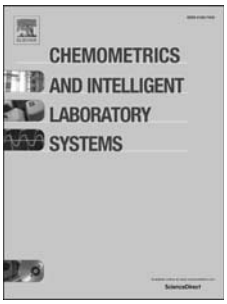
MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing

Puneet Mishra, Jean Michel Roger, Douglas Rutledge, Alessandra Biancolillo, Federico Marini, Alison Nordon, Delphine Jouan-Rimbaud Bouveresse



MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing

Puneet Mishra, Jean Michel Roger, Douglas N. Rutledge, Alessandra Biancolillo, Federico Marini, Alison Nordon, Delphine Jouan-Rimbaud-Bouveresse

PII: S0169-7439(20)30395-6

DOI: https://doi.org/10.1016/j.chemolab.2020.104139

Reference: CHEMOM 104139

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 2 July 2020

Revised Date: 12 August 2020

Accepted Date: 16 August 2020

Please cite this article as: P. Mishra, J.M. Roger, D.N. Rutledge, A. Biancolillo, F. Marini, A. Nordon, D. Jouan-Rimbaud-Bouveresse, MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing, *Chemometrics and Intelligent Laboratory Systems* (2020), doi: https://doi.org/10.1016/j.chemolab.2020.104139.

Table of Contents

MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing

Puneet Mishra[1,2], Jean Michel Roger[3,4], Douglas N. Rutledge[5,6], Alessandra Biancolillo[7], Federico Marini[8], Alison Nordon[2], Delphine Jouan-Rimbaud-Bouveresse[9]

[1]*Food and Biobased Research, Wageningen University and Research, Bornse Weilanden 9, 6708 WG, Wageningen, The Netherlands.*

[2] *WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, Glasgow, G1 1XL, United Kingdom*

[3]*ITAP, IRSTEA, Montpellier SupAgro, University Montpellier, Montpellier, France*

[4]*ChemHouse Research Group, Montpellier, France*

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

[6]*National Wine and Grape Industry Centre, Charles Stuart University, Wagga Wagga, Australia*

[7] *Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio 67100, Coppito, L'Aquila, Italy*

[8] *Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, 00185, Rome, Italy*

[9] *UMR PNCA. AgroParisTech, INRA. Université Paris-Saclay, 75005 Paris, France*

Corresponding author: *puneet.mishra@wur.nl;* jean-michel.roger@inrae.fr

Abstract

In recent years, due to advances in sensor technology, multi-modal measurement of process and products properties has become easier. However, multi-modal measurements are only of use if the data from adding new sensors is worthwhile, especially in the case of industrial applications where financial justification is needed for new sensor purchase and integration, and if the multi-modal data generated can be properly utilised. Several multi-block methods have been developed to do this; however, their use is largely limited to chemometricians, and non-experts have little experience with such methods. To deal with this, we present the first version of a MATLAB-based graphical user interface (GUI) for multi-block data analysis (MBA), capable of performing data visualisation, regression, classification and variable selection for up to 4 different sensors. The MBA-GUI can also be used to implement a recent technique called sequential pre-processing through orthogonalization (SPORT). Data sets are supplied to demonstrate how to use the MBA-GUI. In summary, the developed GUI makes the implementation of multi-block data analysis easier, so that it could be used also by practitioners with no programming skills or unfamiliar with the MATLAB environment. The fully functional GUI can be downloaded from (https://github.com/puneetmishra2/Multi-block.git) and can be either installed to run in the MATLAB environment or as a standalone executable program. The GUI can also be used for analysis of a single block of data (standard chemometrics).

Keywords: data fusion; multi-sensor; chemometrics; graphical user interface

Introduction

Sensing technologies play a major role in chemical industries where they are implemented to monitor and optimise process and product properties [1]. Sensing technologies do this by rapid estimation of key critical quality attributes of the process and products. However,

sometimes the products or processes are so complex that a single technique fails to obtain sufficient information about the samples. One such case, in the framework of process monitoring applications, is the use of Raman and mid-infrared (MIR) spectroscopy. Both techniques are complementary to one another as they are sensitive to different vibration modes corresponding to molecular groups. Furthermore, they complement one another's drawbacks as Raman signal may get influenced by fluorescence, but it will work with high moisture samples while MIR spectroscopy will be affected by the presence of moisture, but it will work well with fluorescent samples. In such a case, data from both techniques can be utilised in a complementary way: a combination of MIR and Raman spectroscopies could yield better results as each one will compensate for the drawbacks of the other.

Innovations in measurement technologies such as combining multi-spectral techniques for non-destructive estimation of process and product properties is now a major research domain requiring multi-block data analysis techniques. A possible application could be the integration of several different PATs such as MIRS, Raman, near-infrared spectroscopy (NIRS) and fluorescence spectroscopy (FS) into a single measurement probe (Fig. 1A). In the case of a process monitoring application, such a combined probe could be inserted into the process vessel with signals being obtained from multiple sensors, thus enabling the recording of continuous data from multiple techniques (Fig. 1B). However, this will require not only advances in hardware but also in chemometric procedures, such as variable selection methodologies for identification of key spectral regions, and data fusion algorithms for the combination of data coming from multiple techniques.

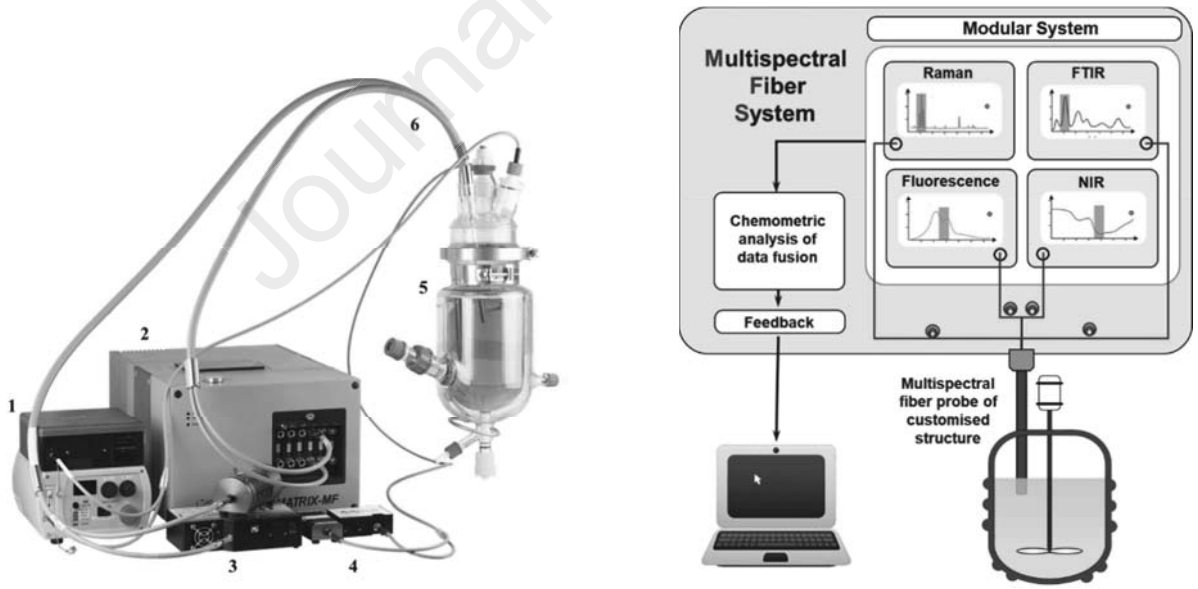
In some cases, it could be preferable for the data analysis to be performed using a sequential approach so as to highlight the added value coming from including new measurement modes, especially for industrial applications where financial justification is needed for new sensor purchase and integration. Data fusion is a feasible approach to combine all the information

from multiple sensors [2-7]. Data fusion can be dealt with in different ways, depending on the scientific domain. In the machine learning domain, data fusion can be performed at three levels, i.e. low, mid and high-level [8]. Low-level data fusion implies taking all the data together and performing an analysis in a similar way as for data from a single sensor. Mid-level fusion involves some preliminary data refining step where interesting features are extracted from each dataset and subsequently fused together. High-level data fusion is the fusion of the conclusions drawn from the output of the models created on each dataset, using decision rules such as majority voting or averaging. The main aim of data fusion in the machine learning domain is to improve the model accuracy with less attention being paid to the background process. However, in the chemometrics domain, data fusion, or multi-block data analysis, aims not only to improve the model accuracy but also to have a better understanding of the underlying characteristics involved. The aim is to identify the common and the specific hidden factors between and within multiple blocks of measurements, and to subsequently use them to build explainable and explanatory models [9-14].

There are two main tasks that need to be accomplished in multi-block data analysis, i.e., to enhance the data visualisation and improve the predictive performance of models. For data visualisation, a summary and comparison of methods can be found in [9, 11, 12]. Recently, a new data visualisation approach was presented for exploration of designed experiments in a multi-block scenario [10]. For multi-block predictive modelling, partial least squares regression-based methods are summarised in [15]. Multi-block methods have also been extended to incorporate variable selection [16, 17], which is important, for example, when different spectral sensors are used to study the same set of objects. In such a case, variable selection looks for subsets of variables that are important in each of the spectral techniques, which can then be useful to have a better understanding of the process, and to orient the development of cheap multi-spectral sensors. Multi-block methods are also emerging to

perform fusion of data of very different types, such as when fusing a 3 way data array (3D tensor) with a 2-way matrix (2D tensor) [18, 19].

In the present work, the first version of a MATLAB-based graphical user interface for multi-block data analysis (MBA-GUI) is presented. The MBA-GUI can perform data visualisation, sequential regression and variable selection on up to 4 different data sources. However, the algorithms implemented in the GUI are not limited to 4 data sources and can be used for any number of data sources. The MBA-GUI can also be used to implement a technique called sequential pre-processing through orthogonalization (SPORT). The MBA-GUI can also be used for a single block scenario. In addition, cases are presented showing how to use the MBA-GUI for data visualisation, regression, classification, variable selection and SPORT in the multi-block scenario. This is the first version of the MBA-GUI and, as multi-block analysis methods progress, the toolbox will be updated to incorporate new algorithms.



(A) **(B)**

Fig. 1: Scheme of multispectral fibre system (figure courtesy of Art Photonics GmbH, Germany). (A) Raman system (1); FTIR absorption System (2); NIR reflection System (3);

fluorescent System (4); chemical Reactor (5); fibre optic probes (6), and (B) A schematic of the multi-block data generated in a four blocks scenario.

Similar works

When looking at the available resources to perform multi-block data analysis, three main toolboxes can actually be found. multi-blockThe first one is the multi-block toolbox by the University of Copenhagen, Denmark (http://www.models.life.ku.dk/~courses/MBtoolbox/mbtmain.htm), which, having been last revised in 2001, focuses only on the two data fusion approaches which were most popular in that period, i.e, multi-block principal component analysis and multi-block partial least squares regression. The second one is the 'multi-block regression by parallel and sequential partial least-squares regression' toolbox by NOFIMA [20]. Both these toolboxes provide command line functionalities (which may be difficult to implement for people unfamiliar with the Matlab environment) and, anyway, consist of a limited number of tools. There is also a basic GUI available for performing multi-block component analysis in the domain of Behaviour research [21]. However, the GUI can only perform multi-block component analysis for data visualisation. Therefore, there exists a need for a GUI which has updated tools and a complete set of functionalities to perform multi-block data analysis.

Software description

The MBA-GUI was built utilising the application builder in MATLAB version 2018b (Natick, MA, USA). The application can be downloaded and installed in MATLAB (preferred version 2018b or higher) or can be used as a stand-alone executable or can be run through the '.mlapp' files in MATLAB command line. If user does not have MATLAB version 2018b or greater then it is recommended to install free MATLAB runtime tool and run the app as standalone. All the executable and MATLAB function can be downloaded

from (https://github.com/puneetmishra2/Multi-block.git). In the GitHub repository, the standalone toolbox executable files can be downloaded as *'Multi-block_toolbox.zip'* and the function for running the tools in command line as '*Toolbox.zip'*. The dataset demonstrated in this article can be downloaded as '*Dataset.zip'* from the same GitHub repository. All three files are available in the link (10/June/2020). To run the toolbox from command line user should use the toolbox folder as the current folder and type T1 on the command line which will start the main GUI interface. The user should put the password: 'welovedata' without comma and click run. Then user can load data and run the analysis. See also supplementary file to have a visual understating on how to download and setup the GUI. The GUI supports data format of .csv, .xlsx and .mat. A summary of the functionalities is presented in Fig. 2. In summary, the toolbox has options for loading data, three levels of pre-processing, i.e., smoothing, scatter correction and normalisation, and derivative estimation, as well as multi-block data visualisation, regression, classification, variable selection and SPORT. Multi-block variable selection methods are available for both regression and classification cases. Two main types of regression and classification are available, i.e., sequential orthogonalization [15], and common components and specific weights analysis (CCSWA), aka common dimensions (ComDim) [22].

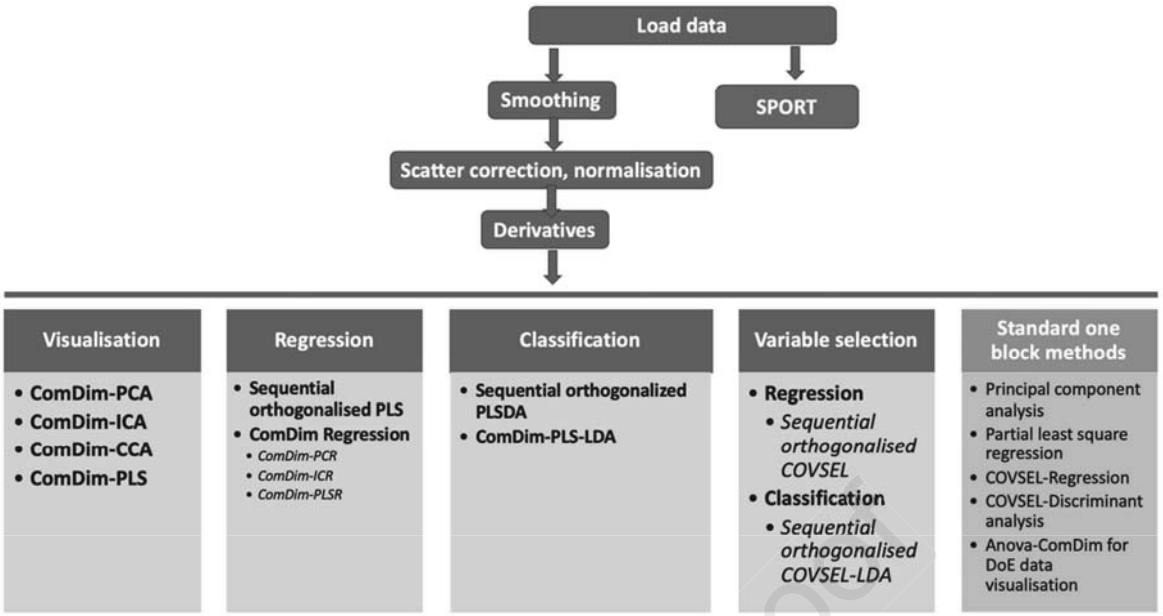


Fig. 2: Schematic of the tasks that can be performed with the MBA-GUI.

Software architecture and brief mathematical background of techniques available

Pre-processing

Data pre-processing is an important step to clean and homogenise the data prior to analysis. Proper data pre-processing can improve data modelling dramatically. There can be multiple steps in data pre-processing such as smoothing, scatter correction, normalisation and many others [23-25]. In the toolbox, we have provided a collection of common pre-processing methods.

Smoothing operations

Data smoothing reduces high-frequency noise from datasets prior to modelling. In the toolbox, several techniques are provided for performing smoothing in the variable domain. Three window-based smoothing techniques, i.e., Savitzky-Golay (SAVGOL), moving average and moving polynomial are provided. Further, two data decomposition and reconstruction techniques, i.e., principal components reconstruction and independent

components reconstruction are provided. In this toolbox, it is up to the user to choose a technique and decide, based on the model performance, which is best for their data. Pre-processing can also be explored automatically with the SPORT approach. All the spectral smoothing techniques are implemented using the codes explained in [23].

Scatter correction, Baseline correction and Normalisation

Multivariate data, and especially spectral data, suffer from a range of physical and chemical factors leading to baseline, additive and multiplicative effects. Prior to data modelling, it is always recommended to perform correction and normalisation of the data. However, depending on the data, the correction or normalisation method may be different. In the toolbox, the user may select several scatter and spectral normalisation techniques, including detrending [26], offset correction, multiplicative scatter correction [27], spline correction (where spline fitting is used to approximate the baseline which is then subtracted from the signal), asymmetric least-squares (ALS) correction [28], standard normal variate (SNV) [26], variable sorting for normalisation (VSN) [29], probabilistic quotient normalisation (PQN) [30], robust normal variate (RNV) [31], logarithm transform, autoscaling, 1st derivative (Savitzky-Golay), 2nd derivative (Savitzky-Golay), min-max, norm, range and max correction. All the correction and normalisation methods were implemented using the codes presented in [23]. It is recommended to perform the smoothing step, if required, before baseline correction and normalisation as these techniques can be affected by high-frequency noise. Detailed understanding of pre-processing methods in chemometrics can be found in [23-25].

Derivatives

Many normalisation and baseline correction techniques can remove effects like baseline shift, additive and multiplicative scatter. Derivatives, however, are also able to reveal underlying peaks. Therefore, the user can choose to perform 1st and 2nd derivatives or define the order of

the derivative as a 3rd pre-processing step. The algorithm for this operation is based on the codes provided in [23].

Pattern recognition and data visualisation tool

Data visualisation is the primary step prior to data analysis to gain an insight into the data [32]. Many factors, such as the presence of outliers, any groupings of samples or any strange patterns, can be detected by visualising the data. Based on the type of data, its visualisation can be very simple or very complex. If the data comprises only a few variables then major patterns can be observed via 1D, 2D and 3D plots. However, if the data consists of many variables, then dimension reduction techniques are used to transform the high-dimensional data into a space defined by interesting properties such as the variance. Once projected into a subspace (dimensionality reduction), the samples can be visualised via 1D, 2D and 3D plots together with plots of the transformed variables. Since spectral data usually contain many variables, ranging from hundreds to thousands depending on the spectral resolution, data transformation is almost always required [33]. In the toolbox, we implement a range of data transformation tools to enhance data visualisation task in such cases. The techniques are mainly based on capturing the major sources of variance in the sample domain. The multivariate data transformation techniques are specific to the multi-block scenario as they can highlight the contributions from the different blocks [9].

ComDim- based methods

ComDim belongs to the family of multi-block methods, which aim to extract the global components that highlight the important dimensions as well as the local (block specific) components [22]. Originally, the ComDim method was developed with the name common components and specific weights analysis (CCSWA) [34]. ComDim extracts the global and local components from multiple blocks of data in a sequential way. Each data block has a specific contribution to each common component which is called its '*Salience*'. ComDim

starts by normalising each block, $\mathbf{X_i}$, by its Frobenius' norm so that they all have the same total variance. Details regarding the algorithm can be found in [22, 35]. Application of the original ComDim (which, for reasons which will become apparent later, we will call ComDim-PCA here) in the case of 2-blocks is presented in Fig. 3. The common components (CCs) are extracted sequentially in an iterative fashion, by extracting the eigenvector associated with the largest eigenvalue from a matrix **W**, which is the created by concatenating the weighted individual matrices, $\mathbf{X_i}$. The weightings (the *saliences*) are initially all set to 1, but they are recalculated during the iterations to reflect the contribution of each block to the dispersion of the individuals along that CC. After one CC is extracted, the $\mathbf{X_i}$ matrices are deflated, and the procedure is repeated to obtain the next CC.

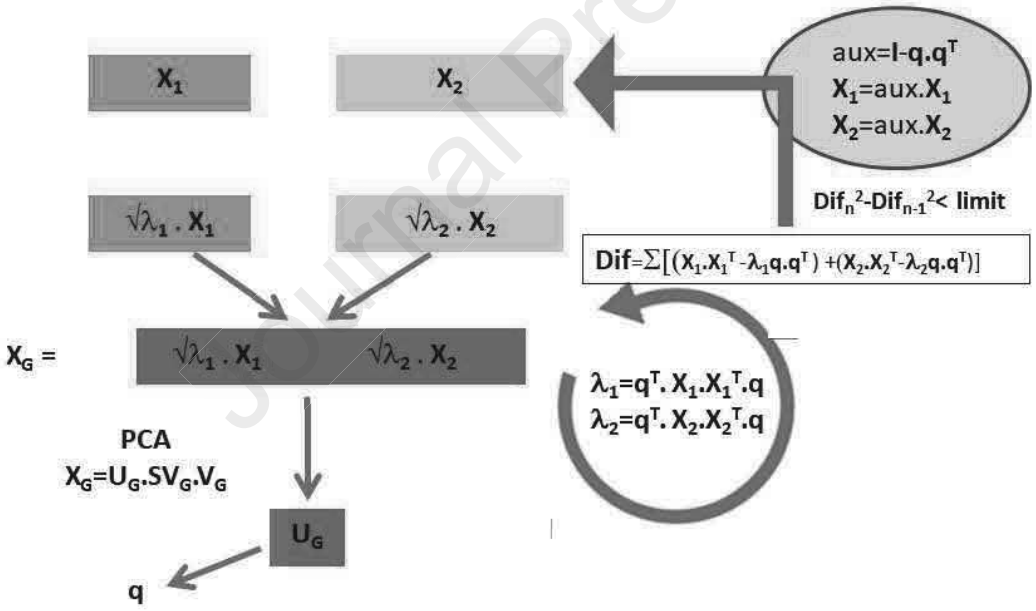


Fig. 3: The ComDim algorithm applied in the 2-block case ([36]): In the first step of the algorithm, X_1 and X_2 have been normalised.

In the MBA-GUI, ComDim is provided in several variants, i.e., common dimensions-principal components analysis (ComDim-PCA), common dimensions-independent

components analysis (ComDim-ICA), common dimensions-common components analysis (ComDim-CCA) (unsupervised decomposition methods), common dimensions-partial least squares-independent components analysis (ComDim-PLS-ICA) (a semi-oriented decomposition) and common dimensions-partial least squares (ComDim-PLS) (an oriented decomposition). The difference between the first three ComDim variants is the way the concatenated matrix of blocks is decomposed, i.e., using singular value decomposition (SVD), independent components analysis (ICA) or common components analysis (CCA). ComDim-PLS-ICA is a slightly more complex version of ComDim-ICA as it replaces the PCA decomposition of **W** by an ICA algorithm where the initial estimates of the independent components have been determined using a partial least squares (PLS) regression. In supervised ComDim-PLS, the PCA decomposition within ComDim is simply replaced by a PLS regression. The iterative procedure is identical in all cases, but the resulting CCs differ somewhat as a function of the criteria that determine the different decompositions of **W**.

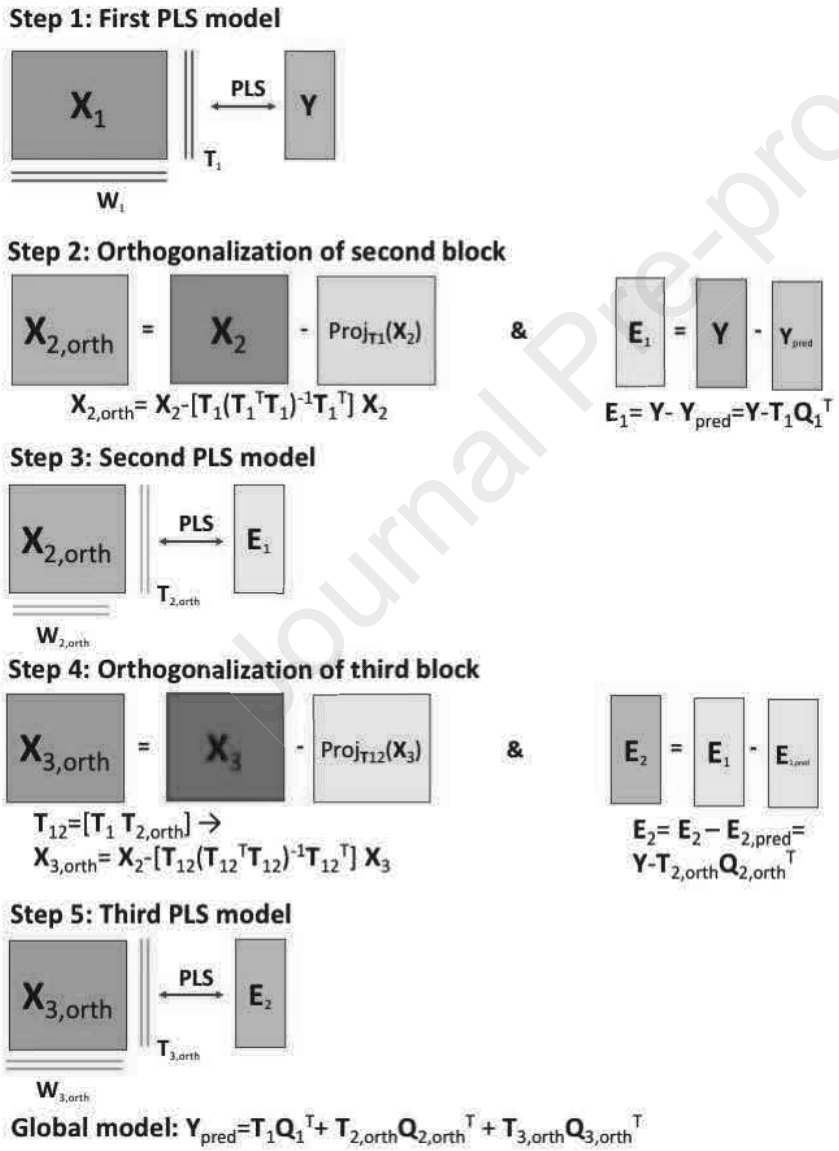
Regression

SO-PLS Regression

Sequential and orthogonalized PLS (SO-PLS) regression belongs to the family of multi-block PLS methods; the centrepiece of the method is the orthogonalization step, which ensures the removal of redundancies among modelled data blocks [15]. In SO-PLS regression, the extraction of information is sequential, meaning that the aim is to incorporate blocks of data one at a time and to assess the incremental contribution. A PLS regression is calculated between the first block $\mathbf{X}_1$ and **Y**, yielding scores $\mathbf{T}_1$. Then, all the remaining blocks $\mathbf{X}_2$, ... $\mathbf{X}_k$ and the **Y** block are orthogonalized with regards to $\mathbf{T}_1$. Then, the process is repeated on the second block, and so on for all the blocks. A scheme showing sequential extraction of information by SO-PLS regression is presented in Fig. 4. The major advantages of SO-PLS

311 are linked to orthogonalization, which removes redundant information, and to its sequential
312 nature, which allows the interpretation of the incremental contributions provided by each data
313 block. For more details, the reader is directed to [15, 37, 38]. The SO-PLS regression
314 function integrated into the toolbox is the freely available one at
315 https://www.chem.uniroma1.it/romechemometrics/research/algorithms/

316



317

318 *Fig. 4:A scheme presenting the sequential orthogonalized partial least squares (SO-PLS)*
319 *regression method.*

ComDim regression

The aim of ComDim regression methods is to sequentially extract the global component from multi-block data and to later use the scores to perform the regression on a **y** vector. To perform multi-block regression, four ComDim variants are integrated, i.e., ComDim-Principal component regression (ComDim-PCR), ComDim-Independent component regression (ComDim-ICR), ComDim-Partial least squares regression (ComDim-PLSR) and ComDim-Partial least squares – Independent component regression (ComDim-PLS-ICR). The difference between these regression methods is simply that the global components that are used were obtained using the different multi-block ComDim methods presented above. ComDim-PCR and ComDim-ICR use global scores that were extracted in an unsupervised way to perform multi-linear regression (MLR). ComDim-PLSR and ComDim-PLS-ICR use scores that were extracted in a supervised way by the PLS step within ComDim or within the ICA which is nested inside ComDim. To use new data, the ComDim models developed on the calibration dataset are used to calculate the scores for the new dataset and these are then introduced into the trained MLR model for the prediction. MLR regression on the ComDim scores is performed using in-house codes freely downloadable at https://www.chem.uniroma1.it/romechemometrics/research/algorithms/.

Classification

SO-PLS-LDA

Sequential orthogonalized partial least squares linear discriminant analysis (SO-PLS-LDA) is the natural extension of SO-PLS to the classification field [39]. To create a SO-PLS-LDA model, a SO-PLS model is first created using a dummy class matrix as **Y**, and then LDA can be applied either to the predicted **Y** or to the concatenated scores. Due to it being closely related to the SO-PLS algorithm, the steps of the SO-PLS-LDA approach are the same as those already sketched in Fig. 4; the main differences being that the response matrix **Y** should

be coded to account for class membership and that, as a further step, linear discriminant analysis is applied to either the predicted response or the concatenated scores block. The function integrated into the GUI is the one freely downloadable at https://www.chem.uniroma1.it/romechemometrics/research/algorithms/.

ComDim - based linear discriminant analysis

The aim of ComDim based classification methods is to sequentially extract global components form multi-block data and then to use the scores to perform a linear discriminant analysis (LDA). In the GUI, two variants of ComDim-LDA are included. ComDim-PLS2-LDA simply uses a PLS2 inside ComDim to orient the decomposition of the **W** matrix. On the other hand, ComDim-PLS2-ICA-LDA uses PLS2 to orient the extraction of the independent components from the **W** matrix by ICA, which has replaced PCA nested inside ComDim. In both cases, the scores can be directly used for LDA. Test set prediction is performed by first transforming the new data to the same space using the ComDim calibration model and then inputting into the trained LDA model.

Variable selection

SO-CovSel

SO-CovSel is a multi-block variable selection technique recently developed by [16]. The SO-CovSel is an extension of the CovSel technique to the multi-block scenario [40]. CovSel extracts the '*k*' variables from a matrix **X** that are most correlated to **Y** and independent to each other. SO-CovSel performs CovSel in a sequential orthogonalized way. It works as SO-PLS, replacing the PLS-scores by the COVSEL selected variables. CovSel in such an approach performs variable selection so that extraction of the variables from the consecutive block improves the model. SO-CovSel is designed for both regression as well as classification cases. SO-CovSel for regression and SO-CovSel-LDA are integrated into the

MBA-GUI utilising in-house codes freely downloadable at https://www.chem.uniroma1.it/romechemometrics/research/algorithms/.

SPORT

Novel use of multi-block methods can also be understood as boosting of different pre-processing techniques. Such a methodology called SPORT was recently developed by [41], where the sequential orthogonalization approach was used to fuse the information from different pre-processing techniques. Thus, instead of choosing between pre-treatments, SPORT allows us to make optimal use of the advantages of all pre-treatments. In this GUI, up to four different pre-processing techniques can be used for boosting. Boosting with sequential orthogonalization is a recent approach and has proven to be of high value to improve the prediction accuracies of the models. To perform SPORT, the same data can be loaded in four blocks and then different pre-processing can be applied to them. Further, based on the case of regression/classification, SO-PLS/SO-PLS-DA is used. More details on SPORT can be found in [41].

Standard one block chemometric analysis

Apart from multi-block analysis, the MBA-GUI provides an option to perform standard chemometric analyses. For a single block, the MBA-GUI has options to do PCA, PLS as well as CovSel variable selection for regression and discriminant analysis. The one-block analysis will automatically start when the MBA-GUI detects that only one block of data is loaded.

Datasets for MBA-GUI demonstration

Use of the MBA-GUI is demonstrated with three datasets. All the three datasets can be accessed in the same GitHub repository. The first dataset for data visualisation consists of the FTIR spectra of olive oils of 4 different origins (4 classes) measured in transmission mode [42, 43]. The second dataset to demonstrate the regression and variable selection task relates

to dry matter prediction in olive fruits measured with a portable spectrometer in diffuse reflectance. The reference dry matter was measured by the hot air oven method by noting the fresh and the dry weight of the samples. More details on the olive fruits data can be found in [44]. The third dataset to demonstrate the classification task consists of the NIR spectra of mayonnaise samples made from oils of 6 different origins. The classification task can be understood as a 6-class problem. The mayonnaise dataset was obtained from the official website of ChemHouse (www. chemproject.org). To make the data fit for multi-block analysis, each dataset was split into two blocks in the spectral domain. A further description of the datasets is provided in Table 1.

407 *Table 1: Description of the datasets used in the demonstration.*

Samples	Task	Calibration Block 1	Calibration Block 2	Test Block 1	Test Block 2	Y calibration	Y test
Olive oils (Discrete response)	Visualisation	83×300 (798-1375 nm)	83×270 (1377-1896 nm)			83×1	
Olive fruits (Continuous response)	Regression and variable selection	350×75 (708-930 nm)	350×60 (933-1113 nm)	145×75 (708-930 nm)	145×60 (933-1113 nm)	350×1	145×1
Mayonnaise (Discrete response)	Classification	72×150 (1100-1696 nm)	72×201 (1700-2500 nm)	72×150 (1100-1696 nm)	72×201 (1700-2500 nm)	72×1	72×1

408

409

410

411

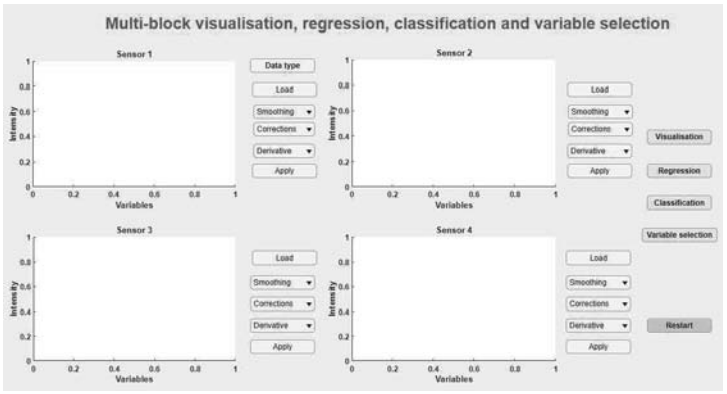
Operating procedure and demo analysis

The operating procedures are presented in separate sections to demonstrate the use of the MBA-GUI for the analysis of different datasets so that a person with minimal experience can repeat the analysis and use the GUI in their day-to-day multi-block data analysis tasks. The sections are data loading and pre-processing, data visualisation, regression, classification, variable selection and SPORT. All the figures presented in the analysis come directly from the MBA-GUI.

Data loading and pre-processing

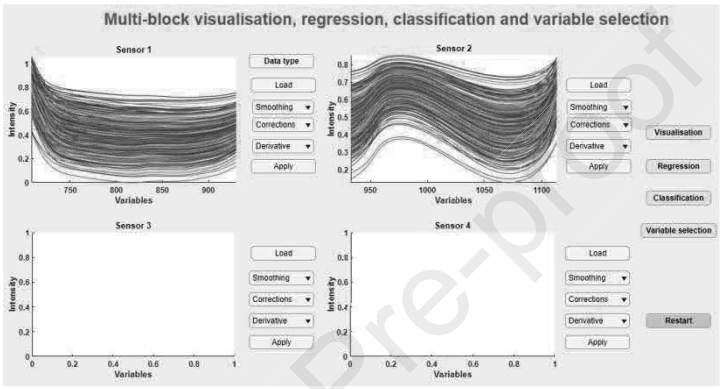
Fig. 5A shows the MBA-GUI interface for loading datasets. Currently, two data formats can be loaded, i.e., .xls and .csv. It is currently possible to load up to 4 blocks of data. The datasets can either come from several sensors or can be the same dataset imported multiple times to perform SPORT fusion. The datasets should be loaded in a logical order since sequential loading of the dataset may improve the performance of sequential methods where the natural order of blocks can provide meaningful insights into the datasets. Once the data are loaded, the figure in the MBA-GUI will be updated to show the data. Fig. 5B shows an example where two data blocks are loaded. Each data block can be separately pre-processed. The pre-processing can be performed in three steps - smoothing, normalisation or scatter correction, and derivatives. Derivatives are of particular use for NIR data where they can reveal underlying peaks. An example of pre-processing is shown in Fig. 5B. The pre-processing choice in Fig. 5C involved SAVGOL smoothing, SNV normalisation and no-derivative. Fig. 5C shows that once the pre-processing is selected the MBA-GUI will show the new pre-processed spectra. The user is free to explore multiple pre-processing methods by visualising how pre-processing affects the spectra. After pre-processing, the data are ready for multi-block analysis. There are five push-button options provided - visualisation,

regression, classification, variable selection and SPORT. There is also an option to restart the complete analysis by clearing all the previous data and operation logs. A point to note is that if there is only one data block, the MBA-GUI options will take the user to standard chemometric analysis where analyses such as PCA, PLS and variable selection can be performed. Due to limited space, the one block chemometric analysis is not presented in this article.



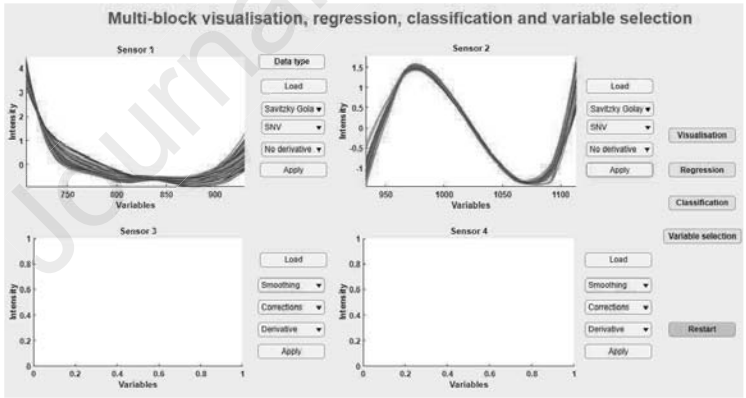
445

446 **(A)**



447

448 **(B)**



449

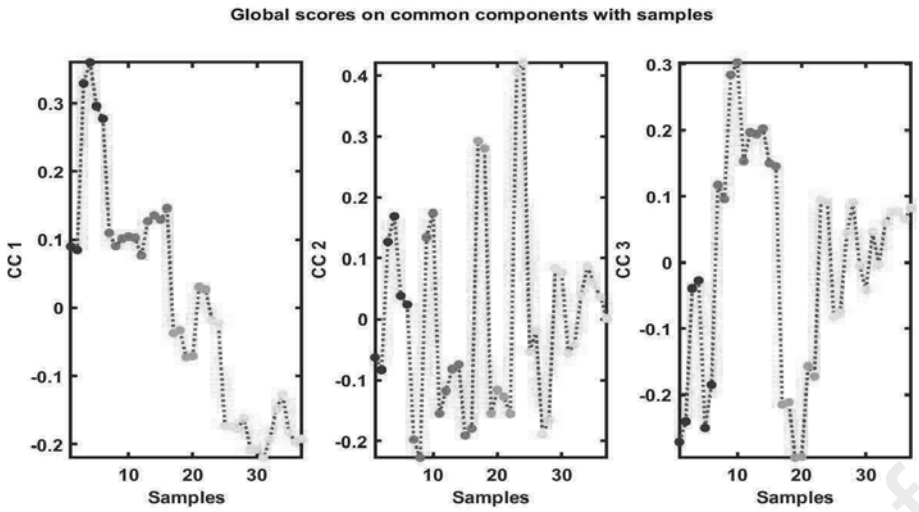
450 **(C)**

451 *Fig. 5: MBA-GUI interface for data loading. (A) Up to four different data blocks can be*
452 *loaded and analysed. (B) GUI interface once the data are loaded using the Load button. The*
453 *spectra can be loaded in a sequential order i.e. 1, 2, 3 and 4. Once the data are loaded, all of*
454 *the pre-processing methods can be applied. (C) MBA-GUI interface after pre-processing.*
455 *Once the pre-processing option is selected and the Apply button is pressed, the figures will*
456 *contain the pre-processed spectra.*

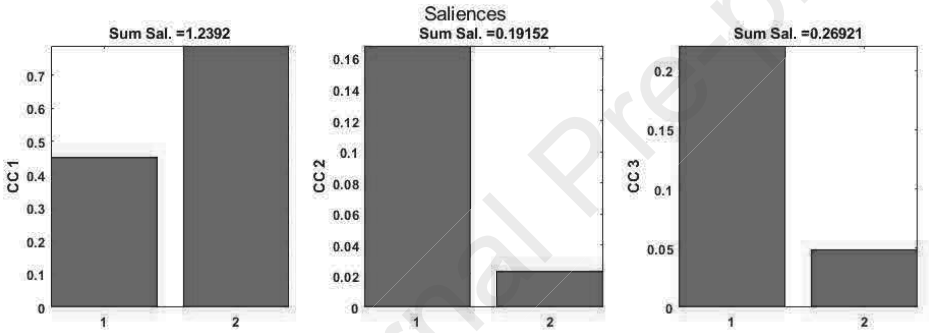
Data visualisation

Multi-block data visualisation is a challenging task where the interest is not only in the extraction of latent variables but also in how different blocks are linked with each other or contribute to the extraction of the global LVs. In the present case, the objective is also to limit the extraction of redundant information from the different blocks. In the MBA-GUI, multiple multi-block data visualisation methods are implemented. The presentation here will be limited to the use of ComDim-PLS, which is a supervised common dimension extraction method requiring a response variable to orient the decomposition of the blocks. It will be applied to the olive oil dataset.

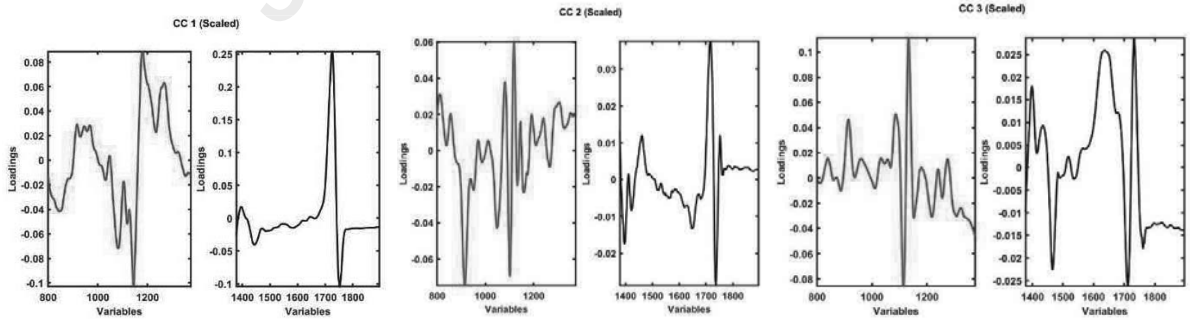
Fig. 6A, 6B and 6C show the scores when a model with three common components (CCs) is selected. Different coloured points in the figure indicate samples belonging to different classes as defined by the Y vector. It can be seen that with 1st and the 3rd CCs, a clear distinction of different classes is possible. Fig. 6D, 6E and 6F show the saliences for each block that contributed to the CCs. In this case, the higher saliences for block 2 show that CC1 was dominated by the information from that block, whereas CC2 and CC3 were dominated by the information from block 1. Since they were normalised, each block has a total salience of 1 which means, in the present case, a total salience of two. In the figure, we can see the amount of salience extracted by each block for each CC as well as the total salience extracted by each CC. It is also possible to see that the total amount extracted from the 2 blocks by 3 CCs is about 1.75. Fig. 6G, 6H, 6I show the loadings for each CC presented in two figures each, corresponding to the two-blocks. Such loading plots can be used to understand which variables are of interest.



(A). **(B).** **(C).**



(D). **(E).** **(F).**

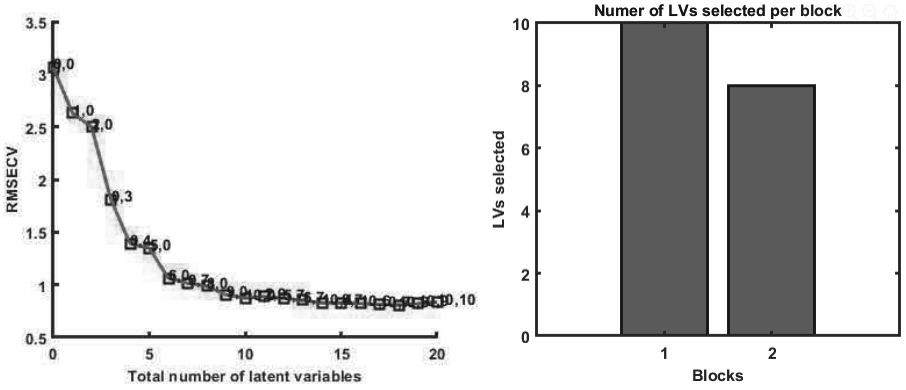


(G). **(H).** **(I).**

Fig. 6: Output from ComDim-PLS performed on the olive oil dataset. (A) Scores from CC1, (B) scores from CC2, (C) scores from CC3, (D) saliences for CC1, (E) saliences for CC2, (F) saliences for CC3, (G) loading for CC1, (H) loading for CC2, and (I) loading for CC3.

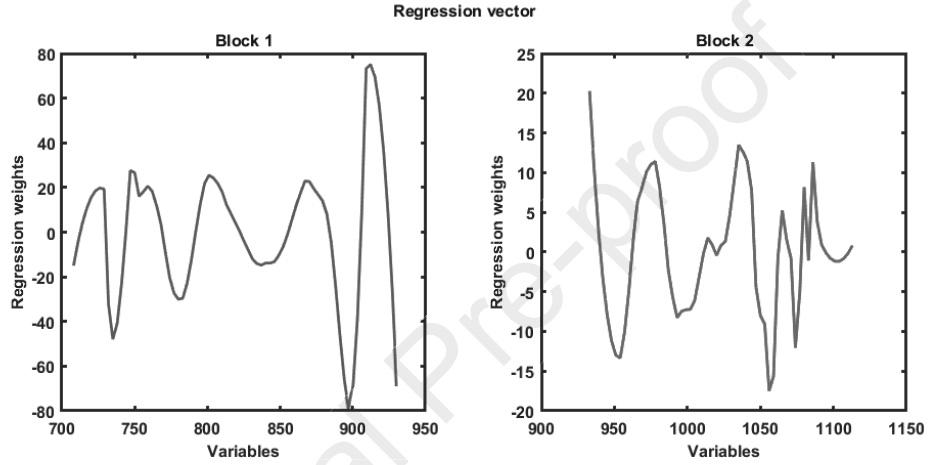
Regression

Multi-block regression is useful when multiple sensors are integrated to improve the predictive performance of the model. In the GUI, two types of multi-block regression are implemented, i.e., SO-PLS and MLR, both based on the ComDim-PLS scores. An example using SO-PLS is presented here. Fig. 7 presents the output of the SO-PLS analysis, where Fig. 8A shows the cross-validation error and Fig. 7B shows the number of LVs selected for each data block, i.e., 10 LVs for block 1 and 8 LVs for block 2. Fig. 7C and 7D show the final regression vector based on the LVs extracted from blocks 1 (Fig. 7C) and 2 (Fig. 7D). The results from the SO-PLS modelling are presented as the calibration plots in Fig. 7E and 7F. The R^2 for calibration and prediction were 0.95 and 0.93, respectively. The calibration and prediction errors were 0.69 and 0.66 % dried matter, respectively. It should be noted that when only one block of data is used, the R^2 is lower and the error is higher (as shown in single block analysis in Fig. 10), showing the benefit of multi-block regression.



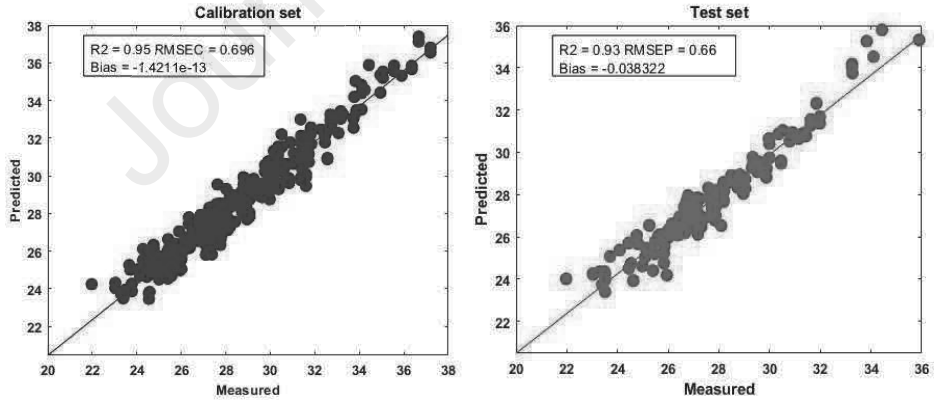
507

508 **(A).** **(B).**



509

510 **(C).** **(D).**



511

512 **(E).** **(F).**

513 *Fig. 7: The output from SO-PLS regression performed on the olive fruit dry matter dataset.*

514 *(A) Cross-validation error, (B) number of LVs selected from each block, (C) regression*

515 *vector from block 1, (D) regression vector from block 2, (E) calibration set results, and (D)*

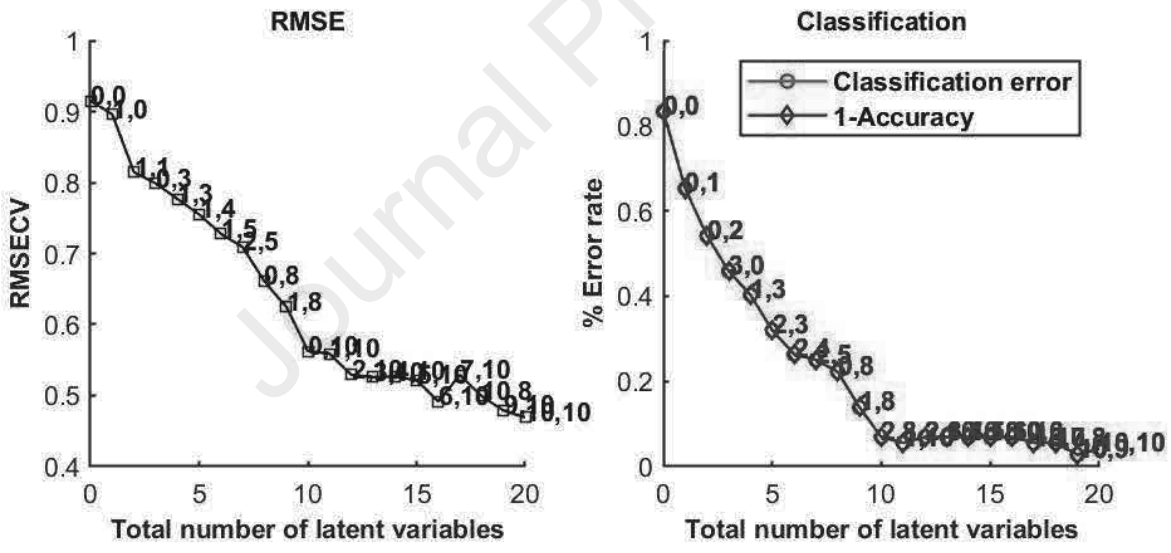
516 *test set results.*

517

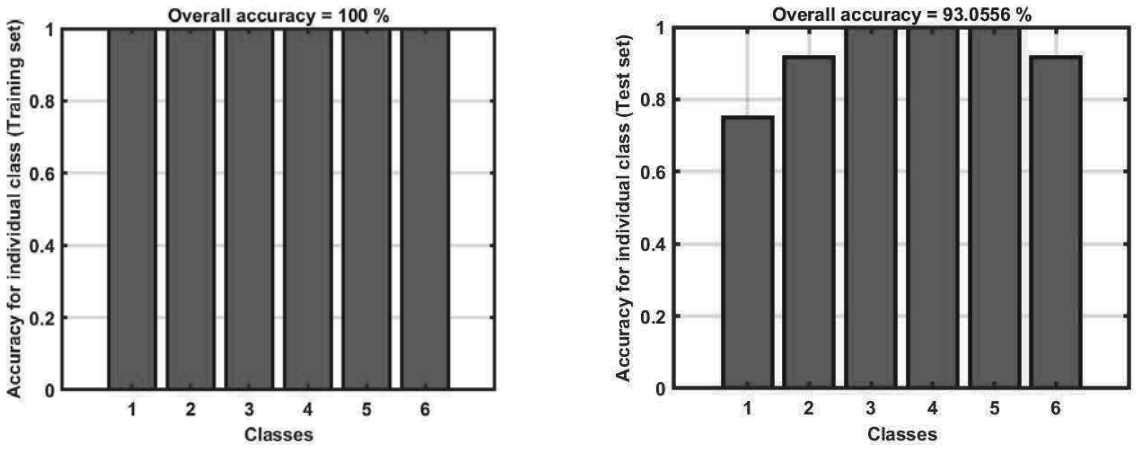
Classification

Multi-block classification involves the use of multiple sensor data to improve the classification accuracies. In the MBA-GUI toolbox, several multi-block classification techniques are implemented. Here an example of SO-PLS-LDA on the mayonnaise dataset is presented.

Fig. 8A and 8B shows the RMSE and the error evolution, respectively, as a function of the number of LVs. The RMSE and error plot were used for automatic selection of the number of latent variables for the two data blocks. The classification results from the calibration and test sets are presented in Fig. 8C and 8D respectively. An overall prediction accuracy of 93% was obtained.



(A). (B).

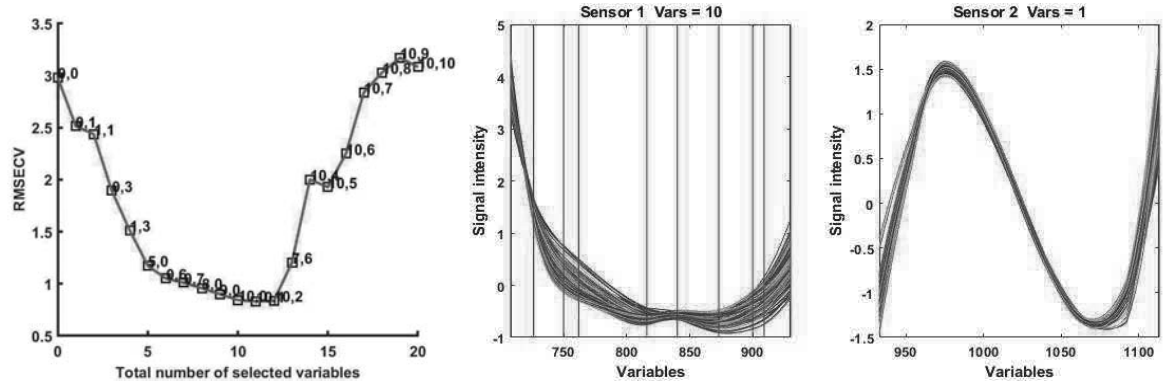


(C). **(D).**

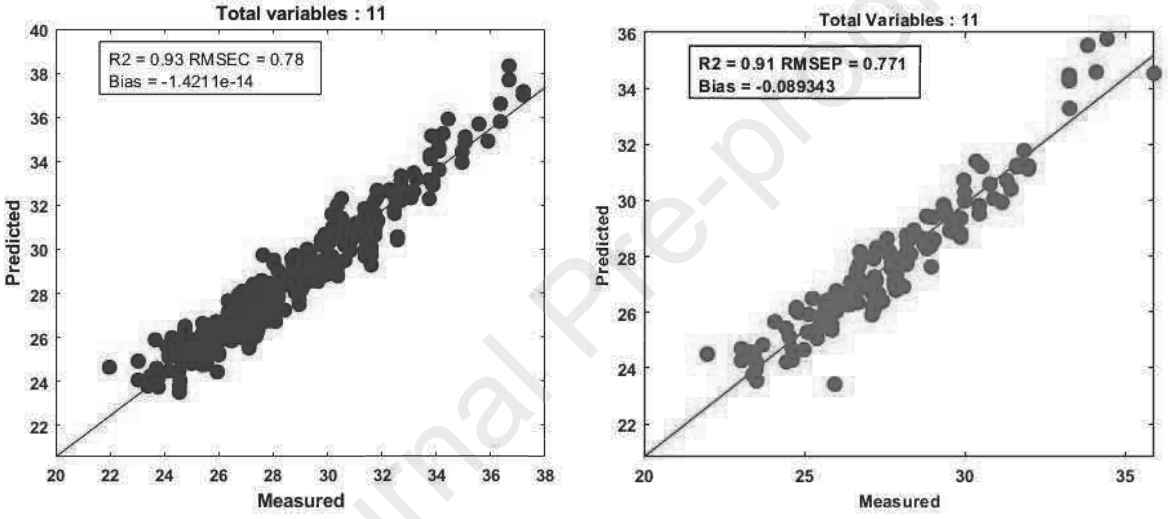
Fig. 8: Results from SO-PLS-LDA of the mayonnaise dataset. (A) RMSE as a function of the number of variables selected, (B) error rate as function of number of variables selected, (C) accuracy on calibration data, and (D) accuracy on test data.

Variable selection

Variable selection is a key step to identify the predictive variables that are most responsible for explaining the response variables. Variable selection can give a better understanding of the important parameters and, in many cases, help in the development of cheap multi-spectral sensor systems. In this work, a demonstration of multi-block variable selection using SO-CovSel is given. The analysis was carried out on the olive fruits dry matter content dataset, where two data blocks in the NIR range are used to predict the dry matter in olives. Currently, only CovSel variable selection is integrated into the MBA-GUI, however, it can be used for both continuous (regression) and discrete (classification) response variables. A cross-validation option is also provided which supports the selection of key variables. Once the response variables are loaded, the calibrate button can be used and the results will appear in new figures. Fig. 9 shows the outcome of the SO-CovSel analysis where Fig. 9A shows the cross-validation error, Fig. 9B shows the variables selected from block 1 and Fig. 9C shows the variables selected from block 2. A total of 11 variables were selected, 10 from block 1 and only 1 from block 2. The number of selected variables is almost 1/10 of the initial 146 in the two blocks. The calibration and prediction R^2 were 0.93 and 0.91, respectively (Fig. 9D and 9E), and the RMSEC and RMSEP values were 0.78 and 0.77 % dried matter. Although there was a slight decrease in R^2 and a slight increase in RMSEP with the models based on selected variables compared to SO-PLS regression, it should be noted that the model is now much simpler as it includes only 11 variables.



(A). **(B).** **(C).**



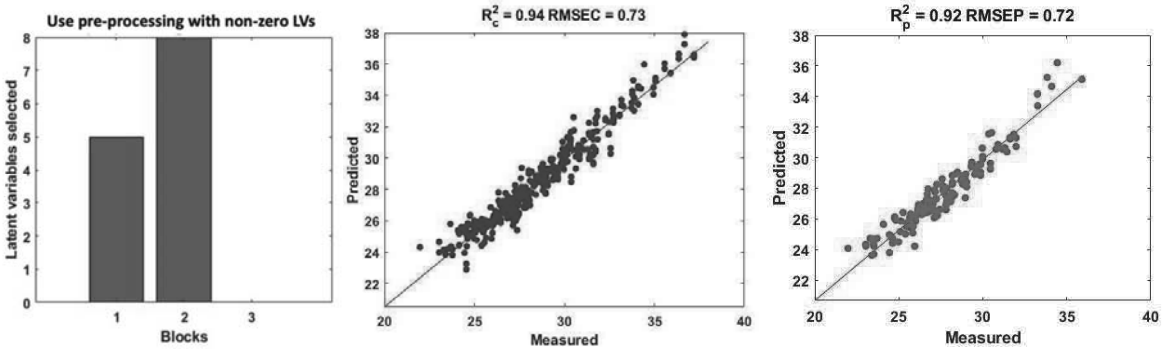
(D). **(E).**

Fig. 9: Output results for SO-CovSel analysis performed on the olive fruits dry matter content dataset. (A) Error plot for variable extraction, (B) variables selected in block 1, (C) variables selected in block 2, (D) calibration set, and (E) test set.

573 SPORT
574

575 Choosing the best pre-processing technique can sometimes be a challenging task such as in
576 the case of spectral data. However, multiple pre-processing techniques can provide
577 complementary information; for example, pre-processing with a derivative can help in
578 revealing the underlying peaks and techniques such as SNV can help to reduce multiplicative
579 effects. In the present MBA-GUI, a newly developed approach called SPORT is also
580 integrated. Using SPORT is completely automated and just requires the user to load the same
581 data in multiple blocks. The user can load the same data up to four times, and therefore, can
582 apply a different combination of pre-processing which the user thinks are the best candidates
583 for the type of data.

584 In the present case, three-blocks were assigned with three different types of pre-processing.
585 The first block was pre-processed using SNV, the second block has a combination of
586 SAVGOL smoothing and VSN, and the third block had a combination of SAVGOL and
587 MSC. After performing SPORT, the best pre-processing options were selected highlighting
588 the LVs extracted from each block (Fig. 10A). The block having zero LVs is not useful, and
589 therefore, the pre-processing associated with this block provides no improvement over the
590 pre-processing previously used. In the present case, SNV and VSN pre-processing both had
591 non-zeros LVs, in contrast to MSC pre-processing which had zero LVs, meaning that MSC is
592 not useful in this case. Further, both SNV (5 LVs) and VSN (8 LVs) pre-processing has a
593 significant number of LVs in the final calibration. The model obtained had a R^2 of calibration
594 and prediction of 0.94 and 0.92, respectively. Further, the RMSEC and RMSEP were 0.73 %
595 and 0.72 %, respectively. In summary, SPORT identified the best pre-processing method.



(A) (B). (C).

Fig.10: Selection of the best pre-processing options and their fusion with the SPORT methodology. (A) Latent variables selected from each block, (B) calibration set modelled with selected pre-processing fusion, and (C) test set.

Conclusion

A MATLAB based GUI for multi-block data analysis (MBA-GUI) is presented. The toolbox can perform a range of common pre-processing methods on blocks of multivariate data. Multi-block data analysis for regression, classification, visualisation, variable selection and SPORT are proposed. The performance of the MBA-GUI for each of the data analysis tasks was demonstrated with several data sets. The results showed that the MBA-GUI performed well, and all the options are fully functional. The main advantage of the toolbox is that it can be easily understood and used by non-experts. The first version of the GUI can be downloaded at (https://github.com/puneetmishra2/Multi-block.git). Other features will be added to the GUI with the development of new methods. All the data analysis presented in this work can be replicated with the supplied data. The GUI supports data format of .csv, .xlsx and .mat. The users are welcome to notify the authors if they find any bug or problem related to the use of the toolbox, so that the toolbox can be continuously improved along time. The app can be directly installed in MATLAB or can be used as stand by installing the MATLAB 2018b run time compiler tool at (https://nl.mathworks.com/products/compiler/matlab-runtime.html). The password to the start the toolbox is “welovedata” without double colon.

Validation

Dr. Raffaele Vitale

1) U. Lille, CNRS, LASIRE, Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l’Environnement, Cité Scientifique, F-59000, Lille, France

2) Molecular Imaging and Photonics Unit, Department of Chemistry, KU Leuven, Celestijnenlaan 200F, B-3001, Leuven, Belgium

Email: rvitale86@gmail.com

The MBA-GUI Toolbox v. 1.0 was successfully executed on three different versions of MATLAB (R2016b, R2017b and R2019a) without any issue to report. The Graphical User Interface (GUI) for data loading and visualisation allows a maximum number of 4 distinct numerical arrays to be imported, represented and pretreated by means of a significant amount of computational tools for smoothing, derivation and other types of corrections (*e.g.*, Standard Normal Variate – SNV – Multiplicative Scatter Correction – MSC – *etc.*). It is intuitive and very well-conceived.

The imported datasets can afterwards be processed by a remarkable collection of multi-block latent variable-based dimensionality reduction methodologies permitting tasks of various nature to be carried out:

1. Variable selection (Sequential Orthogonalised Covariance Selection and Sequential Orthogonalised Covariance Selection-Linear Discriminant Analysis);
2. data exploration/visualisation (Common Dimension – ComDim – Principal Component Analysis, ComDim Canonical Correlation Analysis, ComDim Independent Component Analysis and ComDim Partial Least Squares regression);
3. multivariate regression (Sequential Orthogonalised Partial Least Squares regression and ComDim Regression); and
4. multivariate classification (Sequential Orthogonalised Covariance Selection-Linear Discriminant Analysis, ComDim Independent Component Analysis-Linear Discriminant Analysis, ComDim Principal Component Analysis-Linear Discriminant Analysis and Sequential Orthogonalised Partial Least Squares Discriminant Analysis).

All the toolbox menus are extremely easy to browse and their design is capable of guiding even non-expert users through the sequential steps of multi-block data analysis. They enable not only model training/calibration, but also (when it holds) model optimisation (by two different approaches for cross-validation) and external testing/validation. Furthermore, it is worth mentioning that the presented GUI offers as a valuable add-on an implementation of a recently proposed strategy named SPORT (Sequential Preprocessing through ORThogonalisation) for both the selection of the best pretreatment technique and the extraction of complementary information from the outcomes of multiple preprocessing operations.

In its ensemble, the developed toolbox constitutes a comprehensive software suite to address multi-block data analysis problems, which shows a great potential for attracting practitioners by *making their life easier* in scenarios that might exhibit particularly high complexities.

Disclaimer

The MBA-GUI is free to use for public as it also involves some algorithms from public sources. Great care has been taken while developing the MBA-GUI, however, the authors do not accept any responsibility or liability.

Acknowledgments

PM and AN acknowledge the funding received from the European Union's Horizon 2020 research and innovation program named MODLIFE (Advancing Modelling for Process-Product Innovation, Optimization, Monitoring and Control in Life Science Industries) under the Marie Sklodowska-Curie grant agreement number 675251.

References

[1] L.L. Simon, H. Pataki, G. Marosi, F. Meemken, K. Hungerbühler, A. Baiker, S. Tummala, B. Glennon, M. Kuentz, G. Steele, H.J.M. Kramer, J.W. Rydzak, Z. Chen, J. Morris, F. Kjell, R. Singh, R. Gani, K.V. Gernaey, M. Louhi-Kultanen, J. O'Reilly, N. Sandler, O. Antikainen, J. Yliruusi, P. Frohberg, J. Ulrich, R.D. Braatz, T. Leyssens, M. von Stosch, R. Oliveira, R.B.H. Tan, H. Wu, M. Khan, D. O'Grady, A. Pandey, R. Westra, E. Delle-Case, D. Pape, D. Angelosante, Y. Maret, O. Steiger, M. Lenner, K. Abbou-Oucherif, Z.K. Nagy, J.D. Litster, V.K. Kamaraju, M.-S. Chiu, Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review, Organic Process Research & Development, 19 (2015) 3-62.

[2] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Analytica Chimica Acta, 891 (2015) 1-14.

[3] H. Zheng, A. Cai, Q. Zhou, P. Xu, L. Zhao, C. Li, B. Dong, H. Gao, Optimal preprocessing of serum and urine metabolomic data fusion for staging prostate cancer through design of experiment, Anal Chim Acta, 991 (2017) 68-75.

[4] T.G. Doeswijk, A.K. Smilde, J.A. Hageman, J.A. Westerhuis, F.A. van Eeuwijk, On the increase of predictive performance with high-level data fusion, Anal Chim Acta, 705 (2011) 41-47.

[5] Q. Ouyang, J. Zhao, Q. Chen, Instrumental intelligent test of food sensory quality as mimic of human panel test combining multiple cross-perception sensors and data fusion, Anal Chim Acta, 841 (2014) 68-76.

[6] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication, Anal Chim Acta, 820 (2014) 23-31.

[7] A.R. Martínez Bilesio, M. Batistelli, A.G. García-Reiriz, Fusing data of different orders for environmental monitoring, Anal Chim Acta, 1085 (2019) 48-60.

[8] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Information Fusion, 57 (2020) 115-129.

[9] A.K. Smilde, I. Måge, T. Næs, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, Journal of Chemometrics, 31 (2017) e2900.

[10] M. Alinaghi, H.C. Bertram, A. Brunse, A.K. Smilde, J.A. Westerhuis, Common and distinct variation in data fusion of designed experimental data, Metabolomics, 16 (2019) 2.

[11] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, Journal of Chemometrics, 33 (2019) e3085.

[12] Y. Song, J.A. Westerhuis, A.K. Smilde, Separating common (global and local) and distinct variation in multiple mixed types data sets, Journal of Chemometrics, 34 (2020) e3197.

[13] J. Boccard, D.N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multi-block Omics data fusion, Analytica Chimica Acta, 769 (2013) 30-39.

[14] J. Boccard, D.N. Rutledge, Iterative weighting of multi-block data in the orthogonal partial least squares framework, Anal Chim Acta, 813 (2014) 25-34.

[15] A. Biancolillo, T. Næs. The Sequential and Orthogonalized PLS Regression for Multi-block Regression: Theory, Examples, and Extensions. In: M. Cocchi (Ed.), Data Fusion Methodologies and Applications, Data Handling in Science and Technology, vol. 31, Elsevier, Oxford, UK, 2019, pp. 157-177.

[16] A. Biancolillo, F. Marini, J.-M. Roger, SO-CovSel: A novel method for variable selection in a multi-block framework, Journal of Chemometrics, 34 (2020) e3120.

[17] B. Galindo-Prieto, P. Geladi, J. Trygg, Multi-block variable influence on orthogonal projections (MB-VIOP) for enhanced interpretation of total, global, local and unique variations in OnPLS models, arXiv preprint arXiv:2001.06530, (2020).

[18] E. Acar, M.A. Rasmussen, F. Savorani, T. Næs, R. Bro, Understanding data fusion within the framework of coupled matrix and tensor factorizations, Chemometr Intell Lab, 129 (2013) 53-63.
[19] E. Acar, E.E. Papalexakis, G. Gürdeniz, M.A. Rasmussen, A.J. Lawaetz, M. Nilsson, R. Bro, Structure-revealing data fusion, BMC Bioinformatics, 15 (2014) 239.
[20] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, Chemometrics and Intelligent Laboratory Systems, 124 (2013) 32-42.
[21] K. De Roover, E. Ceulemans, M.E. Timmerman, How to perform multi-block component analysis in practice, Behavior Research Methods, 44 (2012) 41-56.
[22] V. Cariou, D. Jouan-Rimbaud Bouveresse, E.M. Qannari, D.N. Rutledge, ComDim Methods for the Analysis of Multi-block Data in a Data Fusion Perspective. In: M. Cocchi (Ed.), Data Fusion Methodologies and Applications, Data Handling in Science and Technology, vol.31, Elsevier, Oxford, UK, 2019, pp. 179-204.
[23] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, . In: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, 2nd Ed., vol. 3, Elsevier, Amsterdam, The Netherlands, 2020, pp. 1-75.
[24] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing?, TrAC Trends in Analytical Chemistry, 50 (2013) 96-106.
[25] Å. Rinnan, F.v.d. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry, 28 (2009) 1201-1222.
[26] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, Applied Spectroscopy, 43 (1989) 772-777.
[27] T. Isaksson, T. Næs, The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy, Applied Spectroscopy, 42 (1988) 1273-1284.
[28] P.H.C. Eilers, Parametric time warping, Analytical Chemistry, 76 (2004) 404-411.
[29] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: Variable sorting for normalization, Journal of Chemometrics, 34 (2020) e3164.
[30] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics, Analytical Chemistry, 78 (2006) 4281-4290.
[31] Q. Guo, W. Wu, D.L. Massart, The robust normal variate transform for pattern recognition with near-infrared data, Anal Chim Acta, 382 (1999) 87-103.
[32] R. Bro, A.K. Smilde, Principal component analysis, Anal Methods-Uk, 6 (2014) 2812-2831.
[33] P. Geladi, Chemometrics in spectroscopy. Part 1. Classical chemometrics, Spectrochimica Acta Part B: Atomic Spectroscopy, 58 (2003) 767-782.
[34] E.M. Qannari, I. Wakeling, P. Courcoux, H.J.H. MacFie, Defining the underlying sensory dimensions, Food Quality and Preference, 11 (2000) 151-154.
[35] A. El Ghaziri, V. Cariou, D.N. Rutledge, E.M. Qannari, Analysis of multi-block datasets using ComDim: Overview and extension to the analysis of (K + 1) datasets, Journal of Chemometrics, 30 (2016) 420-429.
[36] D. Rutledge, Novel extensions and applications of Common Components Analysis in chemometrics, Twelfth Winter Symposium on ChemometricsSaratov (Russia), 2020.
[37] A. Biancolillo, T. Næs, R. Bro, I. Måge, Extension of SO-PLS to multi-way arrays: SO-N-PLS, Chemometr Intell Lab, 164 (2017) 113-126.
[38] T. Næs, O. Tomic, B.H. Mevik, H. Martens, Path modelling by sequential PLS regression, Journal of Chemometrics, 25 (2011) 28-40.
[39] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, Chemometr Intell Lab, 141 (2015) 58-67.
[40] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, CovSel: Variable selection for highly multivariate and multi-response calibration Application to IR spectroscopy, Chemometr Intell Lab, 106 (2011) 216-223.

776 [41] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization
777 (SPORT) and its application to near infrared spectroscopy, Chemometrics and Intelligent Laboratory
778 Systems, 199 (2020) 103975.
779 [42] W.B. Zheng, H.P. Shu, H. Tang, H.Q. Zhang, Spectra data classification with kernel extreme
780 learning machine, Chemometr Intell Lab, 192 (2019).
781 [43] H.S. Tapp, M. Defernez, E.K. Kemsley, FTIR spectroscopy and multivariate analysis can
782 distinguish the geographic origin of extra virgin olive oils, Journal of Agricultural and Food Chemistry,
783 51 (2003) 6110-6115.
784 [44] X. Sun, P. Subedi, R. Walker, K.B. Walsh, NIRS prediction of dry matter content of single olive
785 fruit with consideration of variable sorting for normalisation pre-treatment, Postharvest Biology and
786 Technology, 163 (2020) 111140.

787

788

789

790

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.