Chemometrics and Intelligent Laboratory Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy

Puneet Mishra [a,*], Jean Michel Roger [b,c], Federico Marini [d], Alessandra Biancolillo [e], Douglas N. Rutledge [f,g]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] ITAP, INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France
[d] Department of Chemistry, University of Rome "La Sapienza", Piazzale, Aldo Moro 5, 00185, Rome, Italy
[e] Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, 67100, Coppito, L'Aquila, Italy
[f] Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 75005, Paris, France
[g] National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia

ARTICLE INFO

Keywords:
Multi-block data analysis
Data fusion
Partial least squares
Pre-processing
Parallel and orthogonalized partial least squares (PO-PLS)

ABSTRACT

Data generated from spectroscopy may be deformed by artefacts due to a range of physical, chemical and environmental factors that are not of interest for the characterization of the samples under study. For example, data acquired by near-infrared (NIR) spectroscopy in the diffuse reflectance mode can be affected by light scattering. This artefact, if not reduced or removed by spectral pre-processing, can complicate the multivariate data analysis. However, different pre-processing approaches correct these effects in different ways. For example, differentiation can reveal underlying bands, while spectral normalization techniques such as standard normal variate (SNV) can correct for multiplicative and additive effects. Combining multiple pre-processing techniques can lead to better results. However, it is not feasible for a user to explore all possible combinations of pre-processing techniques. In the present work, a new pre-processing fusion approach, based on the framework of separating common and distinct components in multi-block multivariate data analysis, is demonstrated. The approach utilizes parallel and orthogonalized partial least squares (PO-PLS) regression for the parallel fusion of multiple pre-processing techniques applied to the same data. The results obtained on 4 different NIR spectroscopic data sets related to the assessment of fruit quality and used as benchmark are compared to those of the recently developed sequential pre-processing through orthogonalization (SPORT) approach: it is found that, in all the cases, the PO-PLS approach leads to slightly better performances. Furthermore, a clear understanding of the common and distinct information present in the data sets after each pre-treatment was obtained. Parallel pre-processing through orthogonalization (PORTO) can be seen as parallel boosting of multiple pre-processing techniques to improve model performances.

1. Introduction

Data generated from spectroscopy may be deformed by artefacts due to a range of physical, chemical and environmental factors that are not of interest for the characterization of the samples under study [1,2]. For example, NIR spectroscopic measurements done in diffuse reflectance mode contain scattering effects which, in the majority of cases, may mask the absorption features related to the chemical components present in the samples [3–6], even if there are a few particular situations, e.g., when the physical properties such as granulometry have to be modeled, in which scattering itself constitutes the relevant information. These artefacts in the measured signal can complicate the multivariate data analysis, especially if predictive models are to be created. Often, pre-treatment methods are used to reduce/remove such artefacts that are unrelated to the property to be predicted [1]. For example, if one would like to use NIR spectroscopy to predict moisture content in a fresh fruit, then, the objective of pre-processing would be to remove the scattering effects while retaining the absorption features related to water so as to use them for the data modelling.

Pre-processing is required to correct for scattering, baseline shifts, noise and other sources of unwanted/spurious variation [2,3]. Several methods are available to remove these effects. In the case of spectral data

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.chemolab.2020.104190
Received 23 July 2020; Received in revised form 16 October 2020; Accepted 18 October 2020
Available online xxxx
0169-7439/© 2020 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article as: P. Mishra et al., Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy, Chemometrics and Intelligent Laboratory Systems, https://doi.org/10.1016/j.chemolab.2020.104190

Table 1
A description of the data sets used in the study.

Food	Spectral range (nm)	Calibration	Test	Reference	Source
Apples	500–1018	437	188	SSC (%)	[23]
Olives	669–1122	171	81	DM (%)	[24]
Mangoes	744–1092	1014	435	DM (%)	[27]
Pears	709–1125	329	142	MC (%)	Generated for the present study

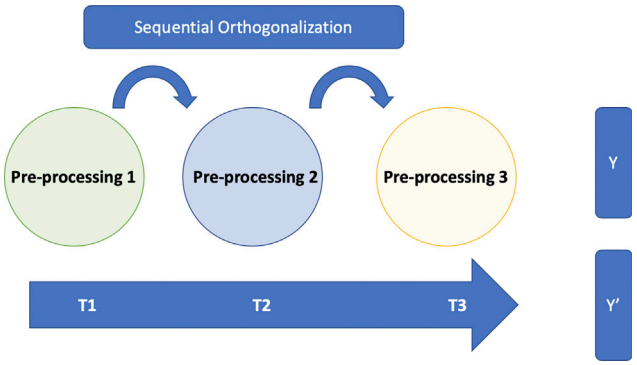


Fig. 1. An illustration of the SPORT approach for pre-processing fusion. SPORT involves sequential PLS and orthogonalization steps to extract scores from differently preprocessed data. Later all the scores are concatenated and used for ordinary regression analysis.

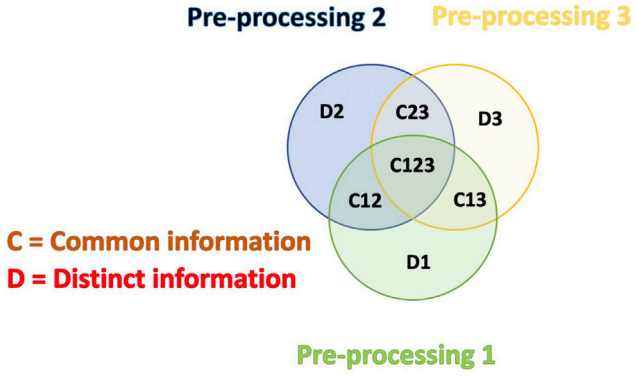


Fig. 2. A schematic illustration of the concept of common and distinctive information which constitutes the conceptual basis of PO-PLS and, as a consequence, of the PORTO approach to pre-processing fusion. The Fig. is adapted from the one presented in Ref. [19].

with broad bands such as UV, visible or near infrared spectroscopy, high frequency noise is commonly removed using smoothing methods such as the Savitzky-Golay filter (SAVGOL) [1,3]. The additive and multiplicative effects are dealt with by spectral normalization techniques such as standard normal variate (SNV) [7], variable sorting for normalization (VSN) [8], and scatter correction methods, such as multiplicative scatter correction (MSC) [9] and its extended version (EMSC), which expands the range of effects which can be removed from the signal by taking into account, e.g., polynomial baselines, known analyte profiles or interferents [10]. Furthermore, differentiation of the signal is commonly used to reveal underlying bands [1]. However, there are so many pre-processing methods available that it is difficult for a user to optimally test all possible combinations of techniques, in order to select the best combination for their data. To deal with this, a design of experiments (DoE) based approach to the selection of combinations of pre-processing techniques has been proposed [11]. The DoE based approach aims to select the pre-processing techniques and their sequential combination by exploring their effects on the model performances.

The complementarity of different pre-processing techniques is apparent and ensemble approaches to pre-processing fusions are emerging [12–15]. Another interesting approach to pre-processing fusion is sequential pre-processing through orthogonalization (SPORT) [14,16,17]. SPORT is based on the concept of multi-block data analysis and is inspired by sequential and orthogonalized partial least squares (SO-PLS) regression. However, SPORT performs the pre-processing fusion in a sequential way, which implies that the order of pre-processing might impact the result. This is inherited from the SO-PLS approach which processes the blocks by means of a sequential extraction of information [18]. However, it is difficult to objectively define this order.

In the present work, a new approach utilizing parallel and orthogonalized partial least squares (PO-PLS) regression for the fusion of multiple pre-processing techniques is proposed. The approach assumes that all pre-processing techniques are of equal importance, and therefore, their parallel fusion can lead to improved outcomes with respect to the sequential fusion which requires a pre-defined order for the sequential inclusion of the pre-processing techniques. Furthermore, the parallel approach is based on identifying the common and distinct information in the blocks corresponding to the different pre-processing techniques [19–21]. This means that pre-processed data carry some common information irrespective of the pre-treatment involved and, at the same time, a distinct information which is specific to the pre-processing technique. The proposed fusion approach has been given the name parallel pre-processing though orthogonalization (PORTO). In the present study, the PORTO approach was tested on 4 different real case data sets related to the use of NIR spectroscopy for quality prediction in fresh fruits. Furthermore, the parallel PORTO method is compared with the recently developed sequential SPORT pre-processing fusion approach.

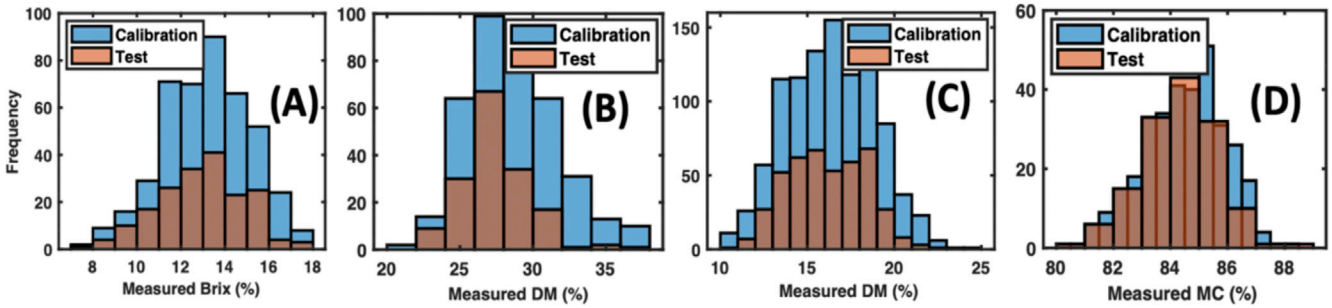


Fig. 3. Distribution of the reference values for each data set used for calibration and test set. (A) Apple data set, (B) Olive data set, (C) Mango data set, and (D) Pear data set.

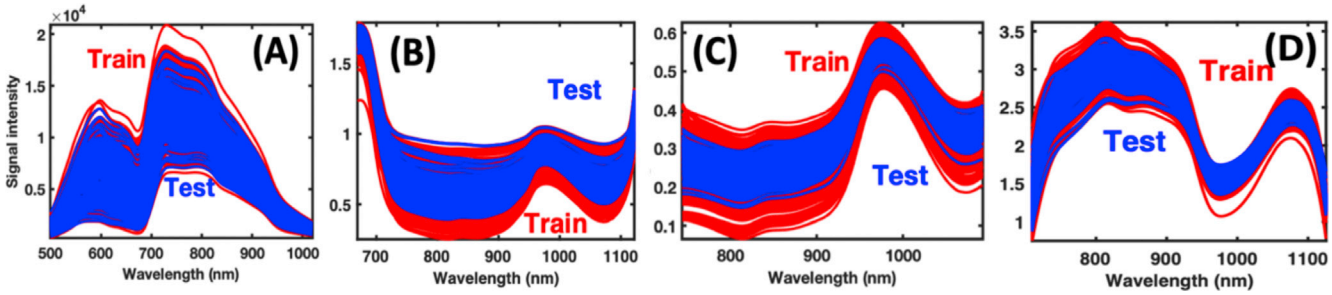


Fig. 4. Calibration (blue) and test (red) spectral profiles for different data sets. (A) Apple data set, (B) Olive data set, (C) Mango data set, and (D) Pear data set.

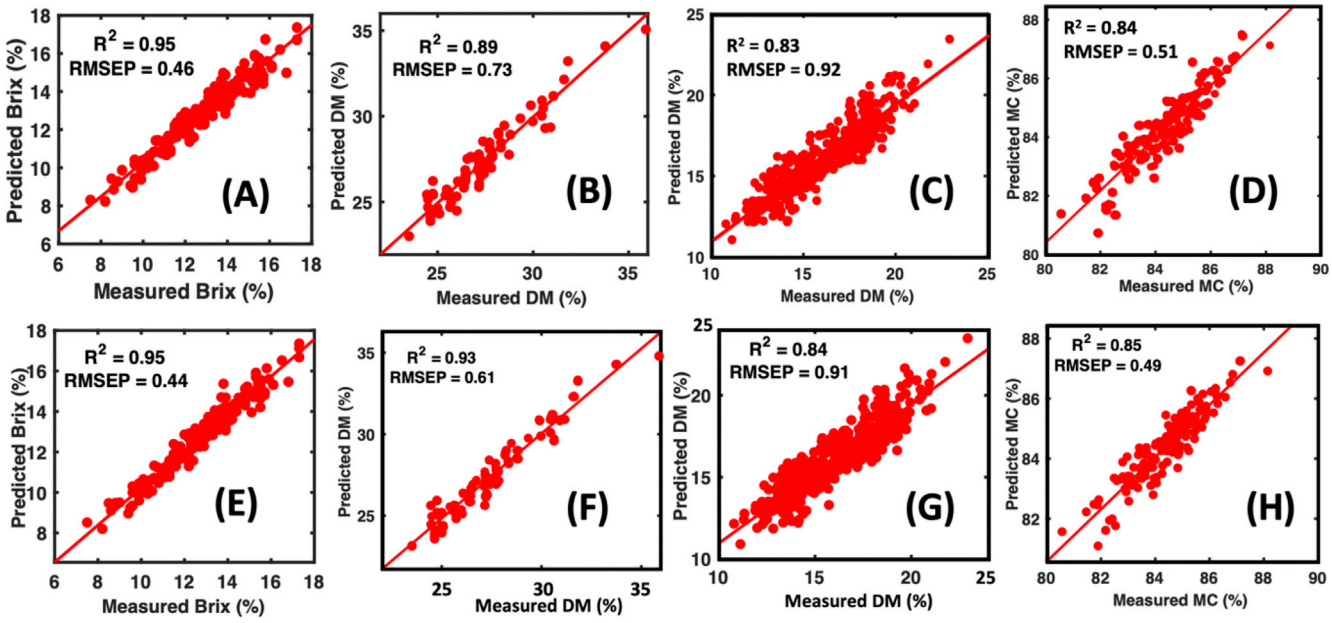


Fig. 5. Results of SPORT (upper row) and PORTO (lower row) modelling. Predictions on the test set samples: (A) and (E) Apple data set, (B) and (F) Olive data set, (C) and (G) Mango data set, and (D) and (H) Pear data set.

Table 2
Summary of the performances of PLS models built on data pre-treated with each of the considered pre-processing techniques and of the SPORT and PORTO approaches.

Pre-processing approach	Apple data set		Olive data set		Mango data set		Pear data set	
	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP
Raw	0.85	0.77	0.70	1.21	0.76	1.09	0.83	0.52
MSC	0.86	0.74	0.90	0.70	0.81	0.96	0.81	0.55
VSN	0.83	0.80	0.92	0.63	0.82	0.96	0.79	0.58
SNV	0.82	0.82	0.91	0.69	0.81	0.96	0.82	0.54
2^{nd} derivative	0.89	0.65	0.90	0.72	0.77	1.07	0.81	0.56
SPORT	0.95	0.46	0.89	0.73	0.83	0.92	0.84	0.51
PORTO	0.95	0.44	0.93	0.61	0.84	0.91	0.85	0.49

Table 3
Number of LVs selected from each pre-processing block by the SPORT approach.

Data sets/Pre-processing	Raw data	MSC	VSN	SNV	2nd derivative
Apple	7				11
Olive		4		4	1
Mango		8	7	1	1
Pear			2	1	10

2. Material and methods

2.1. Data sets

Four different NIR spectroscopic data sets related to the prediction of quality attributes in fresh fruits were used. These data sets were selected as they were all measured in diffuse reflectance mode, which leads to various additive and multiplicative effects due to light scattering. Often scatter correction methods are required to reduce/remove these effects prior to data modelling. A description of the data sets is provided in Table 1. All the samples were split into calibration (70%) and test sets (30%) using the Kennard-Stone algorithm [22]. The apple data set is related to soluble solids content (SSC) prediction in intact fruits [23]. The olive fruits and mango data sets are related to the prediction of dry matter (DM) in intact fruits [24]. The pear dataset was acquired specifically for this study and consists of 471 fruit samples. More details on the fruit samples can be found in Refs. [25,26]. The spectral measurements were carried out with a portable spectrometer Felix F-750 (Camas, WA, USA) with a Carl Zeiss MMS-1 detector (Oberkochen, Germany) to record the reflected light in the spectral range of 310–1130 nm with a spectra sampling at 3 nm, a Xenon Tungsten Lamp for illumination and a built-in white painted reference standard for setting the 100% value of the reflectance scale. The data acquisition was performed by placing the fruit in the sample holder and by manually pressing the scan button on the Felix device. For each pear, the spectral measurements were performed on the largest part of the hypanthium. The moisture content (MC) was measured by the hot air oven drying method.

2.2. Pre-processing methods

NIR spectra of fruit contain a range of scattering effects leading to additive and multiplicative distortion effects. In this work, four different pre-processing methods, i.e. multiplicative scatter correction [9], standard normal variate [7], variable sorting for normalization [8] and 2nd derivative (Savitzky-Golay with a 15 points window and 2nd order polynomial) were used for pre-processing fusion [1]. The set of pre-processings was selected to include both model-based (MSC and VSN) and model-free (SNV and 2nd derivative) techniques. All the pre-processing methods were implemented as discussed in Ref. [1].

3. SPORT

The SPORT approach to pre-processing fusion is based on two steps, i.e., a PLS regression step and a sequential orthogonalization step [14]. Firstly, a PLS regression model is fitted between the **Y** and the first pre-processed block. Then, the second block is orthogonalized with respect to the scores of the first regression. Then the residuals of **Y** are fitted to the orthogonalized second block and the procedure is continued for as many blocks as there are pretreatments. The main steps of the algorithm are schematically represented in Fig. 1. The algorithm for two pre-processing blocks ($\mathbf{X_1}$ and $\mathbf{X_2}$) as presented in Ref. [14] is as follows:

1. The **Y** responses are fitted to the $\mathbf{X_1}$ by the PLS regression
2. $\mathbf{X_2}$ is orthogonalized with respect to the scores obtained from the first regression
3. The orthogonalized $\mathbf{X_2}$ is used to predict the **Y** residuals
4. The overall predictive model is obtained by combining (concatenating the scores) the sub-models calculated in steps 1 and 3

Table 4
Common and distinct components selected by the PORTO approach. The '+' sign indicates that the common component is shared by the indicated blocks.

Data sets/ Pre-processing	Common components[a]		Distinct components	
Apple	4	1. RAW (26.7%, 0.997) +MSC (71.4%, 0.999) + VSN (69.9%, 0.998) + SNV (71.3%, 0.999) + 2nd derivative (27.5%, 0.997) 2. RAW (11.5%, 0.997) +MSC (15.5%, 0.998) + VSN (16.9%, 0.997) + 2nd derivative (35.5%, 0.997) 3. RAW (17.7%, 0.997) +MSC (6.0%, 0.997) + SNV (6.2%, 0.997) + 2nd derivative (6.5%, 0.996) 4. RAW (34.0%, 1.000) + 2nd derivative (27.1%, 1.000)	3	5 RAW (9.0%) 6 MSC (2.1%) 7 2nd derivative (0.5%)
Olive	5	1. RAW (5.1%, 0.996) + MSC (22.7%, 0.998) + VSN (26.8%, 0.994) + SNV (22.9%, 0.999) +2nd derivative (24.5%, 0.995) 2. MSC (12.4%, 0.999) +VSN (12.5%, 0.995) +SNV (12.4%, 0.999) + 2nd derivative (40.6%, 0.994) 3. RAW (29.9%, 0.997) + MSC (26.1%, 0.997) + 2nd derivative (18.6%, 0.995) 4. RAW (14.3%, 0.994) + SNV (2.7%, 0.990) +2nd derivative (3.3%, 0.988) 5. MSC (7.7%, 0.999) + VSN (53.4%, 0.998) + SNV (7.6%, 0.999)	3	6 RAW (42.2%) 7 MSC (10.7%) 8 VSN (0.2%)
Mango	5	1. RAW (4.7%, 0.999) + MSC (12.8%, 0.996) + VSN (4.4%, 0.989) + SNV (12.8%, 0.996) +2nd derivative (26.8%, 0.995) 2. RAW (5.8%, 0.999) + MSC (27.3%, 0.997) + VSN (6.0%, 0.985) + SNV (27.3%, 0.997) 3. RAW (16.2%, 0.998) + MSC (4.6%, 0.990) + VSN (9.8%, 0.980) + 2nd derivative (9.7%, 0.994) 4. RAW (22.7%, 0.998) + MSC (30.4%, 1.000) + SNV (30.4%, 1.000) 5. VSN (11.3%, 0.987) + 2nd derivative (7.6%, 0.987)	2	6 RAW (37.1%) 7 MSC (0.7%)
Pear	6	1. RAW (7.7%, 0.998) + MSC (29.5%, 0.998) + VSN (84.7%, 0.999) + SNV (30.0%, 0.999) + 2nd derivative (34.0%, 0.997) 2. RAW (9.0%, 0.998) + MSC (24.4%, 0.998) + SNV (24.2%, 0.999) + 2nd derivative (11.8%, 0.997) 3. RAW (11.5%, 0.997) + MSC (4.2%, 0.999) + SNV (4.3%, 0.999) 4. MSC (6.0%, 0.998) + VSN (9.6%, 0.999) + 2nd derivative (10.1%, 0.996) 5. MSC (20.6%, 0.998) + SNV (20.2%, 0.999) + 2nd derivative (11.5%, 0.996) 6. RAW (10.8%, 1.000) + 2nd derivative (8.6%, 1.000)	2	7 RAW (27.2%) 8 VSN (0.3%)

LVs: of SPORT vs PORTO for the olive dataset.
[a] The numbers in parentheses indicate the % explained variance for the block and, for the common components, the correlation coefficient, respectively.

The number of LVs is usually optimized in cross-validation using a global approach; all possible combinations of LVs are tested and the optimal one is the one resulting in the lowest RMSECV (or classification error in CV, depending on the property/ies to be predicted). In the present work, the algorithm presented in Ref. [14] was implemented by means of freely available multi-block data analysis graphical user interface [28] in MATLAB (Release 2017b; The MathWorks, Natick, USA).

3.1. Parallel pre-processing through orthogonalization

Parallel pre-processing through orthogonalization (PORTO) is inspired by and relies on PO-PLS regression [29]. The latter is a combination of PLS regression, generalized canonical analysis (GCA) and multiple orthogonalization steps. GCA is the extension of canonical correlation analysis (CCA) to more than 2 data blocks. The main aim of PORTO is to identify and extract common and distinct information within the blocks that result from differently pre-processed data, which can lead to improved data modelling. The concept of common and distinct component is illustrated in Fig. 2, where the three circles represent three pre-processings performed on the same data set and the letters D and C indicate the distinct and the common parts of the information. Like PO-PLS, the aim of PORTO is to first identify the information common to differently pre-processed data blocks and then orthogonalize each individual pre-processing block with respect to the sub-space identified as the common information. This leaves only the distinct information in each differently pre-processed data block, highlighting the distinctiveness of that pre-processing. The PORTO approach is basically performing the multi-block PO-PLS on the differently pre-processed data sets. The main steps of the PO-PLS algorithm, which constitutes the basis of PORTO, as presented in Ref. [29], are the following:

1. The same data set **X** is pre-treated with a specific number of user-defined pre-processing techniques (P): P matrices, each corresponding to an individual pre-processing applied to **X**, are obtained and altogether they constitute the multi-block data set.
2. Standard PLS models are calculated between the **Y** and each of the differently pre-processed blocks $\mathbf{X_p}$ ($p = 1,\ldots,P$), so as to obtain the corresponding scores matrices $\mathbf{T_p}$. This is just a preliminary data compression to filter out from each block noise or other unwanted sources of variability, so to stabilize the successive search for common and distinct components: all the successive steps are carried out on the block scores $\mathbf{T_p}$ instead of the full data matrices $\mathbf{X_p}$.
3. GCA is performed on all possible subsets of blocks to identify globally and locally common components ($\mathbf{T_C}$, $\boldsymbol{T_{Lk}}$) as those linear combinations of the block scores with high correlation (close to 1). This is an iterative process which starts with extracting common components at the upper level (i.e., among all blocks), and continues by examining all smaller subsets (in decreasing order of number of matrices involved) to extract locally common components. Prior to each GCA sub-step, all the individual block scores $\mathbf{T_p}$ are orthogonalized with respect to the common components extracted at the previous sub-steps. In detail the following steps are involved:
   a. An ordered list of all the possible subsets k of the blocks is defined ($k = 1$:K). Usually, one starts with a subset including all the blocks, then with subsets including all the blocks but one and so on. For instance, in a case with four blocks, there would be 10 subsets ordered as follows: 1 2 3 4; 1 2 3; 1 2 4; 1 3 4; 1 2; 1 3; 1 4; 2 3, 2 4; 3 4.
   b. GCA is performed on the scores $\mathbf{T_p}$ of all the blocks ($k = 1$) to identify the globally common components $\mathbf{T_C}$ as those linear combinations of the block scores with high correlation (usually above 0.90). To allow such components to be also predictive the common components are considered for the successive modeling step only if they explain at least a certain amount of X- and Y variance (usually 5%).
   c. If the number of extracted common components is greater than zero, then the common scores are saved and the individual block scores are orthogonalized with respect to the globally common component, to obtain the orthogonalized block scores $\mathbf{T_{po}}$: $\boldsymbol{T_{po}} = \boldsymbol{T_p} - \boldsymbol{T_C}\boldsymbol{T_C^T}\boldsymbol{X_p}$.
   d. GCA is performed on the orthogonalized scores $\mathbf{T_{po}}$ of a subset k of the blocks ($k = 2$:K) to identify the locally common components $\mathbf{T_{Lk}}$ as those linear combinations of the block scores with high correlation (usually above 0.90). To allow such components to be also predictive, the common components are considered for the successive modeling step only if they explain at least a certain amount of X- and Y variance (usually 5%).
   e. If the number of extracted common components is larger than zero, then the locally common scores are saved and all the individual block scores are orthogonalized with respect to the locally common component, to obtain the orthogonalized block scores $\mathbf{T_{po}}$: $\boldsymbol{T_{po}} = \boldsymbol{T_{po}} - \boldsymbol{T_{Lk}}\boldsymbol{T_{Lk}^T}\boldsymbol{X_{po}}$.
   f. Steps (d)-(e) are iterated until all the K subsets of blocks have been investigated.
4. For each block, a PLS regression model is calculated between **Y** and the orthogonalized scores $\mathbf{T_{po}}$, and the scores of the corresponding model $\mathbf{T_{Up}}$ are the distinct scores for that block. Indeed, the orthogonalization steps in (3c) and (3e) have removed from the original block scores $\mathbf{T_p}$ all the contributions from the global and local common components, leaving only the part of information which is unique to that particular block. This further PLS modeling step allows to retain only that part of the unique information which is relevant to predict the response(s).
5. The final model is built by running an ordinary least squares regression between the concatenated scores matrix $\mathbf{T_{all}} = [\mathbf{T_C}, \mathbf{T_{L2}}, \ldots, \mathbf{T_{Lk}}, \mathbf{T_{U1}}, \ldots \mathbf{T_{Up}}]$ and the **Y**: $\mathbf{Y} = \mathbf{T_{all}}\boldsymbol{\beta}$, $\boldsymbol{\beta}$ being the regression coefficients.

As described above, development of a PORTO model requires multiple selection steps, to optimize the initial number of LVs for each block and to identify the most appropriate number of common and distinct

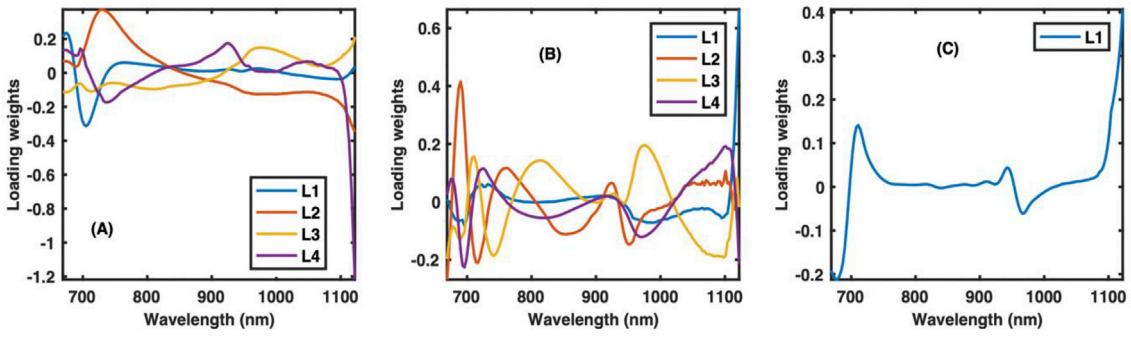


Fig. 6. Loading weights of the LVs extracted by SPORT for the olive data set. (A) MSC, (B), SNV, and (C) 2nd derivative.

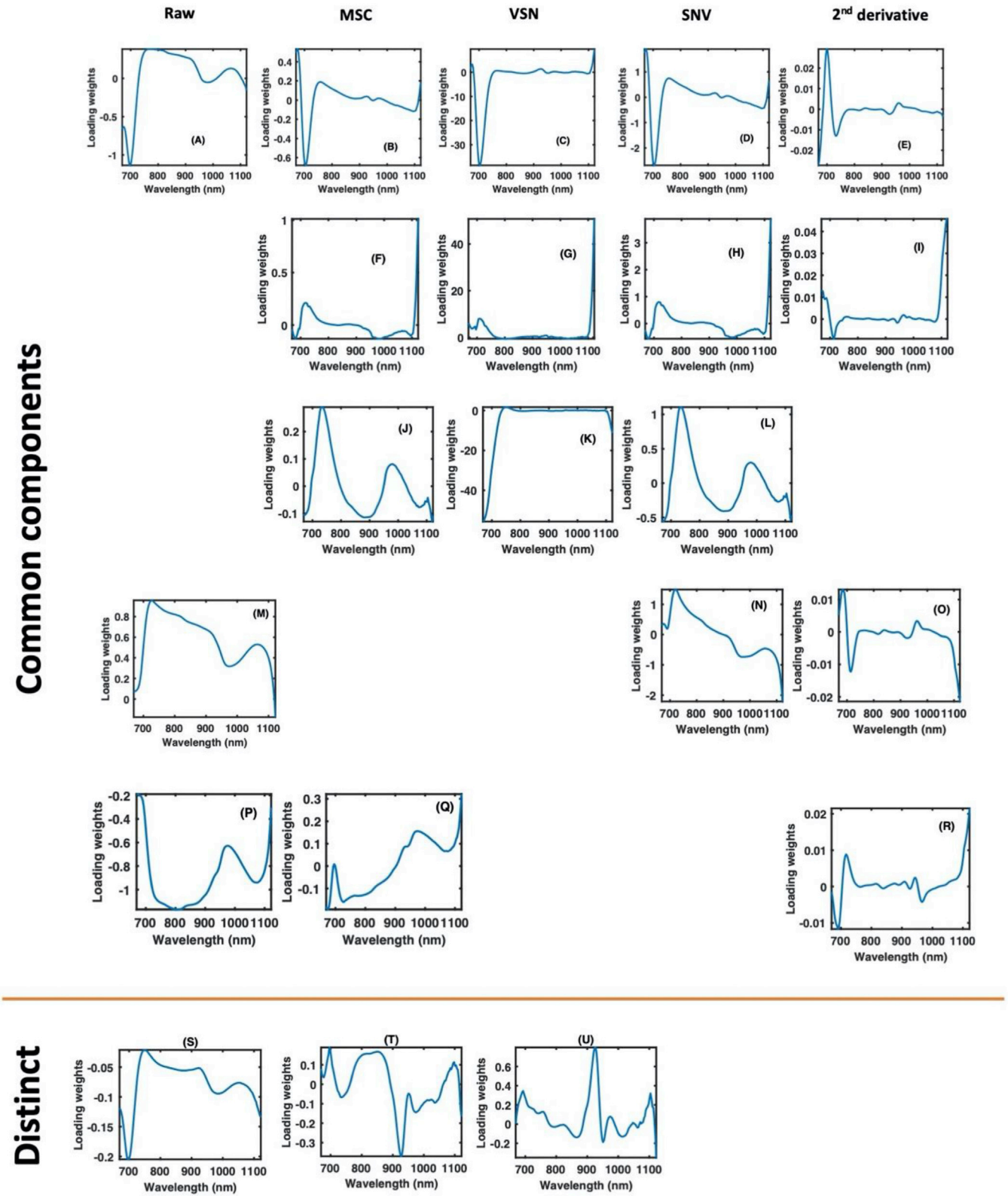


Fig. 7. Loading weights of the common and distinct components extracted by PORTO for the olive data set. Common component 1 (A–E), Common component 2 (F–I), Common component 3 (J–L), Common component 4 (M–O), Common component 5 (P–R), Distinct component corresponding to Raw (S), Distinct component corresponding to MSC (T), and Distinct component corresponding to VSN (U).

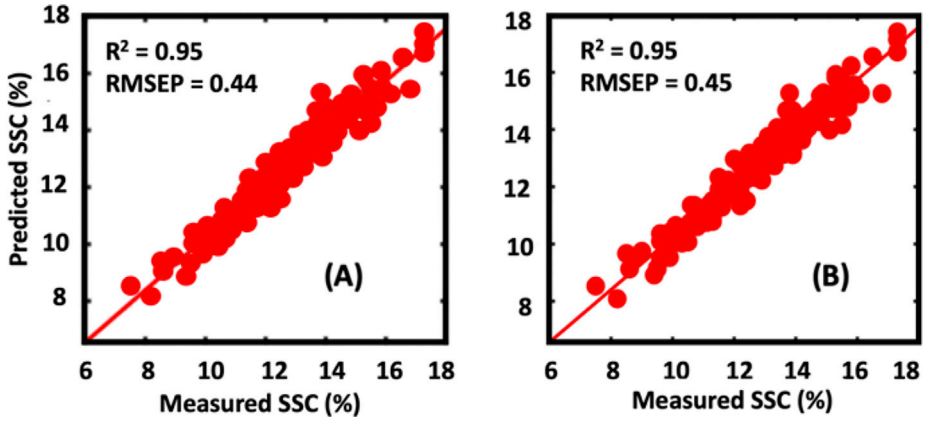


Fig. 8. PORTO model before and after pre-processing selection. (A) Preprocessings used: Raw+MSC+VSN+SNV+2nd derivative, and (B) Raw+MSC+2nd derivative.

components; to this purpose, several local cross-validations (CV) are performed in sequence, as discussed in Ref. [29] (Steps 2 and 4) or threshold-based criteria are adopted (Step 3). Moreover, since all the orthogonalization steps are extended also to the blocks not directly involved in the calculation of the common and distinct components, the final set of components are all mutually orthogonal, irrespective of their nature. The functions for calculating and validating the PORTO approach were developed in MATLAB (Release 2017b; The Mathworks, Natick, MA) by the authors but rely on the multi-block data analysis codes from NOFIMA [30] for the implementation of PO-PLS.

4. Results

4.1. Data description

The reference values for each data set are shown as histograms in Fig. 3. The calibration set is shown in blue and the test set is shown in brown. In all cases, the distribution of both the calibration and test sets were almost normal. Such normal distributions of fruit properties are often obtained in the field of fresh fruit analysis where most of the fruits have average properties and a few fruits have either lower or higher quality. Fig. 4 shows the calibration (Red) and test (Blue) spectra of the different fruit samples used in the study. In the case of apples (Fig. 4A), the bands in the range of 500–670 nm are related to the pigment composition of the skin of the fruits. The region above 670 nm corresponds to the 3rd overtones of C–H, O–H and N–H bonds and is widely used in the analysis of fruit products to predict Brix and DM content [5,6]. In the spectra of all the fruits (apples, olives, mangoes and pears), a difference in the global spectrum intensities can be observed. Such a difference indicates the presence of scattering effects (additive and/or multiplicative) due to the interaction of light with the complex structure of the fruit.

4.2. Data modelling

The results of regression modelling on all the 4 data sets are displayed in Fig. 5. In particular, the outcomes of the sequential pre-processing fusion done using SPORT are shown in panels A–D, whereas those of the parallel pre-processing fusion done using PORTO are reported in panels E–H. It can be noted that in all the cases, the PORTO models showed slightly better performances compared to modelling with the SPORT approach. In the case of the apple data set, for the PORTO approach the R^2_p was the same but the RMSEP decreased by 4%, compared to the SPORT approach. In the case of the olive data set, the PORTO approach increased the R^2_p by 4.5% and decreased the RMSEP by 16%, compared to the SPORT approach. In the case of the mango data set, the PORTO approach increased the R^2_p by 1.2% and decreased the RMSEP by 1% compared to the SPORT approach. In the case of the pear data set, the PORTO approach increased the R^2_p by 1.1% and decreased the RMSEP by 4% compared to the SPORT approach. For all four data sets, the differences in the predictive performances between PORTO and SPORT were statistically significant based on the outcomes of CV-ANOVA [25] (Apple: $p = 0.002$; Olive: $p = 0.047$; Mango: $p = 0.014$; Pear: $p = 0.007$). A summary of the performances of the PLS models built on data individually pre-treated with each of the pre-processings considered for the fusion as well as those of the SPORT and PORTO approaches is presented in Table 2. The PORTO fusion approach performed better than any individual pre-processing for all the cases considered.

4.3. Number of latent variables SPORT vs PORTO

A summary of the number of LVs extracted by SPORT and PORTO is shown in Tables 3 and 4, respectively. In the case of SPORT, distinct orthogonal LVs were extracted from the differently pre-processed blocks, whereas, in the case of PORTO, the common and distinct LVs were extracted from the whole set of preprocessed matrices or the individual blocks, respectively.

An example of the LVs extracted by the SPORT and PORTO approaches for the olive data set are shown in Figs. 6 and 7, respectively. The SPORT approach selected 4, 4 and 1 LVs for MSC, SNV and 2nd derivative pre-processing, respectively. In the case of PORTO, 5 LVs were selected as common and 3 LVs were selected as distinct (Fig. 7): it can be seen that the common components that are extracted from the differently pre-processed data often have similar profiles or similar bands (except for 2nd derivative which has different shape because 2nd derivative operation affects the overall profile of the spectra). Especially, the loadings of the SNV and MSC blocks for the common components 1, 2 and 3 have very similar profiles: this observation is consistent with what is reported in the literature, where it is indicated that SNV and MSC perform very similarly on sufficiently large datasets [31,32]. In this context, the possibility of identifying relations (similarities, such as in this case, but also dissimilarities) between different pre-processings is one of the advantages of PORTO. For example, the following comments can be put forward for the common components. The first common component (Fig. 7A–E) shows a very sharp peak at 700 nm, which is related to the absorption of chlorophyll. It should be noted that, although chlorophyll absorbs towards 680 nm, this peak at 700 nm actually corresponds to the red edge that is observed in all chlorophyllin plants, which expresses the slope between the chlorophyll absorption zone and the near-infrared reflection plateau, caused by tissue scattering [33]. The loading weights for the second common component (Fig. 7F–I) show on the one hand a sinusoidal profile around 700 nm, related to the variation in position of the red edge, and on the other hand the beginning of a peak, around 1150 nm, which certainly corresponds to the fat absorption, around 1208 nm [34]. The third common component (Fig. 7J–L) seems to be related to variations in light scattering, caused by the size of olive

cells. For MSC and SNV, the component is related to the overall shape of the spectrum; a peak, positive or negative, is found at the location of the steepest areas of the spectra (Fig. 4C). For VSN, which performs a much better correction of the spectra, the weights are very close to 0 everywhere, except in the red edge area. The fourth and fifth components are a little more difficult to interpret; it seems that they are related to the absorption of water at 960 nm [34] and to various interaction phenomena. All these findings are consistent with the fact that the DM of olives is very much related to their maturity. Indeed, during olive ripening, the chlorophyll regresses, the proportions of water and oil change and the cell size changes [35].

4.4. Pre-processing selection: an example on the apple data set

The PORTO approach supports pre-processing selection by deselecting some of the preprocessings if they are redundant in both common and distinct components. As an example, in the case of the apple data set (Table 3), the distinct LVs were identified as corresponding to the raw data, MSC and 2nd derivative pre-processing techniques. In the case of the common components, all of them either have 2nd derivative preprocessing technique or raw data linked to the common information. Therefore, the SNV and the VSN pre-processings are not needed as the common information is already present in the 2nd derivative preprocessed data. Hence, SNV and VSN can be deselected. The model before (Fig. 8A) and after deselecting (Fig. 8B) SNV and VSN are shown in Fig. 8. It can be noticed that the performances before and after deselecting SNV and VSN are practically identical.

5. Discussion

NIR spectra of fruit are affected by phenomena such as light scattering and the efficient removal of the effect of this scattering can improve the data modelling. Different scatter correction techniques can reduce/remove scattering to a certain extent. However, a single technique is usually not sufficient to remove all the effects, so a fusion of multiple techniques can often perform better. This is because different scatter correction techniques can highlight complementary information [36]. In the present work, modelling with the SPORT and PORTO approaches also showed that the optimal models were based on the information (LVs) corresponding to multiple pre-processing techniques used to pre-process the same data.

In the case of SPORT, the LVs are orthogonal, selected sequentially and represent distinct information. Whereas, the PORTO method first extracts the common LVs and later, once all the common information is extracted, the remaining distinct information is extracted from each differently pre-processed block. As in SO-PLS [18], the performances of SPORT are dependent on the order of the pre-processing blocks as it performs a sequential extraction of LVs where the blocks (pre-processing techniques) of high importance should be included first [14]. However, PORTO does not have such a disadvantage as it processes all the matrices in parallel, so that the relevance of the individual blocks does not matter as it doesn't affect the outcomes. Not needing to choose a priori the pre-processing order is the main advantage of PORTO over the SPORT approach. By extracting the common and distinct information, the PORTO approach was able to achieve slightly better performances compared to SPORT on all the data sets presented in this study. The better performances of PORTO could be the fine tuning of the model by extraction of the common and distinct information, where extraction of the common information was guided by multiple pre-processing techniques all together. However, when the pre-processing order for the olive dataset was changed, the SPORT approach achieved similar results compared to the PORTO approach. This indicates the SPORT approach is not inefficient; rather, based on the "correct" block order it can perform as well as the PORTO approach.

Pre-processing selection is also feasible with the PORTO approach: indeed, if a certain pre-processing does not contain distinct components, the corresponding pre-processing can be deselected, thus reducing the computational load for future analysis. One other challenge frequently encountered with the use of different pre-processings is related to the interpretation of the models. Often the different mathematical transformations make the chemical interpretation of the models based on the model coefficients not straightforward [37]. In this regard, the PORTO approach does not provide a clear solution but can support in gaining better insight into how the data are affected by different pre-processings. The common components of the PORTO approach can be used to explore how the same information is represented by differently shaped LVs extracted from differently preprocessed data, thus, giving a detailed insight into the associated mathematical transform. For example, in Fig. 7 the 5th common component (5th row of Fig.) has a very different shape for the LVs corresponding to RAW and MSC preprocessed data, compared to the 2nd derivative preprocessed data, but all three components explain the same information.

Another advantage of PORTO can also be understood as a means of selecting between the combinations of multiple preprocessing. Often multivariate data generated from analytical sensors suffers from a range of effects and several preprocessing operations are required to correct the data, such as smoothing + scatter correction +2nd derivative for revealing underlying peaks. Often, several combinations of pretreatments (and, within each technique, of parameters/meta-parameters) need to be tested to select the best option. To this purpose, approaches that rely on DoE-based exploration of preprocessing techniques are available [11] but they are based on defining a specific order of the operations and the selection of at most one technique for each family of preprocessing strategies. On the other hand, the common and distinct information extracted by PORTO can also be used to select specific combinations of preprocessing operations in a more versatile way. Indeed, the user can just remove the preprocessing combinations which do not carry unique information, similar to what was demonstrated with the apple data set in the present study (Fig. 8).

6. Conclusion

In the present work, PORTO, a new approach for parallel orthogonalized synergistic fusion of pre-processing strategies was presented. The results showed that the performances of the PORTO approach were slightly better than those of SPORT, a competing approach for pre-processing fusion. The improvement in R^2_p and decrease in RMSEP were up to 4.5% and 16%, respectively. Furthermore, the key benefit of the PORTO approach is to explore the common and distinct information related to the use of the different pre-processing techniques. The presence of distinct information underlines the distinct variability highlighted by each pre-processing technique. Also, pre-processing techniques can be deselected if they only explain the common information but do not carry any distinct information. Moreover, it has been shown [14] that the results of the SPORT approach may be dependent on the pre-processing order, especially as far as the selected pre-processings and, in general, the number of latent variables extracted from each block are concerned; on the other hand, the order of the pre-processings has been shown not to influence relevantly the predictive performances. Unlike SPORT, the PORTO approach is almost unaffected by the pre-processing order as, in most of the steps of the algorithms, blocks corresponding to different pre-processings are all processed in parallel. The PORTO approach is not limited to spectral data but can also be used in the case of any data where pre-processing is required. This study does not conclude that the SPORT approach is inefficient, but rather it stresses that, in SPORT, there is always the need to decide on the order of the pre-processing, which for some data sets may not be trivial; however, if the correct order is chosen it can provide similar performances compared to PORTO. It is also worth stressing that, although, for the sake of an easier presentation of the algorithm and a more consistent discussion of the results in this paper, the PORTO approach has only been discussed in the regression context, it can also be used to deal with classification

problems. In particular, to extend the PORTO approach to do this, it is only necessary to introduce a dummy coding of the response, just as in PLS-DA [33] and substituting the final ordinary least squares model (step 5 of the PO-PLS algorithm in Section Parallel pre-processing through orthogonalization) by a linear discriminant analysis (LDA) on the concatenated scores.

Lastly, it is fundamental to point out that, being a supervised approach, PORTO could suffer from the risk of overfitting: to minimize such a risk, it therefore essential that all modeling aspects (predictions, interpretation, coefficients) be properly validated [34].

Author statement

Puneet Mishra: Conceptualization, Data curation, Software, Investigation. **Jean Michel Roger:** Writing - review & editing, Software. **Federico Marini:** Formal analysis, Software, Visualization. **Alessandra Biancolillo**: Writing - review & editing, Software, **Douglas N. Rutledge:** Writing - review & editing.

Declaration of competing interest

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.

References

[1] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, Pre-processing Methods. In: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3., Elsevier, Oxford, UK, 2020, pp. 1–75.
[2] 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 Anal. Chem. 50 (2013) 96–106.
[3] Å. Rinnan, F.v.d. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trac. Trends Anal. Chem. 28 (2009) 1201–1222.
[4] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36.
[5] B.M. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review, Postharvest Biol. Technol. 46 (2007) 99–118.
[6] R.F. Lu, R. Van Beers, W. Saeys, C.Y. Li, H.Y. Cen, Measurement of optical properties of fruits and vegetables: a review, Postharvest Biol. Technol. 159 (2020).
[7] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[8] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: variable sorting for normalization, J. Chemometr. 34 (2020) e3164.
[9] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (1988) 1273–1284.
[10] H. Martens, J.P. Nielsen, S.B. Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures, Anal. Chem. 75 (2003) 394–404.
[11] J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den Heuvel, L.M.C. Buydens, Simple and effective way for data preprocessing selection based on design of experiments, Anal. Chem. 87 (2015) 12096–12103.
[12] L. Xu, Y.-P. Zhou, L.-J. Tang, H.-L. Wu, J.-H. Jiang, G.-L. Shen, R.-Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta 616 (2008) 138–143.
[13] X. Bian, K. Wang, E. Tan, P. Diwu, F. Zhang, Y. Guo, A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples, Chemometr. Intell. Lab. Syst. 197 (2020), 103916.
[14] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometr. Intell. Lab. Syst. 199 (2020), 103975.
[15] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, Trac. Trends Anal. Chem. (2020), 116045.
[16] P. Mishra, F. Marini, A. Biancolillo, J.-M. Roger, Improved Prediction of Fuel Properties with Near-Infrared Spectroscopy Using a Complementary Sequential Fusion of Scatter Correction Techniques, Talanta, 2020, 121693.
[17] P. Mishra, A. Nordon, J.-M. Roger, Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques, J. Pharmaceut. Biomed. Anal. (2020), 113684.
[18] A. Biancolillo, T. Næs, The sequential and orthogonalized PLS regression for multiblock 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.
[19] 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, J. Chemometr. 31 (2017), e2900.
[20] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemometr. 33 (2019), e3085.
[21] Y. Song, J.A. Westerhuis, A.K. Smilde, Separating common (global and local) and distinct variation in multiple mixed types data sets, J. Chemometr. 34 (2020), e3197.
[22] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[23] J.-M. Roger, F. Chauchard, V. Bellon-Maurel, EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits, Chemometr. Intell. Lab. 66 (2003) 191–204.
[24] X. Sun, P. Subedi, R. Walker, K.B. Walsh, NIRS prediction of dry matter content of single olive fruit with consideration of variable sorting for normalisation pre-treatment, Postharvest Biol. Technol. 163 (2020), 111140.
[25] P. Mishra, E. Woltering, B. Brouwer, E. Hogeveen-van Echtelt, Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach, Postharvest Biol. Technol. 171 (2021), 111348.
[26] P. Mishra, F. Marini, B. Brouwer, J.M. Roger, A. Biancolillo, E. Woltering, E.H.-v. Echtelt, Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit, Talanta 223 (2021), 121733.
[27] X.D. Sun, P. Subedi, K.B. Walsh, Achieving robustness to temperature change of a NIRS-PLSR model for intact mango fruit dry matter content, Postharvest Biol. Technol. (2020) 162.
[28] 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, 104139.
[29] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, Food Qual. Prefer. 24 (2012) 8–16.
[30] 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, Chemometr. Intell. Lab. 124 (2013) 32–42.
[31] I.S. Helland, T. Næs, T. Isaksson, Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data, Chemometr. Intell. Lab. Syst. 29 (1995) 233–241.
[32] M.S. Dhanoa, S.J. Lister, R. Sanderson, R.J. Barnes, The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra, J. Near Infrared Spectrosc. 2 (1994) 43–47.
[33] D.N.H. Horler, M. Dockray, J. Barber, The red edge of plant leaf reflectance, Int. J. Rem. Sens. 4 (1983) 273–288.
[34] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, Longman scientific and technical, 1993.
[35] D.J. Connor, E. Fereres, The physiology of adaptation and yield expression in olive, Hortic. Rev. 31 (2005) 155–229.
[36] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials, Postharvest Biol. Technol. 168 (2020), 111271.
[37] P. Oliveri, C. Malegori, R. Simonetti, M. Casale, The impact of signal pre-processing on the final interpretation of analytical outcomes - a tutorial, Anal. Chim. Acta 1058 (2019) 9–17.