (3rd data block) and 2nd derivative (4th data block). In total, SPORT modelled 4 latent variables from raw reflectance and 3 from the 2nd derivative, and (B) SPORT predictions.*

4 Conclusion

The SPORT attained better results than those of the deep learning approach used in [2]. This is of particular interest since the SPORT method is simpler, using linear algebraic operations, thus saving the time and resources required for training deep learning models. The other main benefit of SPORT in relation to the deep learning model (based on stacked autoencoders and fully connected neural network) is that it furnishes all the usual multivariate data analysis results, such as loadings, scores and regression vectors, which facilitate the spectrochemical interpretation of the models.

5 References

[1] C. Cui, T. Fearn, Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration, Chemometrics and Intelligent Laboratory Systems, 182 (2018) 9-20.

[2] X. Yu, H. Lu, Q. Liu, Deep-learning-based regression model and hyperspectral imaging for rapid detection of nitrogen concentration in oilseed rape (Brassica napus L.) leaf, Chemometrics and Intelligent Laboratory Systems, 172 (2018) 188-193.

[3] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometrics and Intelligent Laboratory Systems, 199 (2020) 103975.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Toward automated machine learning in vibrational spectroscopy: contribution of genetic algorithm for optimal pre-processing and regression

B. Brunel[1] F. Alsamad[1] O. Piot[1]

[1] Université de Reims Champagne-Ardenne, BioSpecT EA-7506, France, benjamin.burnel@univ-reims.fr

Keywords: genetic algorithm, regression, pre-processing, Raman, NIR.

1 Introduction

Vibrational spectroscopy (VS) has become a valuable tool in many fields as it provides a molecular signature with a single non-contact measurement. To extract valuable molecular information from the spectra, the data must be pre-processed to remove unwanted sources of variability. Pre-processing usually comport smoothing, baseline correction, and normalization, with many methods available for each step. Defining the right methods can significantly increase the performances of the subsequent regression or classification, while not appropriate ones can be worse than no pre-processing at all. The best pre-processing depends on the dataset, thus needs to be tuned for each application. Instead of testing all possible combinations, genetic algorithms (GA) have been suggested as a faster way to find an appropriate pre-processing sequence, in NIR [1], MIR [2] and Raman [3] spectroscopies. Genetic algorithms are optimizing techniques inspired by the natural selection and evolution. They are particularly efficient in solving problem with a high dimensionality, and are versatile enough to be adapted to specific problems. In VS, they were mostly used for selecting the most relevant wavenumbers of spectra for classification [4], or regression [5] tasks.

Similarly to the pre-processing choice, the predictive model have to be chosen carefully. This can also be done by a GA, evaluating both the machine learning models and their parameters. Combining pre-processing and model selection, a GA can optimize simultaneously the full sequence of data processing. This end-to-end automation of machine learning implementation is called automated machine learning (AutoML) and have been an active field lately. While machine learning is getting increasingly complex, AutoML gives the opportunity to non-expert to benefit from machine learning algorithms application. For near infrared and mid-infrared data, Devos et al.[6] showed how a GA could simultaneously choose the best pre-processing methods and the best parameters for a SVM classification. In this paper, we show how to best use GAs as an AutoML for regression on vibrational spectra data. GAs themselves comport many parameters, directly affecting the speed at which they can find an optimal solution. However, the question of tuning them in the context of VS was only marginally addressed by Bangalore et al.[5]. In this study, our objective was to determine the best GA parameters and provide a proficient GA, readily usable on vibrational data, without further tuning.

2 Results and discussion

GAs were evaluated on how fast they could find a solution with a root-mean-square error (RMSE) smaller than a given percentile of the solutions RMSE distribution. Four rarities of solutions were tested: in the top 5%, 1%, 0.1% or 0.01% of all possible solutions. The score used to quantify GAs performances was the time to have a 95% probability to reach a solution of the required rarity, measured in terms of number of solutions evaluated.

We determined the optimal parameters of the GA for three different datasets (Raman or NIR spectra from food industry or biological samples) and found that they were almost similar. Therefore, we were able to determine common parameters performing well on the three datasets. Fig. 1 shows that the GA with the common parameters ('common) scores close to the GA with parameters individually optimized ('individual'). GAs scores were compared to random search and Tree-structured Parzen Estimator (TPE), a Bayesian optimization method, commonly used in machine learning for tuning hyperparameters. For a rarity of 5 % GAs, random and TPE are comparably fast. However, for better qualities, GAs outperform both random and TPE.

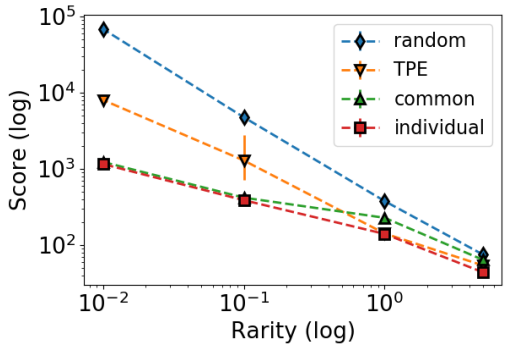


Figure 1 – scores as a function of the rarity of the required solution (top n% of the possible solutions) for random search, TPE and GA optimized for a given dataset ('individual') or for the three datasets together ('common').

3 Conclusion

In this paper, we propose a GA optimizing the combination of pre-processing, dimension reduction (using PCA or PLS), and regression model selection for VS datasets. As an end-to-end optimizer, our GA contributes to the recent field of automated machine learning. The best parameters of the GA hardly depend on the dataset nor the spectroscopic modality (MIR, NIR, Raman). For solution in the top 5 % of the search space, random search is enough, but the rarer the solution, the more GAs outperform random search and TPE. Consequently, for regression tasks on VS datasets, we encourage the use of the GA with the appropriate parameters determined here.

4 References

[1] Devos, O. & Duponchel, L. Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression. *Chemometrics and Intelligent Laboratory Systems*. 107, 50–58, 2011.

[2] Jarvis, R. M. & Goodacre, R. Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics*. 21, 860–868, 2005.

[3] Bocklitz, T., Walter, A., Hartmann, K., Rösch, P. & Popp J. How to pre-process Raman spectra for reliable and stable models? *Analytica chimica acta*. 704, 47–56, 2011.

[4] Li, S., Chen, Q.-Y., Zhang, Y.-J., Liu, Z., Xiong, H., Guo, Z., Mai H.-Q. & Liu S. Detection of nasopharyngeal cancer using confocal Raman spectroscopy and genetic algorithm technique. *Journal of biomedical optics*. 17, 125003, 2012.

[5] Bangalore, A. S., Shaffer, R. E., Small, G. W. & Arnold M. A. Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy. *Analytical Chemistry*. 68, 4200–4212, 1996.

[6] Devos, O., Downey, G. & Duponchel, L. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food chemistry*. 148, 124–130, 2014.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Classification of Brazilian *Coffea canephora* cultivated by natives in the Amazon rainforest using portable near-infrared spectroscopy

M. R. Baqueta[1], E. A. Alves[2], P. Valderrama[3,*], J. A. L. Pallone[1,**]

[1] University of Campinas – UNICAMP, Faculty of Food Engineering, Department of Food Science, Campinas, São Paulo, Brazil. michelbaqueta@gmail.com

[2] Empresa Brasileira de Pesquisa Agropecuária - Embrapa Rondônia, Porto Velho, Rondônia, Brazil. enrique.alves@embrapa.br

[3,*] Universidade Tecnológica Federal do Paraná – UTFPR, Campo Mourão, Paraná, Brazil. pativalderrama@gmail.com

[1,**] University of Campinas – UNICAMP, Faculty of Food Engineering, Department of Food Science, Campinas, São Paulo, Brazil. jpallone@unicamp.br

Keywords: Specialty coffee; Geographical identity; Miniaturized spectrometer.

1 Introduction

Brazil represents an important coffee producer of *Coffea canephora*. Among the main producers in the country, Rondônia is a state in the North region known for its localization in the Amazon rainforest region. In recent years, in addition to the production of traditional coffee growers in Rondônia, indigenous communities have received support from national agencies for the production of high-quality coffees. Brazilian *Coffea canephora* cultivated by natives in the Amazon rainforest receives several appeals and arguments due to its sustainable agroforestry production, agrochemical-free, with forest protection and financial support to indigenous. In this preliminary chemometric study, the major objective was to verify whether or not Brazilian *Coffea canephora* cultivated by natives could be discriminated from other producers in the same state using portable near-infrared spectroscopy (portable NIR). Supervised pattern recognition was applied using the partial least squares with discriminant analysis (PLS-DA) method [1].

2 Material and methods

100 genuine *Coffea canephora* samples representing the diversity of coffee in Rondônia were investigated. These samples were produced during 2020 and belonged to the following classes: n =50 cultivated by natives (class 1) and n=50 produced by other producers within the same state (class 2). Green coffees were subjected to the same roasting process (medium degree), were ground and the particle size was standardized to 20 Tyler mesh.

Reflectance NIR spectra of roasted and ground coffees were obtained using a portable microNIR™ 1700 from JDSU, in the range of 906–1676 nm, with a mean of 32 scans and 125 variables per spectrum. Three different aliquots of the sample were prepared and the spectrum of each aliquot was recorded. Pre-treatments and data analysis was carried out using Matlab R2019a software with the PLS_Toolbox computational package version 8.6. In the present study, each replicate was considered as one sample and different pre-treatments were applied to the original data matrix: spectra transformation into absorbance, multiplicative scatter correction (MSC), Savitzky–Golay smoothing with a window size of 5 points applying the first-order polynomial and first derivative with first-order polynomial. The chemometric classification was done by using PLS-DA method,

selecting randomly 70% of the samples from each class for the training set. The remaining samples were used as the prediction set (test set). A cross-validation method with contiguous blocks with 10 data splits was also applied. The optimal number of latent variables (LVs) was chosen according to the class border evaluation by Bayesian decision, variance explained in dependent variable (**y**), and sensitivity and specificity plots as a function of threshold [1]. Samples potentially outliers showing simultaneously high leverage and Q residuals in training and prediction sets were evaluated and removed [2]. The performance of the optimized model was estimated by the sensitivity and specificity. Moreover, using the variable importance in projection (VIP) scores, it was tried to verify the main regions of the spectrum responsible by the modeling [3].

3 Results and discussion

Figure 1 shows chemometric plots and Table 1 shows parameters related to the PLS-DA classification.

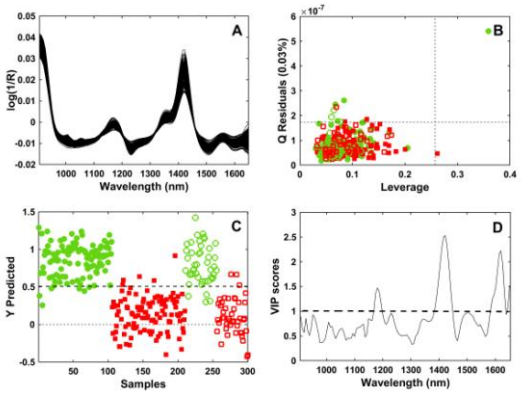


Figure 1 – Pre-treated NIR spectra (A); Evaluation of outliers (B); Estimated class values for training and prediction sets in the PLS-DA (C); Respective VIP scores (D). Different symbols and colors represent different samples: class 1 – coffee cultivated by natives (● training; ○ prediction); class 2 – coffee cultivated by other producers within the same state (■ training; □ prediction).

Table 1 – Variance explained in dependent variable "**y**" (V. E. in **y**), number of latent variables (N° LVs), sensitivity and specificity obtained in the final PLS-DA model.

V. E. in **y**	N° LVs	Training set		Cross-validation set		Prediction set	
		Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
72.22	18	0.942	0.952	0.846	0.924	0.911	0.933

4 Conclusion

It is possible to trace the identity, quality and origin of Brazilian *Coffea canephora* cultivated by natives in the Amazon rainforest in relation to the others *Coffea canephora* produced in the Rôndonia state. Using VIP scores, it is possible establish that the main components found in the coffee, such as chlorogenic acids, lipids, caffeine, proteins, and mainly carbohydrates, are the responsible by discrimination. Coffees have difference from a molecular point of view and PLS-DA is robust to distinguish them.

5 References

[1] Ferreira, M. M. C. *Quimiometria: conceitos, métodos e aplicações*, 1st ed., 2015.

[2] Baqueta, M. R., Coqueiro, A., Março, P. H., Valderrama, P. Multivariate classification for the direct determination of cup profile in coffee blends via handheld near-infrared spectroscopy. Talanta, 222:121526, 2021.

[3] Barbin, D. F., Felicio, A. L. de S. M., Sun, D.-W., Nixdorf, S. L., & Hirooka, E. Y. Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. Food Research International, 61:23–32, 2014.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Pixel-based identification of Raman hyperspectral data: Application to pharmaceutical tablets impurities detection

L. Coïc[1] P.Y. Sacré[1] A. Dispas[1] C. De Bleye[1] M. Fillet[2] C. Ruckebusch[3]
Ph. Hubert[1] E. Ziemons[1]

[1]University of Liege (ULiege), CIRM, ViBra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Liege, Belgium, laureen.coic@uliege.be

[2]University of Liege (ULiege), CIRM, MaS-Santé Hub, Laboratory for the Analysis of Medicines, Liege, Belgium

[3]University of Lille, CNRS, UMR 8516 LAboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement (LASIRE), F-59000 Lille, France

Keywords: Hyperspectral imaging; Spectral identification; Pixel selection; Essential Spectral Pixels.

1 Introduction

In the pharmaceutical field, analysis of tablets by Raman hyperspectral imaging is widely used for quality control purpose and has been now included in the general chapters of the European Pharmacopeia. However, data obtained can be consequent to analyze and implies to use appropriated chemometric tools. Most of the time, factorial decomposition methods (i.e Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS)) can be applied, excepted for the analysis of big data matrices, as well as in the presence of many constituents. Moreover, even when the composition is known, the MCR resolution can be challenging because some low variances sources can be diluted in the process of unmixing and can hardly be resolved unless information on the expected sample composition [1]. Moreover, it can exist minor compounds presents in a few pixels which can be missed in the MCR process. To bypass these limitations, one possibility can be to step back to the analysis of individual pixels, which somehow would be the most efficient method for database matching. The objective of the present study is to develop a pixel-based identification (PBI) approach to elucidate chemical composition of Raman hyperspectral images of complex pharmaceutical formulations. The proposed approach relies on the identification of Essential Spectral pixels (ESP) [2].

2 Material and methods

The proposed study was evaluated on both known and unknown tablets composition. The known formulations were made of polymorphic forms of carbamazepine (case 1) and piroxicam (case 2) to mimic minor compounds (from 0.1%w/w to 5%w/w). The seven unknown samples were falsified chloroquine (case 3) which were seized during the COVID-19 pandemic [3]. Raman hyperspectral imaging analyses of samples were performed with a LabRAM HR Evolution (Horiba scientific) equipped with an EMCCD detector (1600 × 200-pixels sensor) (Andor Technology Ltd.), a Leica 50x Fluotar LWD objective and a 785 nm laser with a power of 45mW at sample (XTRA II single frequency diode laser, Toptica Photonics AG). For both case 1 and 3, the whole tablet surface was analyzed with a 150 x 150 mapping and a step size of 87µm (total map size of ~13 x 13 mm²). For case 2, the middle of the tablet surface was mapped with a step size of 5.5 µm over a 5.5 x 5.5 mm²,

providing a 1000 x 1000 mapping. Three different approaches were then evaluated on each map to select pixels: (i) Kennard-Stone randomized, (ii) randomized selection and (iii) ESP selection, which were subsequently matched with the in-house database by using correlation coefficient (CC).

3 Results and discussion

For case 1 and case 2, the ESP approach compared to the other pixel-selection algorithms has shown the best results in terms of correlation coefficient but also with the smallest analysis time, with 50 seconds for the classical data size and 2 minutes for the big map size. The ESP approach was thus applied on falsified medicines and enabled to get the entire sample composition even for complex formulation (from 4 to 9 chemical compounds) with correlation coefficient superior to 0.80.

After gathering the ESP, a classical least squares was applied and allowed to show that even chemical information localised in a unique pixel had been elucidated, as it can be seen in Figure 1.

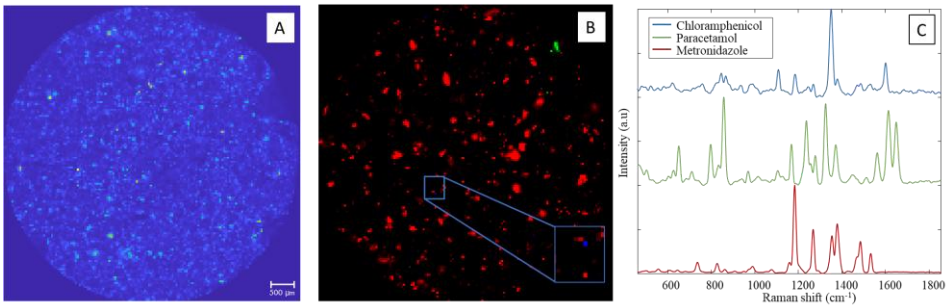


Figure 1 – Results of the ESP approach. A) Initial map. B) Representation of 3 compounds elucidated by the strategy. C) Representation of the different pure spectra obtained by the strategy.

Moreover, thanks to the inherent property of the convex hull, it had been also possible to not lose any chemical information. Indeed, after removing 3 compounds from the database, a PCA on the not-matched ESP allowed to show that they were kept during the approach and easily separated from the noise.

4 Conclusion

The proposed study highlighted the potential of the PBI approach for chemical identification purposes. It has been shown that, for known samples, both tiny and huge amount of data can be analyzed without the need of the entire map, by selecting only a few percentages of pixels (~8% of the initial data). The proposed methodology allowed to keep even the chemical information which was not in the in-house database which is very interesting in case of falsified medicines purposes. The global conclusion of this study is about the potential applicability of the methodology to other hyperspectral imaging techniques or matrices. Indeed, thanks to the inherent properties of the essential spectral pixel algorithm, the only requirement for PBI is to have at least one pure pixel by component. In case of mixed spectra, the use of the ESP could be a pre-processing step like, to reduce data dimensionality, which has been successfully demonstrated in the study [2].

5 References

[1] M. Boiret, A. de Juan, N. Gorretta, Y.M. Ginot, J.M. Roger, Distribution of a low dose compound within pharmaceutical tablet by using multivariate curve resolution on Raman hyperspectral images, J. Pharm. Biomed. Anal. 103 (2015) 35–43.
[2] M. Ghaffari, N. Omidikia, C. Ruckebusch, Essential Spectral Pixels for Multivariate Curve Resolution of Chemical Images, Anal. Chem. 91 (2019) 10943–10948.
[3] C.A. Waffo Tchounga, P.Y. Sacre, P. Ciza, R. Ngono, E. Ziemons, P. Hubert, R.D. Marini, Composition analysis of falsified chloroquine phosphate samples seized during the COVID-19 pandemic, J. Pharm. Biomed. Anal. (2020) 113761.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Product design based on Latent Variable Model Inversion: new tools for process exploration and optimization

D. Palací-López[1], P. Facco[2], M. Barolo[2], A. Ferrer[3]

[1] International Flavors & Fragrances Inc. (IFF), Avda. Felipe Klein, 2, 12580, Benicarló, Spain; daniel.palaci@iff.com

[2] Computer-Aided Process Engineering Laboratory (CAPE-Lab), Department of Industrial Engineering, University of Padova, Via Marzolo 9, Padova, PD, 35131, Italy; pierantonio.facco@unipd.it (P.F.), max.barolo@unipd.it (M.B.)

[3] Multivariate Statistical Engineering Group, Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Camino de Vera s/n, 7A, 46022, Valencia, Spain; aferrer@eio.upv.es

Keywords: Partial least-squares (PLS), Latent Variable Regression Model Inversion, Quality by design (QbD), Optimization.

1 Introduction

The quality-by-design (QbD [1]) initiative promotes the use of science-based methodologies to deliver products that meet the desired specifications by adapting to changes in raw materials and processing conditions without compromising the outputs' quality. This can be achieved by identifying the so-called design space (DS), i.e. "the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality." [2]. To this end, methods based on latent variables (LVs) can deal with big data in the Industry 4.0, where happenstance data (i.e. neither from experimentation nor causal in nature) are routinely collected and first principle models cannot be properly defined or verified, thus making data-based empirical models the best alternative.

In order to apply the QbD initiative in this context, two distinct strategies have been proposed in the literature: (a) estimating the DS as a whole though the so-called null space (NS [3,4]), while accounting for the uncertainty in its estimation [5,6], and (b) solving an optimization problem [7] to obtain a single set of process conditions within the DS. However, the former approach, as presented in past works, does not provide an analytical expression for either the NS or the confidence region within which the DS is expected to be found, and relies on the existence of at least one set of inputs that provides exactly the desired values for all of the considered outputs. The later, consequently, cannot make use of such information. In this work, an analytical definition of the estimation of the DS and its confidence region limits are suggested, which can also be applied to quality attributes defined as linear combinations of outputs. Afterwards, the implications and advantages of these tools for the DS estimation and the optimization problem are presented.

2 Theory

Let $\mathbf{X}$ [$N \times M$] be matrix of input variables and $\mathbf{Y}$ [$N \times L$] a matrix of output variables. Partial Least Squares (PLS) regression [8] can be used to predict $\mathbf{Y}$ from the A-dimensional subspace associated to $\mathbf{X}$ such that its covariance with $\mathbf{Y}$ is maximized. The model structure can be expressed as:

$$\begin{aligned} \mathbf{X} &= \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{T} \cdot \mathbf{Q}^{\mathrm{T}} + \mathbf{F} \\ \mathbf{T} = \mathbf{X} \cdot \mathbf{W} \cdot (\mathbf{P}^{\mathrm{T}} \cdot \mathbf{W})^{-1} &= \mathbf{X} \cdot \mathbf{W}^{*} \end{aligned} \quad (1)$$

where $\mathbf{T}$ [$N \times A$], $\mathbf{P}$ [$M \times A$] and $\mathbf{E}$ [$N \times M$] are the matrices associated to the $\mathbf{X}$ scores, loadings and residuals, respectively; $\mathbf{Q}$ [$L \times A$] and $\mathbf{F}$ [$N \times L$] the corresponding loadings and residuals matrices associated to $\mathbf{Y}$; and $\mathbf{W}$ [$M \times A$] is the weighting matrix. Here, A is, at most, equal to the rank of $\mathbf{X}$.

When the rank of $\mathbf{Y}$ is lower than A, and given a desired set of inputs $\mathbf{y}_{DES}$ [$L \times 1$], the direct inversion provides a single set inputs $\mathbf{x}_{DI}$ [$M \times 1$] that, according to the model, provides $\mathbf{y}_{DES}$, and can be obtained [3] as $\mathbf{x}_{DI} = \mathbf{P} \cdot \boldsymbol{\tau}_{DI} = \mathbf{P} \cdot \mathbf{Q}^{T} \cdot (\mathbf{Q} \cdot \mathbf{Q}^{T})^{-1} \cdot \mathbf{y}_{DES}$. Additional sets of inputs belonging to the NS, $\mathbf{x}_{NS}$ can be obtained by obtained as $\mathbf{x}_{NS} = \mathbf{P} \cdot \mathbf{G} \cdot \mathbf{r}$, where $\mathbf{G}$ is a matrix that contains the left singular vectors associated to the zero singular values of $\mathbf{Q}$, and $\mathbf{r}$ is a vector of random real values [3], or by moving from $\boldsymbol{\tau}_{DI}$ in an increment equal to $\Delta\boldsymbol{\tau}$ such that $\mathbf{Q} \cdot \Delta\boldsymbol{\tau} = \mathbf{0}$ [4]. Alternatively, an optimization problem as formulated in [7] can be resorted to in order to obtain similar results.

3 Material and methods

The proposed method for the estimation of the DS, based almost exclusively on Eq.1 itself, allows obtaining an analytical expression for the NS and the confidence region that contains it, for both outputs and linear combinations of them. This method can also be used to transfer constraints on the original variables to the latent space, and their application for the estimation of the DS, the definition of an experimental region within which it is expected to lay, and the re-formulation of the optimization problem in the latent space are compared to the methods in [3-7], as done in [9,10].

4 Results and conclusions

The tools presented in this work present several advantages when compared to previous approaches. They provide an analytical expression for the NS and its (non-linear) confidence limits for each and all outputs and their linear combinations (the advantages of this are presented in [9]) considered that does not depend on the existence of the direct inversion. These analytical expressions permit their use for both DOE (delimitation of the experimental space) and optimization (definition of constraints within and outside the objective function) and, as discussed in [10], allow explicitly addressing some issues that are implicit, but unapproachable with previous tools, in the formulation of the optimization problem in the latent space.

5 References

[1] J.M. Juran: *Juran on Quality by Design: The New Steps for Planning Quality Into Goods and Services*. S. Schuster, 1992.

[2] ICH Expert Working Group. ICH Pharmaceutical Development Q8. *ICH Harmon. Tripart. Guidel*. 8, 1-28, 2009.

[3] Jaeckle, C.M. & MacGregor, J.F. Industrial applications of product design through the inversion of latent variable models. *Chemom. Intel. Lab. Syst*. 50(2), 199-210, 2000.

[4] García-Muñoz, S., Kourti, T., MacGregor, J.F., Apruzzese, F. & Champagne, M. Optimization of batch operating policies. Part I. handling multiple solutions. *Ind. Eng. Chem. Res*. 45(23), 7856-7866, 2006.

[5] Facco, P., Dal Pastro, F., Meneghetti, N., Bezzo, F. & Barolo, M. Bracketing the design space within the knowledge space in pharmaceutical product development. *Ind. Eng. Chem. Res*. 54(18), 5128-5138, 2015.

[6] Bano, G., Facco, P., Bezzo, F. & Barolo, M. Probabilistic design space determination in pharmaceutical product development: a Bayesian/latent variable approach. *AIChE J*. 64(7), 2438-2449, 2018.

[7] Tomba, E., Barolo, M. & García-Muñoz, S. General framework for latent variable model inversion for the design and manufacturing of new products. *Ind. Eng. Chem. Res*. 51(39), 12886-12900, 2012.

[8] Wold, S. & Sjostrom, M. PLS-Regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst*. 58, 109-130, 2001.

[9] Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. New tools for the design and manufacturing of new products based on latent variable model inversion. *Chemom. Intel. Lab. Syst*. 194, 103848, 2019.

[10] Palací-López, D., Villalba, P., Facco, P., Barolo, M. & Ferrer, A. Improved formulation of the latent variable model inversion–based optimization problem for quality by design applications. *Journal of Chemometrics*. 34, e3230, 2020.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Beneficial Features of Procrustes Cross Validation

O. Rodionova[1], S. Kucheryavskiy[2], S. Zhilin[3], A. Pomerantsev[4]

[1] Semenov Federal Research Center for Chemical Physics, RAS, Moscow, Russia, rcs@chph.ras.ru

[2] Aalborg University, Department of Chemistry and Bioscience, Esbjerg, Denmark, svkucheryavski@gmail.com

[3] CSort Ltd., Barnaul, Russia, szhilin@gmail.com

[4] Semenov Federal Research Center for Chemical Physics, RAS, Moscow, Russia, forecast@chph.ras.ru

Keywords: Procustes cross-validation, short data-set, global and local PCA models

Abstract

Short description of the Procrustes cross- validation (PCV) method [1], which is an alternative to the traditional cross-validation (CV), is presented. Its main idea is to measure variations among the local CV models and to introduce this variation to the calibration set, creating a new pseudo-validation set (PCV-set).The distinct features of the PCV-set, are as follows

PCV-set enables validation of a global model, unlike the cross-validation, where each segment is validated by a corresponding local model;

- the results of the model performance are identical to what we can get using the conventional CV, but there is no need to repeat CV iterations more than once.

In practice, the PCV set can be thought of as a second copy of the training set, which contains different variability in comparison to the original data. The following characteristics as scores, residuals, orthogonal and score distances can be calculated for the PCV-set. Thus, PCV is a new approach, which is an intermediate method between test validation and CV.

A special case of the PCV application is analysis of short datasets where each sample is important and, therefore, cannot be removed in line with the conventional CV procedure. The advantages of the PCV method are shown on several real-world examples. PCV is implemented as MATLAB and R code and can be freely downloaded from [2].

References

[1] S. Kucheryavskiy, S. Zhilin, O. Rodionova, A. Pomerantsev, Procrustes Cross-Validation -- A Bridge between Cross-Validation and Independent Validation Sets, Anal. Chem. 92, 11842-11850, 2020.

[2] GitHub: https://github.com/svkucheryavski/pcv

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Robust variable selection in the framework of classification with label noise and outliers: applications to spectroscopic data in agri-food

A. Cappozzo[1] L. Duponchel[2] F. Greselin[1] T. B. Murphy[3]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy, andrea.cappozzo@unimib.it, francesca.greselin@unimib.it

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

[3] School of Mathematics & Statistics and Insight Research Centre, University College Dublin, Dublin, Ireland, brendan.murphy@ucd.ie

Keywords: Variable Selection; Robust classification; Label noise; Outlier detection; Near infrared spectroscopy; Mid infrared spectroscopy; Agri-food

1 Introduction

Near-infrared (NIR) and mid-infrared (MIR) spectroscopy have nowadays become a standard analytical practice in countless fields, being fast and non-invasive techniques for promptly characterizing samples of interest [1-2]. Within this context, variable selection methods are notably appealing, as adjacent features in spectroscopic data are often correlated. Unfortunately, standard methods are not robust against noisy samples, and whenever adulterations occur the reliability of the entire analysis may be jeopardized. Motivated by this, we introduce a robust model-based method that simultaneously performs variable selection, outliers and label noise detection.

2 Theory

The proposed robust variable selection method is called *stepwise REDDA*: a classification rule is constructed in a step-wise manner by considering the inclusion of extra variables and also the removal of existing variables to/from the model, conditioning on their discriminating power. Robustness is achieved by means of impartial trimming [3], an appealing technique for robust parameter estimation in which the least likely samples are not included in the fitting process. In this way, no model assumption is imposed to the noise component. The wavelength selection is based on a robust information criterion, that accounts for the possible presence of outliers and label noise in the dataset. The procedure iterates between variable addition and removal until two consecutive steps have been rejected, then it stops. In this way, the number of relevant frequencies necessary to build the classification rule is automatically inferred, and it needs not be a priori specified.

3 Material and methods

Our novel methodology is employed for performing classification of corrupted spectra. Specifically, the considered dataset encompasses NIR measurements of homogenized meat samples belonging to five different types, acquired in reflection mode from 400 to 2498 nm. Four corrupted samples are included in the training set, mimicking the ones considered in the Chimiométrie 2005 chemometric contest [4]. The resulting learning scenario is reported in Figure 1.

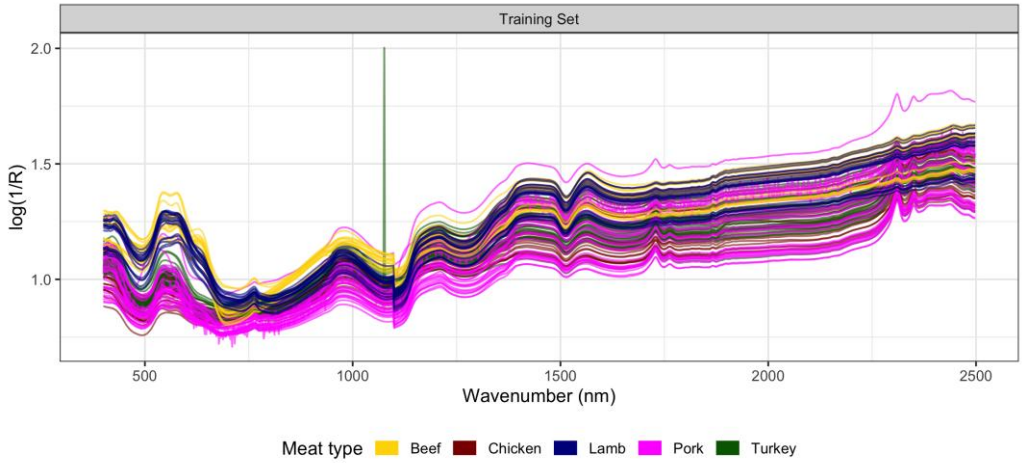


Figure 1 – Meat dataset: visible and near infrared spectra of five homogenized meat types

4 Results and discussion

The classification accuracy on the test samples is detailed in Table 1, where we compare the predictive performance of our methodology with PLS-DA and standard variable selection criteria used in chemometrics, namely variable importance in projection (VIP) and selectivity ratio (SR). Stepwise REDDA not only displays higher predictive power, but the 4 corrupted spectra are also automatically identified within the selection algorithm. In addition, our method selects only six wavelengths (634 nm, 672 nm, 676 nm, 786 nm, 1072 nm and 1076 nm) deemed to be relevant for classification, while VIP and SR retain a total of 407 and 463 frequencies, respectively.

Table 1 – Classification accuracy for different methods, meat dataset

Method	% correctly predicted
Stepwise REDDA	93
PLS-DA	87.7
VIP	89.5
SR	73.7

5 Conclusion

Mislabeling and sample corruption are issues oftentimes overlooked in analytical chemistry: a method like stepwise REDDA that accomplishes variable selection while automatically protecting against potential adulteration seems particularly promising for the spectroscopic field.

6 References

[1] C. Pasquini, Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. *Anal. Chim. Acta. 1026 2018*, pp. 8–36.

[2] R. Valand, S. Tanna, G. Lawson, L. Bengtström, A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations. *Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess. 37, 2020*, pp. 19–38.

[3] A. Gordaliza, On the breakdown point of multivariate location estimators based on trimming procedures. *Stat. Probab. Lett. 11, 1991*, pp. 387–394.

[4] J.A. Fernández Pierna, P. Dardenne, Chemometric contest at 'Chimiométrie 2005': A discrimination study, *Chemom. Intell. Lab. Syst. 86, 2007*, pp. 219–223.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

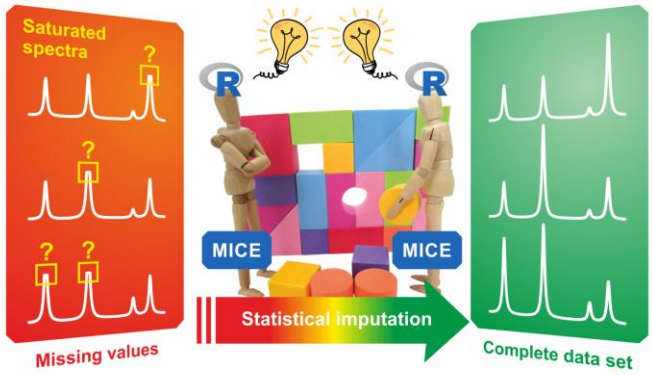
Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon?

Ludovic Duponchel[(1)], Alessandro Nardecchia[(1)], Vincent Motto-Ros[(2)]

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

[(2)] Institut Lumière Matière, UMR 5306, Université Lyon 1 - CNRS, Université de Lyon 69622 Villeurbanne, France.

Keywords: saturation; statistical imputation; multivariate analysis; chemometrics; spectroscopic imaging.



We have all been confronted one day by saturated signals observed on acquired spectra, whatever the technique considered. A saturation, also known as clipping in signal processing, is a form of distortion that limits a signal once it exceeds a threshold. As a consequence, clipped or saturated bands with their characteristic plateau present numerical values that do not correspond to the analytical reality of the analyzed sample. Of course, analysts know that they cannot consider these erroneous values and therefore reconsider either sample preparation or instrument settings. Unfortunately, there are many experiments today (and this is the case of spectroscopic imaging) for which we will not be able to fight against the saturation effect that will undeniably be observed on the acquired spectra. The aim of this presentation is first to show why it is important to correct these saturation effects at the risk of having a biased view of the sample and more specifically in the context of multivariate data analysis. In a second step, we will look at strategies for managing saturated bands. An original concept will then be presented by considering saturated values as missing ones. A statistical imputation strategy[1–5] will then be implemented in order to recover the information lost during the measurement. All imputation calculations have been done under the R environment using MICE, an open source R package. Source code[6] and documentation can be found at https://github.com/amices/mice.

References

(1) Buuren, S. van. *Flexible Imputation of Missing Data*; Chapman and Hall/CRC, 2012. https://doi.org/10.1201/b11826.

(2) Scheuren, F. Multiple Imputation: How It Began and Continues. *The American Statistician* **2005**, *59* (4), 315–319. https://doi.org/10.1198/000313005X74016.

(3) Multiple Imputation for Nonresponse in Surveys. In *Wiley Series in Probability and Statistics*; Rubin, D. B., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1987; pp i–xxix. https://doi.org/10.1002/9780470316696.

(4) Rubin, D. B. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association* **1996**, *91* (434), 473–489. https://doi.org/10.2307/2291635.

(5) Schafer, J. L. *Analysis of Incomplete Multivariate Data*; Chapman and Hall/CRC, 1997. https://doi.org/10.1201/9780367803025.

(6) Buuren, S. van; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **2011**, *45* (3). https://doi.org/10.18637/jss.v045.i03.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Assessing different facets of SIMCA modelling: decision rule, parameter optimization and their interplay

R. Vitale[1] M. Ahmad[1,2] V. Carboni[2] M. Cocchi[2]

[1]Université de Lille, LASIR - Laboratoire de Spectrochimie Infrarouge et Raman, Lille, France, rvitale86@gmail.com.

[2]Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Modena, Italy, marina.cocchi@unimore.it

Keywords: Class Modeling; SIMCA; AUROC; Classification Rules

1 Introduction

SIMCA [1-2] is a very popular and powerful class modeling method, although not yet entirely understood and exploited at its full potential. It is based on building a disjoint principal component analysis (PCA) model for each of the investigated classes and its underlying classification rule is defined on the basis of the distance of every sample from (Orthogonal Distance – OD) and within (Scores Distance – SD) the model space of the concerned category. The way these distance measures are combined, and the distributional assumptions on which this classification rule is based lead to different implementations of the methodology. These aspects have been recently revised in [3], even if such a survey did not directly focus on another critical issue, *i.e.*, the optimization of the SIMCA model tunable parameters: the class subspace dimensionality/complexity and the significance level used to define the distance boundary. For this reason, the main aim of this work is to assess the interplay between SIMCA version and SIMCA model adjustment strategy.

2 Theory

For the purpose outlined in the previous section, a comparative study was conducted as summarized in Figure 1. Data were split into calibration and validation sets several times (by Latin partition bootstrapping) in order to estimate the variation of the prediction performance for each tested approach. Four distinct SIMCA versions (namely, two variants of the so-called *alternative* SIMCA – alt-SIMCA [5] – combined index-based SIMCA – CI-SIMCA [6] – and Data Driven SIMCA – DD-SIMCA [3]) and three different SIMCA model optimization strategies were considered: i) significance level fixed at 95% and class model complexity optimized in cross-validation according to a "rigorous" criterion (*i.e.*, by maximizing the classification sensitivity); ii) significance level fixed at 95% and class model complexity optimized in cross-validation according to a "compliant" criterion (*i.e.*, by maximizing the classification efficiency) and iii) simultaneous significance level and model complexity tuning through the Receiver Operating Characteristic (ROC) curve-based

procedure proposed in [4].

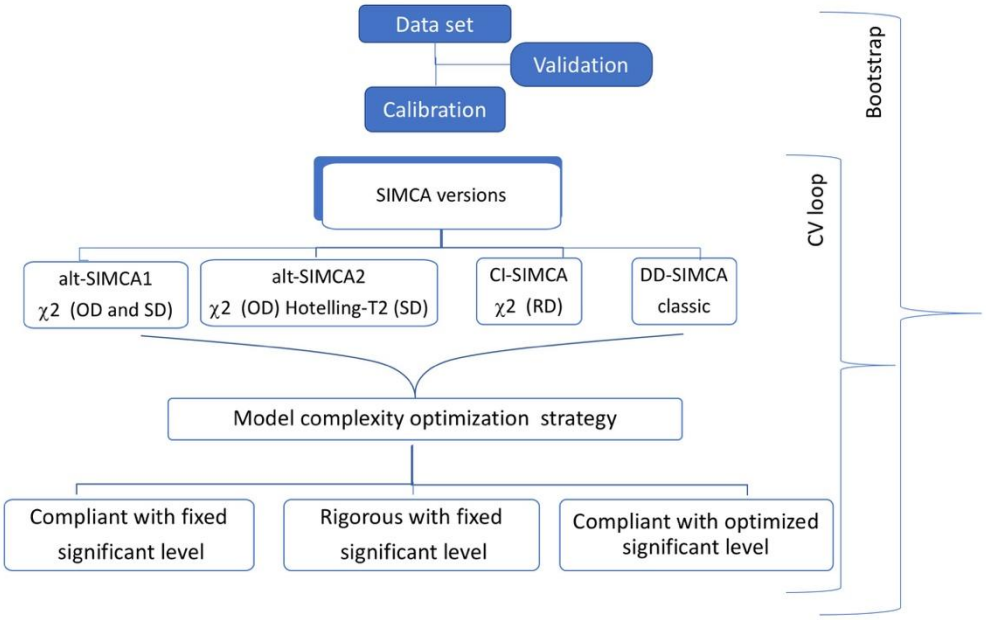


Figure 1 –Flowchart of the comparative assessment.

3 Materials and methods

The comparison was carried out on a dataset consisting of cell morphology descriptors [4]. To evaluate the performance of the different approaches, a single reference class was modelled (the one most overlapping with all the others) and the number of training observations was varied by representative sub-sampling. The effects of the SIMCA implementation, of the SIMCA model optimization strategy and of their interaction on the sensitivity, specificity and efficiency of the final classification were assessed by ANOVA.

4 Results and discussion

The results were evaluated by comparing the SIMCA model predictive performance indices obtained for the external test set at the various bootstrapping iterations. The simultaneous optimization of both SIMCA model dimensionality and significance level was found to be critical to achieve a reasonable compromise between classification sensitivity and specificity and also resulted in more parsimonious class models. On the other hand, the definition of the class acceptance rule seemed to have a limited impact on the classification figures of merit, albeit affecting the optimal class subspace complexity.
The evaluation of the subsampling effect is currently in progress.

5 References

[1] Wold, S. Pattern Recognition by Means of Disjoint Principal Components Models. Pattern Recogn. 1976, 8, 127-136. https://doi.org/10.1016/0031-3203(76)90014-5.

[2] Cocchi, M., Biancolillo, A., & Marini, F. (2018). Chemometric methods for classification and feature selection doi:10.1016/bs.coac.2018.08.006

[3] Pomerantsev A.L., Rodionova O.Y. Popular decision rules in SIMCA: Critical review. J. Chemometrics. 2020;34:e3250. https://doi.org/10.1002/cem.3250

[4] Vitale, R., Marini, F., Ruckebusch, C. Anal. Chem. 2018, 90, 10738−10747. DOI: 10.1021/acs.analchem.8b01270

[5] SIMCA Model Builder GUI (http://wiki.eigenvector.com/index.php?title=SIMCA_Model_Builder_GUI)

[6] Joe Qin S. Statistical process monitoring: basics and beyond. J Chemometr. 2003;17(8-9):480-502.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

p-SIMCA: a non-parametric probabilistic version of the SIMCA classifier

R. Vitale[1,†], F. Marini[2,†], C. Ruckebusch[1], A. Smolinska[3]

[1] Dynamics, Nanoscopy and Chemometrics (DyNaChem) Group, Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, LASIRE, CNRS, U. Lille, Cité Scientifique, F-59000 Lille, France, raffaele.vitale@univ-lille.fr, cyril.ruckebusch@univ-lille.fr

[2] Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy, federico.marini@uniroma1.it

[3] Department of Pharmacology and Toxicology, School of Nutrition, Toxicology and Translational Research in Metabolism, NUTRIM, Maastricht University Medical Center+, 6202 AZ Maastricht, The Netherlands, a.smolinska@maastrichtuniversity.nl

[†] These authors contributed equally

Keywords: class-modelling, non-parametric statistics, probabilistic classification, Soft Independent Modelling of Class Analogy (SIMCA), Kernel Density Estimation (KDE), Bayes' theorem.

1 Introduction

Nowadays, classification is one of the tasks most commonly addressed by users and practitioners of chemometrics. Classification problems are usually coped with by either discriminant (*e.g.*, Partial Least Squares Discriminant Analysis – PLSDA [1]) or class-modelling methods (*e.g.*, Soft Independent Modelling of Class Analogy – SIMCA [2]). These two families of approaches possess distinctive pros and cons, which render their use more or less appropriate depending on the specific nature of the conducted studies. A recent comprehensive survey of such pros and cons can be found in [3]. In recent years, though, in many fields of applied science like forensics, medical diagnostics and food authentication, a growing interest has been raised by so-called *probabilistic classification techniques*. Rather than directly assigning individual samples to a single or multiple categories of interest, these techniques provide, for every investigated object, characteristic probability values determining how strongly the possibility that it belongs to one or another class is supported by the analysed data. However, although several strategies have already been proposed or extended to serve this purpose, they were all originally designed to tackle discrimination issues. On the other hand, the possibility of combining the intrinsic advantages of class-modelling with the increased flexibility guaranteed by probabilistic classification has been far less explored. For this reason, a novel probabilistic version of the SIMCA algorithm, p-SIMCA, is here reported and outlined.

2 Materials and methods

Let $\mathbf{X}$ be an $N \times J$ dataset containing only the measurements collected for a particular category of samples. p-SIMCA is strictly based on the classical modelling approach underlying the original SIMCA classifier and encompasses the following four algorithmic steps:
1) $\mathbf{X}$ is decomposed by Principal Component Analysis (PCA) as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

where **T** ($N \times A$), **P** ($J \times A$) and **E** ($N \times J$) denote the scores, loadings and residuals matrices resulting from the factorisation of **X**, respectively, and A represents the number of retained latent factors;
2) for the objects of the training set, a combined distance index quantifying their degree of *outlyingness* with respect to the class subspace defined in Equation 1 is calculated through a cross-validatory procedure. More specifically, for a generic observation $\mathbf{x}^{\mathrm{T}}$ ($1 \times J$), this combined index is calculated as:

$$d = [(T^2/T^2_{\mathrm{lim}})^2 + (Q/Q_{\mathrm{lim}})^2]^{0.5} \quad (2)$$

being T^2 a statistic reflecting the (Mahalanobis) distance between the origin of the PCA model hyperplane and the projection of $\mathbf{x}^{\mathrm{T}}$ onto it, Q a statistic reflecting the perpendicular (orthogonal) distance between $\mathbf{x}^{\mathrm{T}}$ and the PCA model hyperplane, T^2_{lim} an empirical threshold for the T^2-statistic (usually corresponding to a significance level of 95%) and Q_{lim} an empirical threshold for the Q-statistic (usually corresponding to a significance level of 95%). This operation is also iterated for a representative amount of samples not belonging to the considered category;
3) the probability density functions associated to the two distinct series of d-values (in- and out-of-class, respectively) are retrieved by Kernel Density Estimation (KDE [4]). KDE parameter optimization is internally carried out by Maximum Likelihood Cross-Validation (MLCV [5]);
4) once the in- and out-of-class probability density functions have been obtained, Bayes' theorem is applied in order to compute the *a posteriori* probability that any new unlabelled sample belongs/does not belong to the modelled category.

3 Results and conclusions

p-SIMCA was tested in both simulated and real case-studies. The former allowed assessing the accuracy of the probability estimates returned by the algorithmic procedure in a wide variety of designed and controlled scenarios. The latter enabled the evaluation of the performance of p-SIMCA when applied to datasets produced by analytical platforms of different nature (*e.g.*, infrared spectroscopy, phase-contrast microscopy, *etc.*). The outcomes resulting from its application highlighted how p-SIMCA is capable of dealing with probabilistic classification tasks preserving the advantages inherent to class-modelling approaches – like the possibility of handling situations of pronounced unbalancedness/asymmetry among class sizes – differently from alternative Bayesian versions of classical SIMCA (see, for instance, [6]) which were all originally developed on the basis of discriminant-like sample assignment rules. Options for bypassing the collection of out-of-class sample measurements were also conceived and their feasibility explored and examined.

4 References

[1] Wold, S., Albano, C., Dunn, W., Esbensen, K., Hellberg, S., Johansson, E. & Sjöström, M. Pattern recognition: Finding and using regularities in multivariate data. In *Food Research and Data Analysis*, Applied Science Publishers Ltd., London, United Kingdom, 1st Edition, 1983.

[2] Wold, S. Pattern recognition by means of disjoint principal component models. *Pattern Recogn.* 8, 127-139, 1976.

[3] Biancolillo, A., Marini, F., Ruckebusch, C. & Vitale, R. Chemometric strategies for spectroscopy-based food authentication. *Appl. Sci.* 10, 6544, 2020.

[4] Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* 27, 832-837, 1956.

[5] Habbema, J., Hermans, J. & van den Broek, K. A stepwise discriminant analysis program using density estimation. In *COMPSTAT 1974: Proceedings in Computational Statistics*, Physica Verlag, Wien, Austria, 1st Edition, 1974.

[6] van der Voet, H., Coenegracht, P. & Hemel, J. New probabilistic versions of the SIMCA and CLASSY classification methods: Part 1. Theoretical description. *Anal. Chim. Acta* 192, 63-75, 1987.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

The Needle in the Haystack, or Microplastics in Natural Samples. What is More Complex to Find? Analytical Strategies using Raman and Mid-Infrared Imaging.

José Manuel Amigo[1,2], Imanol Torre[2], Cristina García Florentino[2], Alba Benito[2], Kepa Castro[2]

[1] IKERBASQUE, Basque Foundation for Science. Plaza Euskadi 5. 48009 Bilbao. Spain

[2] IBeA. Department of Analytical Chemistry. University of the Basque country. 48940 Leioa. Spain

Keywords: Microplastics, Raman Imaging, Mid-Infrared imaging, Generalized Least Squares, Grid-Search, DataBases.

Abstract

It is not a novelty that microplastics analysis in natural samples has become one of the major interest and concerns in Environmental Sciences. Despite this fact, the challenges postulated concerning the sampling, detection and identification are far for being well established. This talk will focus on the analytical challenges of detection and identification of microplastics in natural waters and sediments by using Raman Imaging and Mid-Infrared Imaging together with the state-of-the-art algorithmic strategies to analyze the overwhelming amount of information collected in such samples. Concepts like the spatial and spectral limit of detection in a multivariate scenario, construction of databases and algorithms will be revisited and discussed in order to be able to establish new strategies to be more accurate and less time-demanding. Two examples will be presented: The construction of a Transmittance-based Mid-Infrared database and the usage of classification algorithms [1] and the finding of a single-pixel containing a microplastic in a Raman Image. Guided by these two examples, different benefits, drawbacks, challenges, and breakthroughs will be discussed.

References

[1] V. H. da Silva, F. Murphy, J. M. Amigo, C. Stedmon, J. Strand, Classification and Quantification of Microplastics (<100 µm) Using a Focal Plane Array–Fourier Transform Infrared Imaging System and Machine Learning, Anal. Chem. 92 (2020) 13724–13733. https://doi.org/10.1021/acs.analchem.0c01324.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Spectral and spatial fusion: an interesting approach for classification in hyperspectral imaging

A. Nardecchia[1] R. Vitale[2] L. Duponchel[3]

[1] Univ. Lille, CNRS, UMR 8516 – LASIRe – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et l'Environnement, F-59000, Lille, France, alessandro.nardecchia@univ-lille.fr

[2] Univ. Lille, CNRS, UMR 8516 – LASIRe – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et l'Environnement, F-59000, Lille, France, raffaele.vitale@univ-lille.fr

[3] Univ. Lille, CNRS, UMR 8516 – LASIRe – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et l'Environnement, F-59000, Lille, France, ludovic.duponchel@univ-lille.fr

Keywords: 2-D stationary wavelet transform (2-D SWT), PLS-DA, classification, hyperspectral imaging.

1 Introduction

Hyperspectral image analysis is a constant developing branch of chemometrics. Despite its use in various fields, many limitations still hamper its broad utilization. In particular, while data analysis is commonly focused on spectral information, spatial details are still not really taken into consideration. Naturally, this leads to an incomplete use of the full potential that this kind of technique could exhibit for the exploration of complex matrices. Some works on the importance of using also the spatial information, have been published [1,2]. *Inter alia*, an interesting approach that has been recently investigated is the wavelet transform [3–5]. In a recent work, our group has shown the importance of fusing spectral and spatial information in such a way [6]. In fact, by the use of, for example, Principal Component Analysis, new features of the image structure are detectable when the spatial details are extracted by the use of wavelets and fused with the information carried by the spectra. The aim of this work is to describe new possibilities associated to the use of 2-D stationary wavelet transform (SWT 2-D) in this domain: more specifically, it will be shown that classification models (built with PLS-DA) that take into account also the spatial details, can lead to more accurate results.

2 Material and methods

When exploring a hyperspectral image (a three-dimensional data array), an unfolding step is generally required prior to any analysis step. This procedure leads to the complete loss of the spatial information encoded in the image. SWT 2-D is based on the use of particular filters that operate directly on the original array. As a result, four distinct sets of wavelet coefficients (for approximation features and horizontal, vertical and diagonal details) are obtained. The object of this first step is, thus, to capture spatial information. Then, these sets of coefficients are pretreated and fused with the original spectra. In this way, an augmented matrix containing both spectral and spatial information is obtained. A more accurate description of each step of this procedure can be found in [6]. Here, a simulated hyperspectral image will be exploited to assess the advantages of this approach over more standard methods. Four different geometrical shapes (rectangles and circles) were generated and associated to four different classes of pixels. More specifically: i) the

second and the fourth classes exhibit extremely overlapping spectral profiles, but different geometrical shapes, while ii) the third class, constituted by multiple rectangles, encompasses spectra that are linear combinations of those characteristic of the first and the second classes.

3 Results and discussion

As generally outlined in the previous paragraphs, by fusing spectral and spatial information, it is possible to obtain a better identification of specific image properties. As an example, Figure 1 shows how the proposed data analysis strategy permitted to better discriminate pixel classes 2 and 4 than through a standard PLS-DA model built using only the spectral information encoded in the simulated hyperspectral image. Also, class 3, that is completely misclassified by this latter approach, is better discerned when a wavelet transform is used.

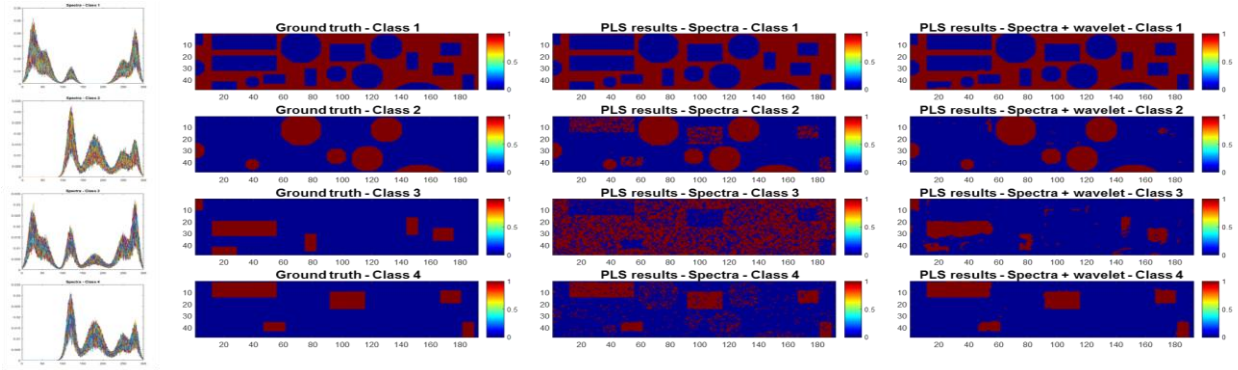


Figure 1 – Graphic representation of the PLS-DA results obtained using 1) only the spectra and 2) both the spectral and spatial information yielded by SWT 2-D.

4 Conclusion

The displayed results give a clear idea about how the combined use of spatial and spectral information extracted from hyperspectral images can lead to better classification models compared to those obtained when only spectral profiles are resorted to. This is especially true when the hyperspectral images under study capture objects of distinct shapes yielding similar instrumental responses. These preliminary results are currently being validated through the assessment of real datasets.

5 References

[1] S. Hugelier, O. Devos, C. Ruckebusch, On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis: Spatial constraints in HSI-MCR-ALS, J. Chemometrics. 29 (2015) 557–561. https://doi.org/10.1002/cem.2742.

[2] F. Jamme, L. Duponchel, Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis, Journal of Chemometrics. 31 (2017) e2882. https://doi.org/10.1002/cem.2882.

[3] G.P. Nason, B.W. Silverman, The Stationary Wavelet Transform and some Statistical Applications, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, Springer New York, New York, NY, 1995: pp. 281–299. https://doi.org/10.1007/978-1-4612-2544-7_17.

[4] M. Li Vigni, J.M. Prats-Montalban, A. Ferrer, M. Cocchi, Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA): Coupling 2D-WT to Multivariate Image Analysis (2D WT-MIA), Journal of Chemometrics. 32 (2018) e2970. https://doi.org/10.1002/cem.2970.

[5] M. Ahmad, R. Vitale, C.S. Silva, C. Ruckebusch, M. Cocchi, Exploring local spatial features in hyperspectral images, Journal of Chemometrics. 34 (2020). https://doi.org/10.1002/cem.3295.

[6] A. Nardecchia, R. Vitale, L. Duponchel, Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images, Talanta. 224 (2021) 121835. https://doi.org/10.1016/j.talanta.2020.121835.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Automated chemical rank determination by hybridizing dependency concept and permutation testing

E. T. Bayat[1] B. Hemmateenejad[1,2] K. Baumann[2] M. Akhond[1]

[1] Chemistry Department, Shiraz University, 71454, Shiraz, Iran, marjanbayat2323@gmail.com , b.hemmateenejad@tu-braunschweig.de, akhond@chem.susc.ac.ir

[2] Institute of Medicinal and Pharmaceutical Chemistry, University of Technology Braunschweig, 38108, Braunschweig, Germany, k.baumann@tu-braunschweig.de

Keywords: Permutation testing, Dependency concept, Nonparametric analysis, Principal component analysis; Chemical rank determination

1 Introduction

PCA is popularly recognized as an unsupervised multivariate exploratory data analysis tool. An important application of PCA is chemical rank analysis for determining the correct number of primary principal components (P-PCs). The redundant information is retained in secondary PCs (S-PCs). Most of the available rank estimation methods suffer from at least one the following three: (i) wide range variability in power to separate P-PCs from S-PCs, (ii) lack of an appropriate stopping rules to figure out of P-PCs and (iii) entrance of user's ingenuity or imagination. Dependency concept is empowered by its hyphenation with permutation testing approach (to directly extract data-driven threshold) as a statistical stopping rule for automating the discovery of P-PCs. So, two non-parametric approaches with minimum assumptions about data nature are combined. In this study, computation time is remarkably mitigated especially in massive data sets. Because this method need data shuffling for *just first PC* unlike other confidence interval methods (permutation based methods) and this decreases the computation time drastically.

2 Theory

Two dependency indies (DIs) are used: maximum information coefficients (MIC) [1] and distance correlation (DC) [2].The algorithm can be divided into 3 phases. In phase I, the null-distribution of the DIs are *just* calculated for *first PC*. The (1-α)×100th percentile (where α is confidence level) of the resultant DIs over n_{perm} times repeats is calculated ($I_{MIC,\alpha}$ and $I_{DC,\alpha}$). Residual data matrices (2nd phase) are calculated sequentially by removing the contribution of subsequent principal components:

$$\mathbf{R}_{i-1} = \mathbf{X} - \sum \quad (1)$$

where **X**, **t**, **p** and *i* stands for original data, score, loading vectors and the iteration index, respectively. Then DIs ($I_{MIC,i}$ or $I_{DC,i}$) are calculated in every $\mathbf{R}_{i-1}$. In the last phase, the significance of $I_{MIC,i}$ or $I_{DC,i}$ at ith iteration is statistically examined different from $I_{MIC,\alpha}$ or $I_{DC,\alpha}$. Null-hypothesis (H_0), where it is accepted if $I_{MIC,i} \leq I_{MIC,\alpha}$ (or $I_{DC,i} \leq I_{DC,\alpha}$). The performance of the proposed algorithm (Deperm) is compared to different confidence interval methods (CIMs) [3-5] to automatically rank estimation. Deperm performance is examined in different types of simulated data sets including: random, chromatography (Chrom), spectrophotometric monitoring of a

stepwise complex formation reaction (Spec-comp), and a two-step first-order consecutive kinetic reaction (Spec-kin). Finally, the different real data sets (Raman, Fluorescence spectroscopy and chronoampermetry) in ill-conditions are examined.

3 Results and discussion

Figure 1 is shown the results of applying Deperm on random (a, b) and Chrom (c, d) data sets. Similar trends are observed in the rest of simulated and real data sets. $I_{MIC,i}$ or $I_{DC,I}$, empirical *null*-distribution and 95% percentile (as a soft cutoff) as a function of PC number are illustrated by bar graph, red cross and yellow dot markers, respectively. As can be easily observed, DIs relying on P-PCs are dramatically stood out from the rest PCs due to their sensitivity to rank ordered shapes. But $I_{MIC,\alpha}$ and $I_{DC,\alpha}$ could be statistically used the significance of each PC. As a results, zero and four chemical rank are correctly estimated. The results are shown both $I_{MIC,\alpha}$ and $I_{DC\alpha}$ represent very small changes (around 1% relative changes) for 100 times repeated permutations of **X** or **R** matrices in all simulated and real situations. So, the high stability of $I_{MIC,\alpha}$ and $I_{DC,\alpha}$ suggests that single permutation on just **X** works well to test the significance PCs. It could lead to reduce computation time in simulated HPLC-DAD, Spec-comp and Spec-kin, respectively: 64%,62%, and 61%. The effect of different parameters is also studied especially in the presence of high level of different types of perturbers (e.g. homoscedastic noise, heteroscedastic noise, resolution in elution profile, minor component, collinearity). The results are proposed that the main controlling parameter in rank determination is collinearity problem. Deperm could successfully estimate true chemical rank especially in high correlated structure even in ill-condition in simulated data sets. In real data sets, the successful rank estimation of proposed method is arisen from stability of $I_{MIC,\alpha}$ and $I_{DC,\alpha}$ and high statistical power in differentiating between P-PCs and S-PCs. While these features are not observed in CIMs.

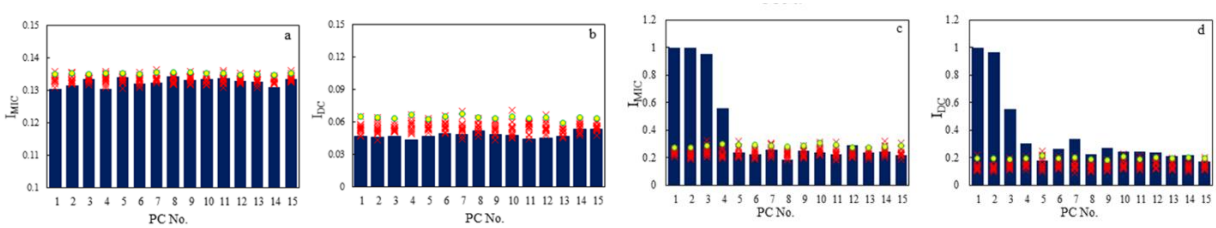


Figure 1 – Statistic dependency indices concerning of random data (a-b) and Chrom (c-d). $I_{MIC,i}$ or $I_{DC,I}$ (blue bars), the empirically estimated *null*-distributions (red cross), and the 95% percentiles of the estimated *null*-distributions (yellow circle) are represented. Permutation number is adjusted at 100 times.

4 Conclusion

The combination two non-parametric approaches including dependency concept and permutation testing could robustly result to *automatically* chemical rank estimation. The high stability of soft thresholds over all PCs even in high level of pertubers in simulated and real data sets could lead to use just single shuffling in original set. This could remarkably result to mitigate computation time.

5 References

[1] Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. & Sabeti, P.C. Detecting novel associations in large data sets, *Science*. 334, 1518–1524,2011.

[2] Székely, G.J., Rizzo, M.L & Bakirov, N.K. Measuring and testing dependence by correlation of distances, *Ann. Stat.* 35, 2769–2794, 2007.

[3] Dray, S. On the number of principal components: A test of dimensionality based on measurements of similarity between matrices, *Comput. Stat. Data Anal.* 52, 2228–2237, 2008.

[4] Endrizzi, I., Gasperi, F., Rødbotten, M &Næs, T Interpretation, validation and segmentation of preference mapping models, *Food Qual. Prefer.* 32, 198–209, 2014. .

[5] Vitale, R., Westerhuis, J.A., Naes, T., Smilde, A.K., de Noord, O.E & Ferrer, A. Selecting the number of factors in principal component analysis by permutation testing-Numerical and practical aspects, *J. Chemom.* 31, e2937, 2017.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Multi-exponential analysis with MCR slicing

O. Devos[1], M. Ghaffari[1], A. de Juan[2], M. Sliwa[1], C. Ruckebusch[1]
[1]U. Lille, LASIRE CNRS, France. [2]U. Barcelona, Department of Chemistry, Espagne
Olivier.devos@univ-lille.fr, m.ghaffari808@gmail.com, Anna.dejuan@ub.edu, Michel.sliwa@univ-lille.fr, Cyril.ruckebusch@univ-lille.fr

Keywords: MCR-ALS, Multiset, Constraints, Trilinearity, Time-resolved Fluorescence.

1 Introduction

Fluorescence spectroscopy encompasses a set of non-invasive, highly specific and extremely sensitive techniques. Among them, time-resolved fluorescence spectroscopy (TRFS) can be used to characterize the fluorescence lifetime, which is a feature of molecules, allowing to probe their interactions with the molecular environment. However, multi-exponential data analysis is required to determine the number of exponential components, their lifetimes and amplitude coefficients. Multi-exponential data fitting is usually applied despite issues existing in practice. Those are related to the ill-conditioned nature of the problem and translate into the need to provide good initial guess of the parameters, non-unique solution, parameter correlation… The larger the number of exponential components, the more similar their associated lifetimes, the lower the signal to noise, the more severe these issues are.

2 Theory

Multivariate curve resolution slicing (MCR slicing) is a new fit-free method that we propose for tailored exponential decomposition, as required for the analysis of TRFS data. MCR slicing relies on the construction of a row-wise augmented multiset data structure to allow the implementation of a bilinear-trilinear constraint setting in MCR-ALS analysis, as illustrated in Figure 1. The two-way bilinear measured data set is first split in two sub-matrices, D1 corresponding to the measurements at short time points (where the effect of convolution with the IRF is significant) and D2 at long time points (where the effect of convolution with IRF can be ignored). The row-wise augmented multiset structure is composed of D1 and of slab matrices returned by the slicing procedure applied to D2. A trilinear constraint is applied on the sliced time decay profiles to enforce exponential behavior. This structure and the implemented constraints allow the decomposition of TRFS data that are only partially describable by single-exponential features, which fits the nature of those data.

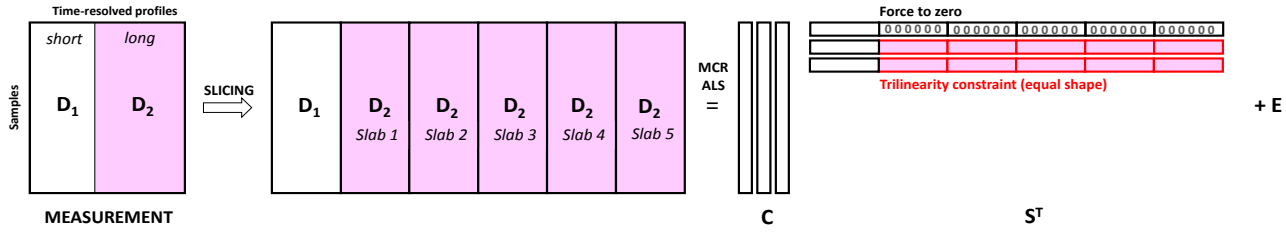


Figure 1 – Scheme of the MCR slicing procedure for multi-exponential curve resolution of TRFS data matrices. A MCR-ALS bilinear-trilinear decomposition is performed. The trilinearity constraint is applied per matrix and per component (only to the row-profiles extracted for $\boldsymbol{D_2}$ *and for the component that follow a multi-exponential shape). A force to zero constraint is applied to the component related to laser artifacts that are only observed at short time points.*

3 Material and methods

Three commercial fluorescent dyes were selected to prepare ternary mixtures. Those dies are carboxy derivatives of namely (i) ALEXA647 (Thermo Fisher Scientific, Invitrogen), (ii) ATTO655 and (iii) ATTO665 (ATTO-TEC GmbH). They show strongly overlapping spectral features for fluorescence, absorption around 640 nm and emission around 670 nm. However, these dyes are known to present mono-exponential fluorescence decay in water solutions with quite distinguishable respective lifetimes of (i) 1.08 ns, (ii) 1.8 ns and (iii) 2.9 ns. The design of experiment consisted of a simplex axial design with 3 extra points, as illustrated in Figure 2.

4 Results and discussion

We validated our approach by decomposing a data set composed of different mixtures of three commercial dyes measured at different emission wavelengths (see Figure 2). We will also show application to the data obtained on the fluorescent protein RsEGFP2. Altogether, a bilinear-trilinear MCR-ALS decomposition is performed, allowing the decomposition into individual components characterized by full time-resolved fluorescence profiles that are only partially describable by single-exponential features.

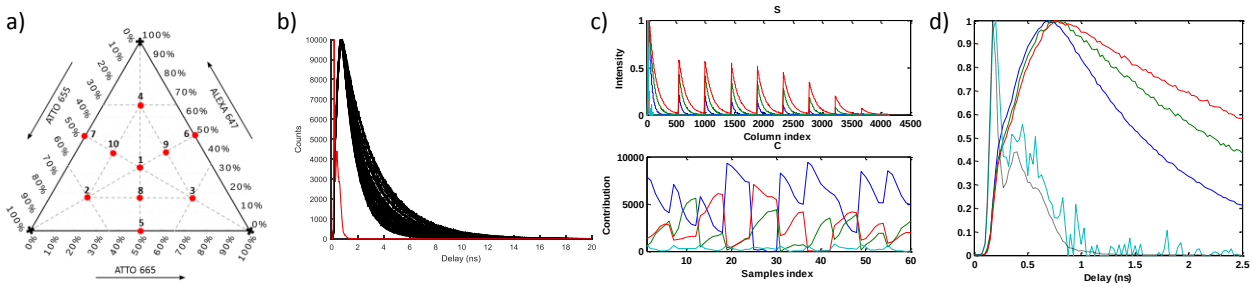


Figure 2 – (a) Dye mixtures, (b) corresponding TRFS data, (c) and (d): MCR-ALS results for the full augmented multiset and focusing on the extracted decays for short time dataset (D1).

5 Conclusion

Data slicing is a way to impose an exponential constraint on the profiles extracted through a trilinear decomposition of an originally bilinear data matrix. This matrix should first be reorganized (sliced). Interestingly, an alternative to trilinear decomposition of three-way data arrays can be provided by bilinear decomposition of multiset data matrix with MCR-ALS applying a trilinearity constraint. In this way, multi-exponential curve resolution is also achieved. Because constraints can be implemented in a very flexible way in MCR-ALS, MCR slicing allows overcoming some limitations that may appear with single set trilinear decomposition. In particular, component profiles that do not follow a purely exponential model can be encompassed in the data decomposition.

6 References

[1] Benabou et al., Nucleic Acids Research, 26 2019 6590

[2] Yramian et al., 326, Letters to Nature, 1987, 169.

[3] Engelsen and Bro, Journal Of Magnetic Resonance, 163, 2003, 192

[4] Gomez et al., Analytical Chemistry, 92, 2020, 9591

[5] Woodhouse et al, Nature Communications, 11, 2020, 4478.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Diesel cetane number prediction by data fusion of near-infrared and nuclear magnetic resonance spectroscopy

J. Buendia-Garcia[1] J. Gornay[2] M. Lacoue-Negre[3] S. Mas-Garcia[4] R. Bendoula[5] J.M Roger[6]

[1] IFP Energies Nouvelles, Rond Point de l'échangeur de Solaize, France, jhon.buendia-garcia@ifp.fr

[2] IFP Energies Nouvelles, Rond Point de l'échangeur de Solaize, France, julien.gornay@ifp.fr

[3] IFP Energies Nouvelles, Rond Point de l'échangeur de Solaize, France, marion.lacoue-negre@ifp.fr

[4] INRAE, 361 rue Jean François Breton - Montpellier, France, jean-michel.roger@inrae.fr

[5] INRAE, 361 rue Jean François Breton - Montpellier, France, ryad.bendoula@inrae.fr

[6] INRAE, 361 rue Jean François Breton - Montpellier, France, silvia.mas-garcia@inrae.fr

Keywords: Data fusion, NIR, NMR, Cetane number, diesel, total effluent, VGO, catalytic conversion.

1 Introduction

Among the different oil refining processes employed to obtain fuel, catalytic conversion processes such as hydrocracking (HCK), hydrotreating (HDT), and fluid catalytic cracking (FCC) play an essential role. These allow the transformation of low-value streams, such as vacuum gasoils (VGOs), into high-value streams, such as middle distillates (Diesel and Kerosene) [1]. Spectroscopic techniques such as near-infrared (NIR) and nuclear magnetic resonance (NMR) can be used to develop predictive models as a reliable way to estimate the physicochemical properties of petroleum-based fuels [2, 3]. When used together, they can provide complementary information to improve the performance of separately generated models [4]. In this study, the NIR and ^{13}C NMR spectra of 93 total effluents obtained from the different catalytic conversion processes outlined were fused for diesel cetane number estimation. Different prediction models were developed using the three most common data fusion strategies (low, mid, and high-level).

2 Material and methods

In this study, 24 different VGOs were used to obtain 93 total effluent streams from the previously mentioned catalytic conversion processes (61 from HCK, 26 from HDT, and 6 from FCC). Each of these total effluent streams was distilled according to ASTM D2892-20 [5] to recover the diesel cut (250°C-370°C). Finally, the cetane number is measured according to the standard ASTM D613-01 [6] on the diesel recovered from each distillation.

The NIR spectra were recorded with a Fourier Transform Near-Infrared spectrometer (FT-NIR) MATRIX-F (Bruker, Optik GmbH) within the range of 9090 - 4600 cm-1 using an immersion probe with an optical path fixed at 2 mm. To ensure the samples liquid state and homogeneity, they were heated to 60°C and manually shaken before performing the NIR analysis.

The ^{13}C NMR spectra were obtained with a Bruker Advance 600 MHz spectrometer operating at 14.1 T, using a 5 mm BBI probe. The spectra acquisition conditions were an analysis temperature of 50°C, 1024 scans, 15 µs pulse, relaxation delay of 5 s, an acquisition time of 0.769 s, and a spectral

width of 42613.64 Hz. The solvent used in the samples' dissolution was deuterated chloroform (CDCl3), with 0.3% wt/wt of Fe(acac)3.

The models generated were classified into two main categories: individual and data fusion models. In constructing the models, two regression methods were used and compared (PLS & SVM). Four methods of variable selection (sRatio, VIP, GA, iPLS) were also applied. From each of the developed models, 6 statistical parameters were obtained (RMSEC, RMSECV, RMSEP, R^2C, R^2CV, R^2P), which were utilized to evaluate and select the best performing model.

3 Results and discussion

Table 1 summarizes the 6 statistical parameters used in the final comparison of the 7 selected models. From this table it can be seen that both individual and data fusion models have a good performance being the mid-level PLS model the one with best performing. The most significant improvement is found in the RMSECV, which has a reduction of 1.1 and 1.3 units with respect to the individual NIR and ^{13}C NMR models respectively. Likewise, the determination coefficient improvement is evident, which increases from 0.95 (NIR model) and 0.94 (13C NMR model) to 0.99. In addition to improving the model's predictive capacity and compensating effects due to the presence of atypical data, this model reduces the number of latent variables necessary for the optimal prediction of the diesel cetane number (6 LV Vs. 4 LV).

Table 1 – Statistical parameters of 7 models selected for final comparison

								LV used in data fusion		LV Final model
		RMSEC	R^2C	RMSECV	R^2CV	RMSEP	R^2P	NIR	^{13}C NMR	
Single Models	NIR	1.3	0.98	2.1	0.95	1.5	0.98	N/A	N/A	6
	^{13}C NMR	1.5	0.98	2.3	0.94	1.7	0.97	N/A	N/A	6
Low-level models	Concatenation	1.3	0.98	2.0	0.96	1.5	0.98	N/A	N/A	6
	SO-PLS	1.2	0.98	1.8	0.96	1.5	0.98	5	2	7
Mid-level models	PCA Scores	1.4	0.98	1.9	0.96	2.0	0.96	6	9	9
	PLS Scores	**0.8**	**0.99**	**1.0**	**0.99**	**1.3**	**0.98**	11	11	4
High-level model	PLS predicted Y	1.0	0.99	1.1	0.99	1.3	0.98	8	3	N/A

4 Conclusion

The results obtained demonstrated that ^{13}C NMR spectrum provides detailed and complementary information to improve property prediction. The data fusion of the two spectroscopic techniques employed in this study has potential use for fast and accurate properties prediction where errors of single models are higher than the reference method value. No data fusion model was found in the literature to predict diesel cetane number from spectroscopic information of the total effluent obtained from catalytic conversion processes.

5 References

[1] M. S. Rana, V. Sámano, J. Ancheyta, J. Diaz, Fuel 2007, 86, 1216–1231.

[2] H. Chung, Applied Spectroscopy Reviews 2007, 42, 251–285.

[3] John C Edwards (Ed.) A Review of Applications of NMR Spectroscopy in the Petroleum Industry, 2011.

[4] M. K. Moro, Á. C. Neto, V. Lacerda, W. Romão, L. S. Chinelatto, E. V. Castro, P. R. Filgueiras, Fuel 2020, 263, 116721.

[5] ASTM D 2892-20, Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column), 2020, ASTM International, West Conshohocken, PA.

[6] ASTM D613-01, Test Method for Cetane Number of Diesel Fuel Oil, 2001, ASTM International, West Conshohocken, PA.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Industry 4.0 enablers for pharmaceutical manufacturing

Y. Roggo[1] M. Jelsch[1] L. Pellegatti[1] S. Ensslin [1]M. Krumme[1]
[1] Novartis Pharma AG, Technical Research & Development / Continuous Manufacturing, Basel, Switzerland
yves.roggo@novartis.com

Keywords: Continuous Manufacturing, Solid Dosage Form, Process Monitoring, Process Analytical Technology, Process Data Science, Soft Sensors, Advanced automation, Industry 4.0

1 Introduction

Continuous Manufacturing (CM) of pharmaceutical drug products is a rather new approach within the pharmaceutical industry. CM process requires sophisticated process control strategies, to know at all times the current process state and to ensure consistent product quality at all times.

A continuous wet granulation line (Figure 1) was investigated for the production of solid dosage forms. Four NIRS (Near Infrared Spectroscopy) probes were installed on this CM line. NIRS is a popular qualitative and quantitative PAT-tool (process analytical technologies) in the pharmaceutical industry, as it is a safe, fast, and non-destructive method [1, 2].

The main objective of this paper is to demonstrate that PAT and Process Data Science support the digitalization and the automation of the CM line as enablers of the fourth industrial revolution. The three main aspects are the following: 1- PAT increase the process understanding and process characterization, 2- PAT-tools and soft-sensors deliver real-time information about the process state and product quality, 3- Advanced automation can be implemented based on PAT and soft sensor.

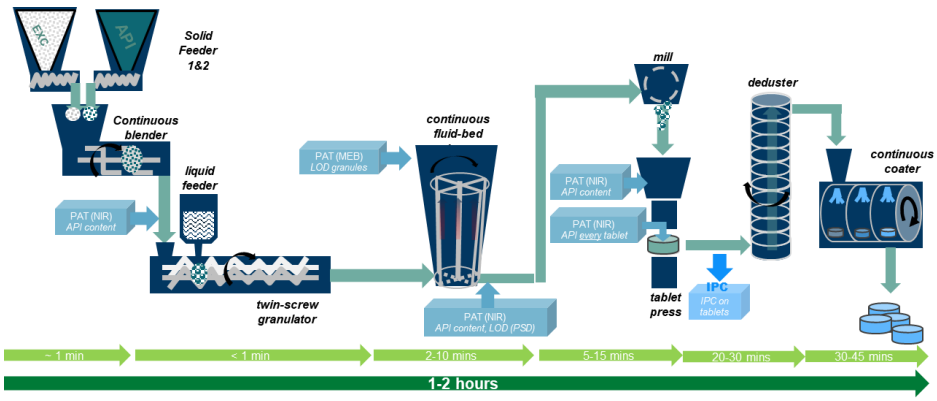


Figure 1 – CM Process and main PAT probes

2 Process Understanding

To characterize the process flow, the residence-time distribution can be determined by mathematical methods using the data provided by the four PAT probes. Example of RTD computation will be described. The understanding of process dynamics and of impacts of process disturbances allows the diversion in real time of out of specification product.

3 Real time process monitoring

The NIRS sensors can monitor the pharmaceutical line. Moreover, soft sensors based on process parameters can be developed in order to predict critical quality attributes. The example of neural network prediction of the LOD (Loss on Drying) will be discussed (Figure 2). Product out-of-specification can be diverted in real time.

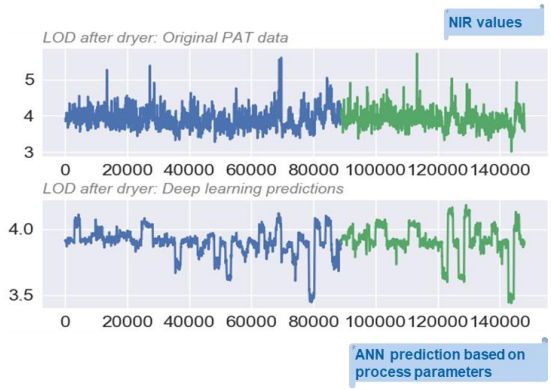


Figure 2 – PAT and Softsensor for LOD prediction

4 Advanced process automation

Model Predictive Control methods is an advanced method of process control with a set of constraints. MPC solves an optimization problem at each time step to find the optimal control actions within a defined prediction horizon. MPC applies these actions to the next time step while it resolves the optimization problem again to drive the predicted plant output to the desired reference. By this way, MPC has the ability to anticipate future events and can take control actions (Figure 3).

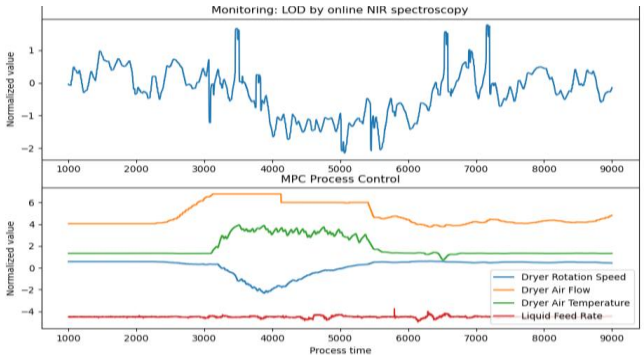


Figure 3 – Regulation of LOD by advanced automation (MPC)

5 Conclusion

Process Analytical Technology (especially NIRS), Soft sensors and advanced process automation (*e.g.* MPC) allows the digitalization and the control of the production line. Pharmaceutical continuous manufacturing is ready for the fourth industrial revolution.

6 References

[1] Roggo, Y., et al., *Deep learning for continuous manufacturing of pharmaceutical solid dosage form.* Eur. J. Pharm. Biopharm., **153**(1): p. 95-105, 2020.

[2] Roggo, Y., et al., *Continuous manufacturing process monitoring of pharmaceutical solid dosage form: A case study.* Journal of Pharmaceutical and Biomedical Analysis, **179**: doi.org/10.1016/j.jpba.2019.112971, 2020.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Bois imprégnés avec des produits de préservation commerciaux : Utilisation de stratégies de fusion de données pour améliorer leur discrimination

M. Rubini[1*] H. Issaoui[1] P. Dulucq[1] D. Talaga[2] J-L. Bruneel[2] A. Desmedt[2] B. Charrier[1]

[1] Xylomat (UMR5254), Université de Pau et des Pays de l'Adour, UPPA, E2S UPPA, CNRS, IPREM, Mont de Marsan, France, *morandise.rubini@univ-pau.fr

[2] ISM - Institut des Sciences Moléculaires, Université de Bordeaux I, 351 Cours de Libération, Talence, France

Keywords: Méthodes multiblocs, *Pinus Pinaster*, Préservation du bois, Recyclage

1 Introduction

Les produits de préservation du bois sont utilisés pour prolonger la durée de vie du bois utilisé dans des situations où il est susceptible d'être biodégradé. Ces produits sont composés d'une multitude de biocides (insecticides ou fongicides), et ces biocides peuvent être de diverses natures chimiques (organiques ou/et inorganiques). Cependant, à la fin du cycle de vie du bois traité, l'identification du produit de préservation est une étape cruciale pour pouvoir adopter la bonne stratégie de recyclage. Actuellement, il n'y a pas de méthodologies non destructives, fiables et rapides qui permettent de discriminer les bois traités, ce qui conduit à un faible recyclage de ces déchets [1]. Plusieurs études ont précédemment suggéré que la spectroscopie, notamment *Proche InfraRouge (PIR)*, a le potentiel de discriminer des bois traités avec différents produits de préservation [2].

Le but de ce travail est de démontrer que différentes spectroscopies, couplées à des stratégies de fusion de données, peuvent être utilisées pour (a) distinguer le bois traité du bois non traité ; (b) distinguer le produit de préservation utilisé sur le bois.

2 Matériels et méthodes

Les produits de préservation à base d'eau sont les plus utilisés dans le monde [3]. Ainsi, cinq produits de préservation commerciaux ont été sélectionnés de manière à avoir des molécules biocides de nature différentes (organiques ou/et inorganiques) [Voir Tableau 1]. Les imprégnations des échantillons de bois (*Pinus Pinaster*) ont été réalisées selon la norme NF EN 350 (2016) à 3 concentrations de produits de préservation (50, 70, 100 %). Au total 360 échantillons ont été préparés.

Chaque échantillon a été analysé par Fluorescence induite [(GSM, Université Bordeaux – CNRS), excitation à 532 *nm*, émission sélectionnée entre 534-891 *nm*], NIR [SCiO (Consumer Physics, Israel), 740-1070 *nm*], et MIR [FT/IR-4700 (Jasco, Italy), 600-4000].

Les blocs de données individuels ont été utilisés pour créer des modèles PLS-LDA, puis ils ont été analysés simultanément par deux ou par trois avec des modèles SO-PLS-LDA [4]. Les spectres ont été prétraités avec différents *preprocessing* dans le but d'éliminer la variabilité parasite éventuellement présente dans les données (principalement due à la diffusion de la lumière). Finalement, la combinaison deux *preprocessing* a été retenue : SNV + SG1 (dérivée 1ère, polynôme

du 2ème ordre et fenêtre de 5 % de points calculés sur le nombre de variables). Le nombre de variables latentes (LVs) optimal a été choisi en minimisant les erreurs de classification calculées à l'aide d'une cross-validation (3-fold). Tous les calculs ont été exécutés dans Matlab 2019a (The Mathworks Inc., Natick, MA) en utilisant des routines basées sur les toolboxes SAISIR et libPLS.

Tableau 1 – Liste et composition des 5 produits de préservation commerciaux utilisés dans cette étude

Nom	Nom commercial	Nom des molécules actives dans les produits
GRI	Wolsit KD-45 + Wolmanit ProColor grey 3501	Propiconazole, Perméthrine
MAR	Wolmanit CX-10 + Wolmanit Colorant marron 2001	Copper carbonate, bis-(N-Cyclohexyldiazeniumdioxy)-copper, Acide borique
VER	Wolmanit CX-10	
3V3	« 3V3 Poutres et charpentes »	Perméthrine
XYL	« Xylophène »	Cyperméthrine, Propiconazole, Tebuconazole, IodoPropynyl Butyl Carbamate (IPBC)

3 Résultats et discussion

La justesse globale (Overall Accuracy) montre que les différentes stratégies de fusion de données ont conduit à de meilleurs résultats [Voir Tableau 2].

Tableau 2 – Justesse globale (Overall Accuracy) des modèles PLS-LDA et SO-PLS-LDA

	PLS-LDA			SO-PLS-LDA			
	Fluo	NIR	MIR	Fluo-NIR	NIR-MIR	Fluo-MIR	Fluo-NIR-MIR
LVs	7	10	9	[10 ;11]	[9 ; 13]	[10 ; 14]	[5 ; 2 ; 6]
O. Acc.(CV)	0.4292	0.6250	0.8125	0.9958	0.9875	0.9750	1.0000
O. Acc. (Test)	0.3500	0.6167	0.7667	0.9583	0.9917	0.9667	0.9917

4 Conclusion

En comparaison avec les modèles PLS-LDA, la fusion des données améliorent considérablement la discrimination, suggérant que ces méthodologies seraient tout à fait adaptées pour discriminer le bois traité du bois non traité, et ainsi, déterminer le produit de préservation qui a été utilisé lors de l'imprégnation du bois (GRI, MAR, VER, 3V3, et XYL). En outre, d'autres analyses statistiques (Test de *Student*, VIP, test de Mc Nemar) ont été utilisées pour identifier les variables qui contribuent le plus à la réussite des discriminations, et confirmer les améliorations significatives des stratégies de fusion des données.

5 Références

[1] F. Berger, F. Gauvin, and H. J. H. Brouwers, "The recycling potential of wood waste into wood-wool/cement composite," *Constr. Build. Mater.*, vol. 260, p. 119786, Nov. 2020.

[2] C.-L. So, S. T. Lebow, L. H. Groom, and T. F. Shupe, "An Evaluation of the Use of Near Infrared (NIR) Spectroscopy to Identify Water and Oil-borne Preservatives," 2003.

[3] T. Koumbi-Mounanga, P. A. Cooper, N. Yan, K. Groves, T. Ung, and B. Leblon, "Prediction of boron content in wood pellet products by near-infrared spectroscopy," *For. Prod. J.*, vol. 66, no. 1–2, pp. 37–43, Apr. 2016.

[4] M. Cocchi, A. Biancolillo, and F. Marini, "Chemometric Methods for Classification and Feature Selection," in *Comprehensive Analytical Chemistry*, vol. 82, Elsevier B.V., 2018, pp. 265–299.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

QbD tools to inscribe PAT control measurement into the process validation lifecycle

Pierre Lebrun
Director Statistics, Princ. Consultant, Pharmalex Belgium, Louvain-la-Neuve

Keywords: QbD, PAT, Qualification

Abstract

In this talk, we'll introduce the concepts of process validation lifecycle and review different tools that can be used to improve process understanding. PAT tools for control of raw material attributes and intermediates should be defined in this risk-based framework. After ranking and identifying attributes and process parameters of critical relevance for the process, an additional decision is made to envisage the analysis of some of them through surrogate's analysis, using measurement systems such as near infrared or Raman spectroscopy, replacing offline measurements such as HPLC or water content, with real time, online measurements. At this process characterization stage, design of experiment is the definitive methodology to analyze if process parameters impact the process and if online measurements are able to make proper control, i.e., detecting when the process is starting to deviate and risks to produce out-of-specification products. Example of design of experiments will be shown, applied to small molecule process and bioreactor follow up.

References

P. Wahl et al., PAT for tableting: Inline monitoring of API and excipients via NIR spectroscopy, European Journal of Pharmaceutics and Biopharmaceutics 87 (2014), 271-278

P.F. Chavez et al., Optimization of a pharmaceutical tablet formulation based on a design space approach and using vibrational spectroscopy as PAT tool, International Journal of Pharmaceutics 486 (2015), 13–20

S. Mercier et al., Process analytical technology tools for perfusion cell culture, Engineering in Life Sciences 16 (2016), 25–35

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Quantitative structure-retention relationship modelling of small pharmaceutical compounds in reverse phase liquid chromatography

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

[1] University of Liege (ULiege), CIRM, Vibra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Avenue Hippocrate 15, 4000 Liege, Belgium

[2] University of Liege (ULiege), CIRM, MaS-Santé Hub, Laboratory for the Analysis of Medicines, Avenue Hippocrate 15, 4000 Liege, Belgium

[#] These authors have equally contributed to this work

Keywords: QSRR, Stepwise regression, Lasso, SVR, Xgboost, Random Forest

1 Introduction

Reverse phase liquid chromatography (RPLC) is still one of the most used analytical technique for the analysis of chemical mixtures. The development step can be very extensive given the different possible stationary phases, mobile phases and other analysis parameters. A thorough screening takes a lot of time and requires many consumables even with a systematic approach using experimental planification. The development of quantitative structure-retention relationship (QSRR) models can advantageously replace this experimental screening phase with *in silico* chromatograms simulations.

QSRR models are statistically derived relationships between chromatographic parameters and computed molecular descriptors characterizing the analytes. Several linear and nonlinear models have been used to build such models (Partial least squares (PLS), Bayesian, Ridge, Lasso, K-nearest neighbors (KNN), support vector machines (SVR), artificial neural network (ANN), etc.) [1]. Ensemble machine-learning models covering boosting, bagging and stacking have shown to generally outperform other algorithms [2].

In the presented work, QSRR models will be built for different chromatographic conditions (pH and gradient time). Subsequently, a response surface model (RSM) will be used allowing predictions of retention times in new conditions within the studied space [3, 4, 5].

2 Material and methods

Ninety-eight molecules were selected to cover a wide range in LogP values (-3.22 – 6.45), molecular weight (46 – 454 g/mol) and includes both non-charged and charged molecules (25 non-charged and 73 charged). Experimental retention times were acquired in house on three different HPLC systems (Waters Alliance) with gradients from 100% buffer to 5% buffer in 20 and 60 minutes considering five different pH levels (2.7, 3.5, 5, 6.5 and 8). These two gradients and five pH conditions represent the ten datasets that will be analyzed. Methanol was selected as the organic modifier.

At first, the weighted average of the molecular descriptors of each present form of the compound are calculated. Then, four machine-learning models were fitted on the ten datasets using the 26

features selected using stepwise regression methods. RMSE and R^2 values were used to compare the different models. Finally, a RSM is fitted for each compound based on the predicted retention times starting from equation (1) while removing pH terms for neutral.

$$\log(t_R) = \beta_0 + \beta_1 \times pH + \beta_2 \times t_{grad} + \beta_{12} \times pH \times t_{grad} + \beta_{11} \times pH^2 + \beta_{111} \times pH^3 \quad (1)$$

3 Results and discussion

Out of the tested models, XGBoost and Lasso were the best performing ones showing R^2 values as high as 0.99 for the training set and R^2 higher than 0.95 for the prediction set shown in Figure1. Their blended prediction performed better over single model predictions.

Using those predictions, RSM models were built. The different predictions of ibuprofen from the external test set can be seen on Figure2.

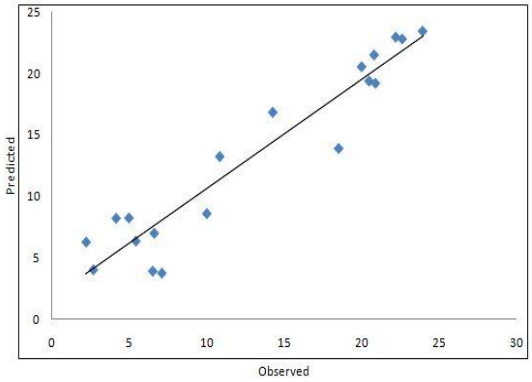


Figure 1: Plot of observed vs. predicted retention times for best prediction model (XGBoost)

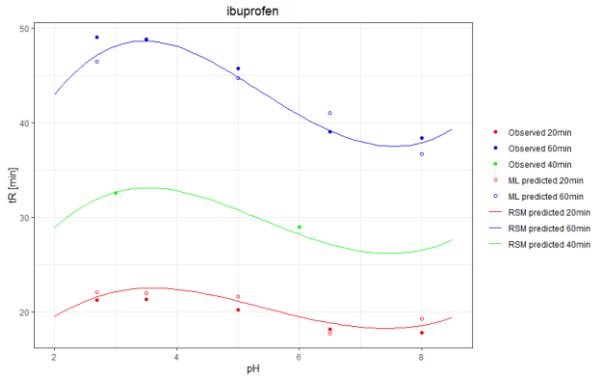


Figure 2: Observed (hollow points), ML predicted (filled points) and RSM predicted (curves) retention times of ibuprofen for 20, 40 and 60 minutes gradients

4 Conclusion

The RPLC retention times predicted by QSRR models followed by a RSM model were close to the experimental ones. This demonstrates that the combination of QSRR and RSM offers the possibility to replace usefully the experimental screening phase by computational methods when developing chromatographic techniques for known sets of molecules. The presented results concern a limited set of test molecules and will be further extended to new molecules and chromatographic modes.

5 Acknowledgments

This work was funded by the FWO / FNRS Belgium EOS grant 30897864 “Chemical Information Mining in a Complex World”.

6 References

[1] P. Haddad, M. Taraji, R. Szücs, Prediction of Analyte Retention Time in Liquid Chromatography, *Analytical Chemistry*, in press

[2] R. Bouwmeester, L.Martens, S. Degroeve, Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction, *Analytical Chemistry* 91 (5): 3694–3703, 2019

[3] M. Taraji, P. Haddad, R. Amos, M. Talebi, R. Szücs, J. Dolan,C. Pohl, Rapid Method Development in Hydrophilic Interaction Liquid Chromatography for Pharmaceutical Analysis Using a Combination of Quantitative Structure-Retention Relationships and Design of Experiments,*Analytical Chemistry* 89(3): 1870–78, 2017

[4] E. Tyteca, M. Talebi, R. Amos, S. Hyun Park, M.Taraji, Y. Wen, R.Szücs, C. Pohl, J. Dolan,P. Haddad Towards a Chromatographic Similarity Index to Establish Localized Quantitative Structure-Retention Models for Retention Prediction: Use of Retention Factor Ratio,*Journal of Chromatography A* 1486: 50–58, 2017.

[5] L. Kubik,P. Wiczling, Quantitative Structure-(Chromatographic) Retention Relationship Models for Dissociating Compounds,*Journal of Pharmaceutical and Biomedical Analysis* 127: 176–83, 2016.

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

From complex real-world data to process understanding and monitoring, a use case in the chemical industry

S. Preys[1] A. Zenner[2] F. Gaulier[3] M. Davezac[4]

[1] Ondalys, 4 rue Georges Besse, 34830 Clapiers, spreys@ondalys.fr

[2] Elkem Silicones, 55 avenue des Frères Perret, 69190 Saint-Fons, alexis.zenner@elkem.com

[3] Elkem Silicones, 55 avenue des Frères Perret, 69190 Saint-Fons, florine.gaulier@elkem.com

[4] Elkem Silicones, 55 avenue des Frères Perret, 69190 Saint-Fons, magali.davezac@elkem.com

Keywords: Process monitoring, fault diagnosis, MSPC, multivariate control charts.

1 Introduction

Process analytics using all kinds of data collected along the production line, from process parameters, univariate sensors, to more complex process analyzers, such as in-line NIR or Raman probes, is now of a great importance within Operational Excellence and Continuous Improvement in different process industries, for both continuous and batch processes. Within a pilot plant or production line and using the same initial database, several goals can be achieved by choosing the appropriate data analytics/chemometrics tools: process understanding, process optimization, real-time process monitoring and troubleshooting, forecasting and control, and even predictive maintenance.

This use case will show how complex real-world industrial data has been handled to get insight into a chemical process for understanding, monitoring and troubleshooting. The singularity of this study lies in the different data analysis challenges arising from the complex nature of the process studied.

2 Material and methods

Spectra were collected on production line during one year from an in-line Kaiser Raman probe RXN4 located between two steps of a complex process (reactor and treatment) for silicone polymer manufacturing. Two different quality parameters were measured on final product.

A well-known Multivariate Statistical Process Control (MSPC) approach was implemented ([1], [2] and [3]). However, different challenges had to be addressed in order to be able to deploy the methodology:

- Massive data due to a high frequency and high resolution measurement
- High level of noise due to starts and ends of production lots
- Weak correlation between the intermediate product measured and the averaged end-product quality, since spectra acquisition was done in the middle of the whole process course
- High process variability due to solvent recycling.

3 Results and discussion

Handling of massive data was resolved by tuning an optimal data compression rate. Iterative specific data cleaning for starts and ends of production lots was performed using different chemometrics diagnostic tools. Despite weak correlation with end-product quality, end-product information was used to refine NOC (Normal Operating Conditions) dataset. And high variability of the process has been modeled within the monitoring model.

Thus, specific data cleaning, preprocessing, and modeling strategy have been the key to achieve a new insight into the process and to build a real-time monitoring tool bringing added value for process understanding, monitoring, and chemical interpretation.

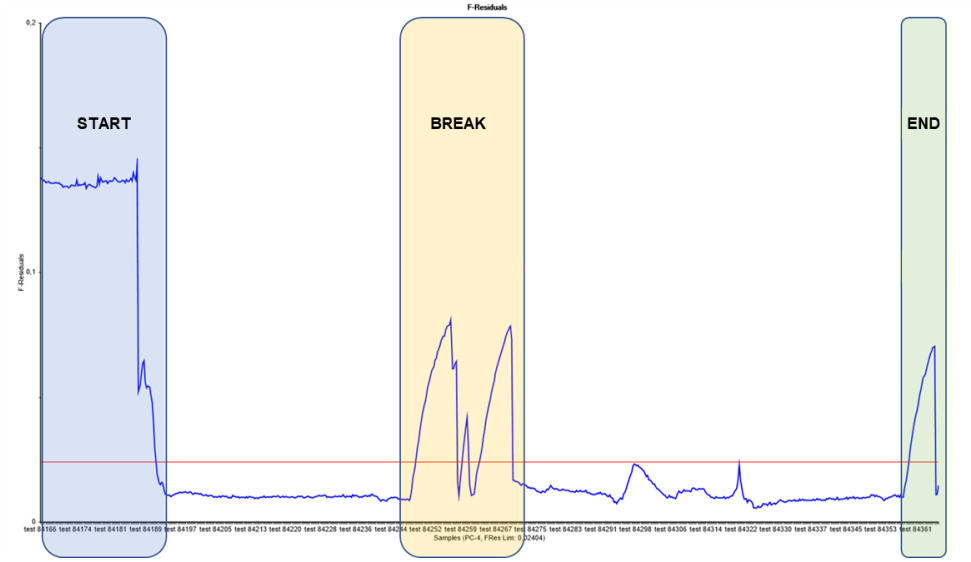


Figure 1 – Real-time monitoring tool (F-Residuals vs. time).

4 Conclusion

Multivariate data analytics has brought high value in this work for extracting useful information from on-line analyzer and for building a real-time monitoring tool of the process, despite the numerous challenges due to this complex process. Relevant dashboards useful for the operators were implemented for real-time monitoring and troubleshooting.

5 References

[1] C. Eisenhart, M. W. Hastay and W. A. Wallis : *Multivariate quality control, illustrated by air testing of sample bombsights*. A. Wallis (Eds), McGraw-Hill, New York, 1947.

[2] Tracy N. D., Young J. C. and Mason R. I. Multivariate control charts for individual observations, *Journal of Quality Technology* 24, 88-95, 1992.

[3] Kourti T. and MacGregor J. F. Tutorial: Process analysis, monitoring and diagnosis, using multivariate projection methods, *Chemometrics and Intelligent Laboratory Systems* 28, 3-21, 1995.

6 Acknowledgments

The authors would like to thank the Axel'One collaborative innovation platform, Lyon/France, for financial support.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Using prediction bands for near-infrared spectra for authentication and verification of drug products

T. H. Avohou[1] P.-Y. Sacré[1] Ph. Hubert[1] E. Ziemons[1]

[1] University of Liège (ULiege), CIRM, Vibra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Department of Pharmacy; Address: Avenue Hippocrate 15, 4000, Liège, Belgium

Keywords: Class-modelling, prediction bands, Bayesian chemometrics, drug product identification.

1 Introduction

Near-infrared spectroscopy (NIR) is a powerful analytical tool approved by the EU and US pharmacopeias. It can provide an accurate description of the physicochemical composition of samples, and hence can be used to fingerprint a drug product. With the fast-development and miniaturization of handheld spectrophotometers, this vibrational spectroscopy technique is more and more used in a large range of research and industrial applications involving characterization, identification and quality control of drug products.

These applications however require the use of accurate, robust, risk-oriented, computationally efficient decision-making tools to statistically compare high dimensional spectra to references in order to identify or control the quality of pharmaceutical products.

We propose a novel and probabilistic one-class classification strategy based on newly emerging chemometric techniques of (Bayesian) functional data analysis, for the identification and quality control of medicines. The strategy uses the concept of prediction bands as acceptance region.

2 Material and methods

A representative training set of spectra of a target product is sampled from several batches of that product using a MicroPhazir® (ThermoFisher Inc) reflection NIR spectrophotometer. Based on this set and using Bayesian (functional) principal component regression [1], a statistical prediction band is constructed so that it contains a high proportion, say at least 90% or 95% of future spectra of the product (see Figure 1 for illustration). The upper and lower limits of the band are used as critical trajectories or reference spectra that would enable testing the deviation from regular behavior or excursions out of the bands of any future unit from the product batch based on its spectrum, while controlling the risks of errors [2].

3 Results and discussion

The proposed one-class classification methodology has been applied to the identification of Dafalgan® 1g. Four other paracetamol-based drugs were used to evaluate the specificity. Spectra were measured with the handheld NIR device. The predicted trajectories of future Dafalgan® 1g spectra and their limiting behaviors (band limits) are illustrated on Figure 2A and B. The pattern of deviation from the band limits (acceptance region) are illustrated on Figure 2C and D for a

Dafalgan® 1g spectrum and an Excedryn® spectrum respectively. The method compares favorably with existing methods like the SIMCA, with high sensitivity between 90% and 98% and similar specificity.

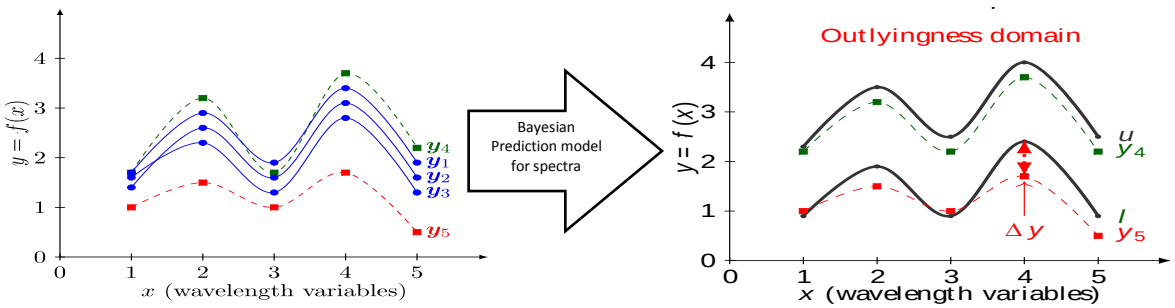


Figure 1 - Concept of prediction bands for one-class classification of NIR spectra for authentication and characterization of drug products. Notes: Blue solid curves ($\mathbf{y_1}$-$\mathbf{y_3}$) are segments of calibration spectra. Dotted curves are segments of test spectra, the green ($\mathbf{y_4}$) being a target class test spectrum and the red ($\mathbf{y_5}$) a non-target class test spectrum; curves ***u*** and ***l*** are respectively the upper and lower band limits. Δy is the deviation of $\mathbf{y_5}$ from the band at *x=4*

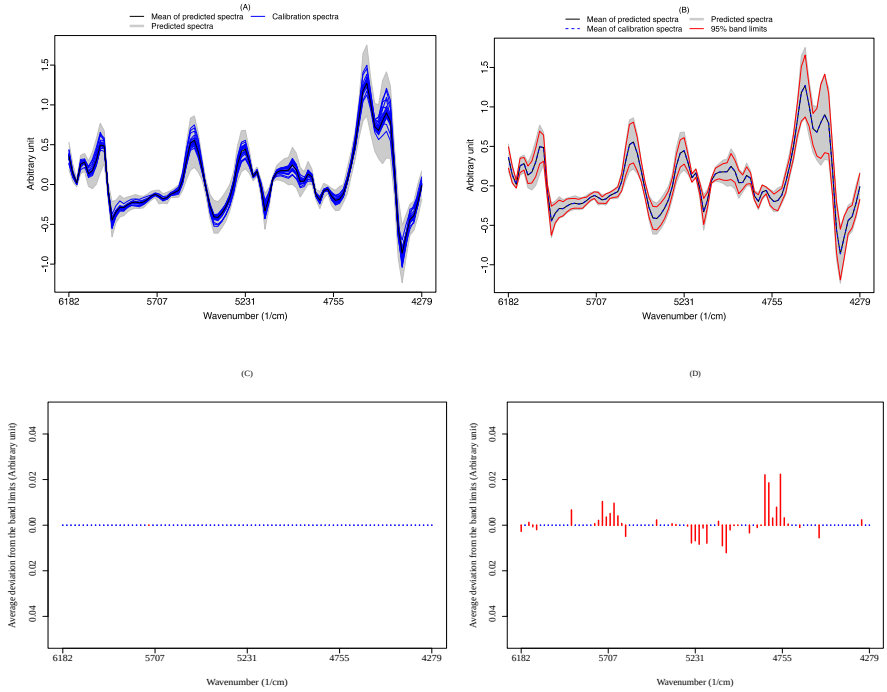


Figure 2 - Predicted NIR spectra of Dafalgan® 1g (A) and prediction band limits (B); deviations from the limits for a Dafalgan® 1g spectrum (C) and an Excedryn® spectrum (D).

4 Conclusion

A new one-class classification method for identification and quality control of drug products is proposed, using prediction bands for NIR spectra. Compared with existing spectral matching models, the proposed approach is fully predictive, with more intuitive interpretation of classification results.

5 References

[1] J.S. Morris, Functional regression. *Annual Review Statistics and its Applications* 2, 2015, pp. 321-359.

[2] TH Avohou, et al. A probabilistic class-modelling method based on prediction bands for functional spectral data: Methodological approach and application to near-infrared spectroscopy. *Analytica Chimica* Acta 1144, 2021, pp. 130-149.

e- CHIMIOMÉTRIE 2021
Une conférence 100 % en ligne et gratuite
2-3 Février 2021
SFdS GFC

Pseudo-univariate calibration by UV spectroscopy in the determination of resveratrol in grape juice

L. Valderrama[1], E. Carasek[1], A. Coqueiro[2], P.H. Março[3], P. Valderrama[3]
[1] Universidade Federal de Santa Catarina (UFSC) – Florianópolis – SC - Brazil, leovalderrama6@gmail.com
[2] Universidade Tecnológica Federal do Paraná (UTFPR) – Ponta Grossa – PR - Brazil, alinedqi@gmail.com
[3] Universidade Tecnológica Federal do Paraná (UTFPR) – Campo Mourão – PR - Brazil, pativalderrama@gmail.com

Keywords: MCR-ALS, HPLC validation, paired t test.

1 Introduction

Humanity has been increasingly concerned with the food it consumes. This has driven research on bioactive compounds, and new analytical methodologies are needed to ensure the quality of food and the amount of bioactive compounds present. Resveratrol is a non-flavonoid compound from polyphenol families [1], which in the grape is synthesized in the skin in response to the stress caused by the attack of fungi, and mechanical damage [2], for example. Due to the health benefits associated with this bioactive compound [3] and considering that the most analytical methods to determine resveratrol in juices are time-consuming, expensive, and generating waste [2,3], the aim of this work was to propose a method based on ultraviolet (UV) spectroscopy and multivariate curve resolution with alternating least squares (MCR-ALS) in a pseudo-univariate calibration model.

2 Theory

The mathematical steps of the MCR-ALS are described in references [4]. Some important advantages can be highlighted in the quantification through the use of curve resolution methods over the conventional multivariate calibration by partial least squares (PLS): a) with curve resolution methods it is not necessary to know or include interferences in the calibration step that can allow reaching the second-order advantage even for first-order data; b) a reduced amount of calibration samples is required when compared to multivariate calibration by PLS. This is possible because the regression is performed using the extracted relative concentration profile (related to the analyte of interest), recovered by curve resolution, against the known concentration values. In addition, its models are mathematically simpler, being considered pseudo-univariate due to their use being similar to univariate but, in actually related to multivariate data [5].

3 Material and methods

Thirteen standard resveratrol samples (UP water with 0.1% of HCl) covering the concentration range of 0.1 to 1.5 mg L^{-1} were prepared. From these standards, UV spectra were recovered (Ocean Optics equipment, 1 cm quartz cuvette, step 1 nm, 32 scans), and high performing liquid-chromatography (HPLC) analyzes made based on the reference [2]. For juice analysis, 10 μL of juice was diluted to 10 mL (UP water with 0.1% of HCl).

4 Results and discussion

The MCR-ALS results (Figure 1, 4 factors, non-negativity for concentration and spectra) show the recovered spectra for resveratrol. Its relative concentration was used to build a pseudo-univariate calibration curve with a correlation coefficient of 0.9938. Four standard samples were used in the validation steps, and a paired t-test shows no difference between the UV/MCR-ALS and HPLC methods with a 95% confidence level. Then, the UV/MCR-ALS model was used to quantify resveratrol in grape juices, and the concentration achieved varied according to those obtained previously for grape juices produced in Brazil [2].

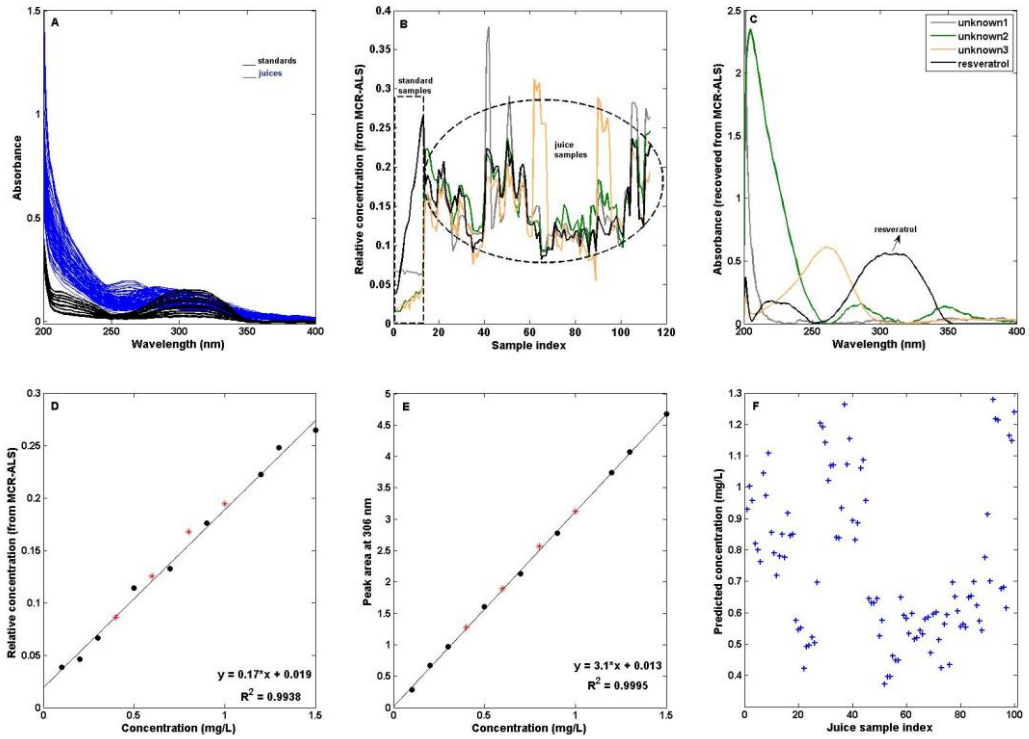


Figure 1 – Results. (A) UV spectra. (B) MCR-ALS relative concentration. (C) MCR-ALS recovered spectra. (D) Pseudo-univariate calibration curve. (E) HPLC curve. (F) Resveratrol prediction in grape juice.

5 Conclusion

A pseudo-univariate calibration model based on UV spectroscopy and MCR-ALS is suitable for quantifying resveratrol in grape juices.

6 References

[1] Airado-Rodríguez, D., Durán-Merá, I., Galeano-Díaz, T. & Wold, J. P. Front-face fluorescence spectroscopy: A new tool for control in the wine industry. *J. Food Comp. Anal.* 24, 257-264, 2011.

[2] Sautter, C. K., Denardin, S., Alves, A. O., Mallmann, C. A., Penna, N. G., Hecktheuer, L. H. Determination of resveratrol in grape juice produced in Brazil. *Ciênc. Tecnol. Aliment.* 25, 437-442, 2005.

[3] Dani, C., Oliboni, L. S., Vanderlinde, R., Bonatto, D., Salvador, M., Henriques, J. A. P. Phenolic content and antioxidant activities of white and purple juices manufactured with organically- or conventionally-produced grapes. *Food Chem. Toxicol.* 45, 2574-2580, 2007.

[4] Março, P.H., Valderrama, P., Alexandrino, G. L., Poppi, R. J., Tauler, R. Multivariate curve resolution with alternating least squares: Description, operation and aplications. *Quim. Nova* 37, 1525-1532, 2014.

[5] Ribeiro, G. M., Madivadua, D. A., Curti, S. M. M., Pantean, L. P., Março, P. H., Valderrama, P. Pseudo-univariate calibration based on independent component analysis for determination of the carbendazim concentration in orange juice. *Microchem. J.* 134, 114-118, 2017.