New data preprocessing trends based on ensemble of multiple preprocessing techniques

Puneet Mishra, Alessandra Biancolillo, Jean-Michel Roger, Federico Marini, Douglas N Rutledge

Trends in Analytical Chemistry 132 (2020) 116045

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

New data preprocessing trends based on ensemble of multiple preprocessing techniques

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

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

ARTICLE INFO

Article history:
Available online 19 September 2020

Keywords:
Multivariate calibration
Ensemble learning
Multi-block analysis
Chemometrics
Preprocessing

ABSTRACT

Data generated by analytical instruments, such as spectrometers, may contain unwanted variation due to measurement mode, sample state and other external physical, chemical and environmental factors. Preprocessing is required so that the property of interest can be predicted correctly. Different correction methods may remove specific types of artefacts while still leaving some effects behind. Using multiple preprocessing in a complementary way can remove the artefacts that would be left behind by using only one technique. This article summarizes the recent developments in new data preprocessing strategies and specifically reviews the emerging ensemble approaches to preprocessing fusion in chemometrics. A demonstration case is also presented. In summary, ensemble preprocessing allows the selection of several techniques and their combinations that, in a complementary way, lead to improved models. Ensemble approaches are not limited to spectral data but can be used in all cases where preprocessing is needed and identification of a single best option is not easily done.

1. Introduction

In analytical chemistry, multivariate data are generally acquired from different measurement techniques for qualitative and quantitative analysis [1,2]. The techniques can range from implementing miniature near-infrared (NIR) spectrometers [3,4] for measuring optical properties to high end complex and costly techniques such as liquid or gas chromatography - mass spectrometry (LC-MS or GC-MS) or even up to their multidimensional implementations (e.g., GCxGC/MS) [5,6]. As a consequence, the data generated by these techniques can range from simple multivariate spectral data (as in the case of NIR) to multi-mode chromatograms (in the case of LC/GC-MS). However, what the data from multiple techniques have in common is that they all suffer from artefacts (unwanted variation) [7]. The presence of these artefacts can be ascribed to several causes, such as the measurement modality, instrumental drifts, sample state and other external physical, chemical and environmental factors [8]. As an example, one could think of NIR measurements performed in diffuse reflection mode on highly scattering materials [9]. In such a case, the acquired NIR spectra will be a combination of absorption and scattering characteristics. The absorption phenomenon will translate to the presence of peaks at specific wavelengths whereas the scattering will manifest itself as additive effects, mainly modifying the baseline, and multiplicative effects which dramatically affect the linear models. Additionally, when the collected spectra are used for calibration, the attenuation in the signals brought by scattering may also affect model quality, not only in terms of predictive accuracy but also of interpretability [10]. As a consequence, it is obvious that, under the above-mentioned circumstances, one should always perform scatter correction in order for the spectra to resemble as closely as possible the absorption profile so that the corresponding models are based on such information only [10,11].

Data preprocessing, in general, is required to correct for scattering, baselines changes, peak shifts, noises, missing values and several other artefacts so that the "true" chemically-relevant

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

https://doi.org/10.1016/j.trac.2020.116045
0165-9936/

Abbreviations

AsLS	Asymmetric Least Square
DoE	Design of Experiments
GAPLS	Genetic Algorithm Partial Least Squares
KDR	Known Data Regression
LC/GC-MS	Liquid Chromatography/Gas Chromatography Mass Spectrometry
MCR	Multivariate Curve Resolution
MSC	Multiplicative Scatter Correction
NIR	Near Infrared
NMR	Nuclear Magnetic Resonance
OPLS	Orthogonal Partial Least Squares
PAT	Process Analytical Technologies
PCR	Principal Component Regression
PLSR	Partial Least Square Regression
RNV	Robust Normal Variate
SNR	Signal to Noise
SNV	Standard Normal Variate
SPORT	Sequential Preprocessing Through Orthogonalization
SS-DAC	Soft Sensor Development, Assessment and Comparison
TSR	Trimmed Scores Regression
VSN	Variable Sorting for Normalization

underlying structure can be highlighted and/or, if required, the property of interest can be predicted correctly [7,10,12]. Thanks to the extensive chemometric researches in the last decades, several preprocessing techniques are now available [8]. However, there are no clear rules to decide when to use a specific preprocessing technique, if a single technique could be enough and, if not, which techniques to combine and how (e.g., in what order) [7,13]. For instance, several techniques are available to perform scatter correction, such as standard normal variate (SNV), multiplicative scatter correction (MSC) and their several variants, calculation of 2nd derivative (which, depending on the algorithm, requires setting additional parameters), or the recently introduced variable sorting for normalization (VSN) [8]. Accordingly, to identify which technique could be the best for her/his data, the user is often required to explore all the available options. Moreover, if there is the need to associate scatter correction with other operations, e.g., noise removal and scaling, the number of combinations to be explored increases exponentially. To overcome this problem, several approaches related to the identification of the best preprocessing and/or of their optimal combination were also developed [14–17]. In this context, until recently, most of the attention in chemometrics was limited to finding the best preprocessing [15] or their optimal combination [14] and almost minimal focus was given to understanding what complementary information the different preprocessing techniques carry.

Preprocessing methods always carry with them the risk of also removing relevant chemical information or variation that is related to the property of interest [7,13,14]. On the other hand, different preprocessing techniques may remove certain types of artefacts while still leaving some other effects behind. Using multiple preprocessing in a complementary way can remove the artefacts that would be left behind by using only one technique. Thanks to recent developments in ensemble [18,19] and data fusion methods [20], the complementary fusion of multiple preprocessing techniques is now possible. The complementary fusion of preprocessing techniques and their combinations has two major benefits: the first one is that it allows all the complementary information linked to different preprocessing techniques and their combinations to be used synergistically to develop models, and the second one is that it takes the user out of the loop of searching for the best preprocessing and their combinations. Several recent works have shown that the complementary fusion of preprocessing techniques can lead to high-quality models [19–21].

This article summarizes the recent developments in new data preprocessing strategies and specifically reviews the emerging ensemble approaches to preprocessing fusion in chemometrics. A demonstration case of using an ensemble preprocessing fusion approach called sequential preprocessing through orthogonalization (SPORT) is also presented.

2. Artefacts in data: background causes

Several artefacts may be present in multivariate data obtained from analytical techniques. Broadly, they can be classified in five main categories, i.e., missing data, noise, baselines shifts, multiplicative effects and peak shifts. Each artefact has its own background cause(s), which, in the case of spectroscopy, can range from a human error during measurements to a complex interaction of light with the physical structure of the sample. In the remainder of this Section, the various typologies of artefacts and their causes will be briefly discussed.

Missing data refers to the lack of one or more entries in the matrix containing the experimental data. The presence of missing data can be for several reasons, such as values outside the instrument range, instrument malfunctioning at a certain point of time, communication failure between the instrument and the digital controller, multiple sensors deployed for the same task but at different sampling rates, saturation of the signal intensity and instrumental errors during data acquisition [22]. Apart from these technical reasons, sometimes the missing values are due to human errors, and arise in cases when repeating a part of the measurements may be impossible or too expensive [23]. In Fig. 1, the occurrence of missing data is represented by the red dotted line, which, compared to the original spectrum (solid blue line), lacks certain spectral bands.

Noise is the unwanted disturbances in a signal [24] and its causes can be traced back to, e.g., the sensitivity of the detectors of instruments or, in the case of optical spectroscopy, the excitation source (electromagnetic light source). The presence of noise is usually expressed as low signal-to-noise (SNR) ratio. Most often the noise is visually identifiable by plotting the dataset. An example of noisy signal is shown in Fig. 1 as a solid purple line: compared to the original signal (solid blue), which does not present any noise, the perturbed spectrum is characterized by a disturbance at the extreme wavelengths. Such a type of noise is commonly encountered in NIR data due to the low detectivity of the sensors at the extreme frequencies.

Variables with huge intensity differences may be encountered in several analytical platforms such as mass spectrometry and NMR. However, such huge differences in signal intensities are rarely encountered in the domain of optical spectroscopy. These are not really artefacts but actually intrinsic characteristics of such data sets. In such cases, the aim is always to transform/normalize the variables so to span a comparable range of variability.

Baselines are structured background effects (“continuous” and low frequency) and the reasons for their occurrence in the signals are, in general, dependent on the specific measurement technique involved. For example, in Raman spectroscopy, the main cause of a non-zero baseline is the background fluorescence of the samples, whereas in NIR, the baselines are caused by stray light and the interaction of the light with the particles or droplets, leading to scattering; on the other hand, in chromatography, the baseline corresponds to the detector response when only the mobile phase emerges from the column. The presence of a linear baseline is exemplified as a solid yellow line in Fig. 1, where it is evident that such an effect results in an increasing difference in intensity with respect to the original signal (solid blue).

Multiplicative effects are artefacts whose extent depends linearly on the intensity of the original signal. They may be caused, among others, by physical phenomena, such as light scattering in the case of optical spectroscopy, or by a non-completely reproducible sample manipulation or presentation (e.g., volume of injected sample in chromatography or dilution in nuclear magnetic resonance). In the case of optical techniques, multiplicative effects are caused by the forced deviation of the photons from a straight trajectory by localized non-uniformities present in the samples. In chromatography, multiplicative effects are related to the very slight variations in the amount of sample being injected and analyzed in the different runs [25]. In NMR-based metabolomics, especially in the analysis of urine, multiplicative effects result from unspecific variations of the overall concentrations of samples, due to different dilutions. In Fig. 1, the dashed green line represents a signal with multiplicative effects: when compared to the original profile (solid blue line), it is evident that the differences in intensity are more pronounced for those wavelengths where the unperturbed spectrum has a higher signal.

Peak shifts are horizontal displacements in the signals which, otherwise, should ideally be aligned (i.e., present maxima at the same frequencies, in case of spectroscopy, or retention times, in the case of chromatography). Accordingly, there are two main types of shifts, namely temporal and spectral shifts. Temporal shifts are common in chromatography where they can be due to the deterioration or aging of the stationary phase, matrix effects, temperature changes, changes in the mobile phase composition, instrumental drift, interactions between analytes, fluctuations in pressure and flow rates and presence of gas bubbles [26]. Spectral shifts are common in techniques like optical spectroscopy, NMR and mass spectrometry, where they can be due to variations in temperature, pressure, viscosity and pH, but also for instrumental reasons [26]. An example of peak misalignment is shown in Fig. 1 where the solid cyan line is the shifted version of the original spectrum (solid blue). When such effect is present, peak alignment is fundamental for optimal modelling.

3. Objectives of data preprocessing

The global objective of data preprocessing is to remove the unwanted variability or effects from the signal so that the useful information related to the property(-ies) of interest can be used for efficient modelling. The specific objectives of the preprocessing techniques are dependent on the type of artefacts to be dealt with. Following the same classification reported in the previous section, in the case of missing data, the objective of preprocessing is to estimate the missing values using data imputation approaches, a detailed description of which can be found in a dedicated review [23]. In the case of noisy data, the objective is to remove/reduce the noise, and this can be achieved through several approaches such as manual removal of the noisy part of the signal or time series, use of filtering algorithms [27] or smoothing functions such as polynomial fitting and spline interpolation [28], and data reconstruction approaches such as principal/independent component reconstruction [8] or wavelet threshold and reconstruction [29]. In the case of the presence of a non-zero baseline, the objective is to identify the type/order of the baseline, estimate it accordingly, and then subtract it from the overall signal. Baseline modelling/removal can be achieved with approaches such as offset corrections, detrending, asymmetric least square (AsLS) and several others [8]. In the case of multiplicative effects, the objective is to remove/reduce the global intensity differences in the signal either by the use of model-free approaches, such as normalization techniques [8], or with model-

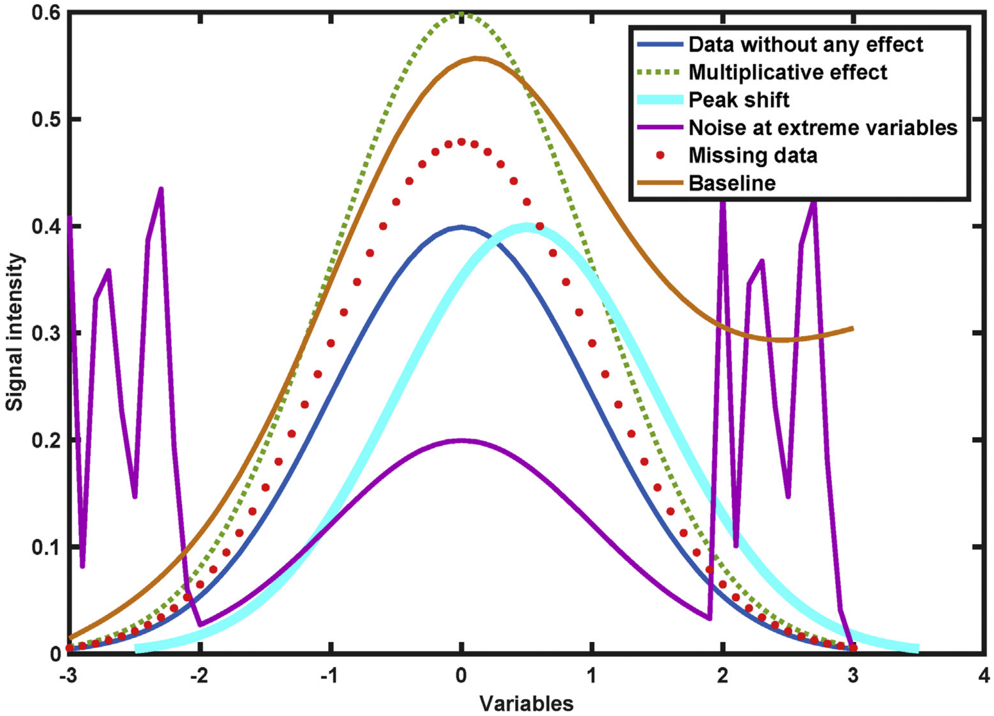

Fig. 1. Example of several simulated artefacts for visible and near-infrared data. Data with no unwanted effect (solid blue line), multiplicative effects (dashed green line), peak shifts (solid cyan line), noise at extreme bands (solid purple line) and missing data (dotted red line), and baseline shift (solid brown line).

based strategies which use a reference signal to estimate the multiplicative effects, in order to remove them by mathematical operations [30]. In the case of peak shifts, the objective is to align the peaks along either the spectral or the time domain. The basic steps of peak alignment are to identify a reference signal (signal with respect to which other signals are aligned) and to warp the experimental profiles along the time or spectral axis, so to be as matched as possible to the reference. In order for alignment to be effective, the target signal should present as many of the peaks present in the experimental profile as possible. Moreover, the extent of matching between the signals and the target can be limited to key peaks, which can be identified by means of approaches such as bucketing or binning and Landmark peak selection, or evaluated across the whole signal. Several approaches to perform peak alignment are available and can be found in a recent review [26]. A further detailed summary of all the most commonly used chemometric preprocessing techniques can be found elsewhere [8], together with the MATLAB codes for their implementation.

4. Recent advances in preprocessing methods

Preprocessing plays a fundamental role in chemometrics. However, there is no preprocessing technique which can be considered as a gold standard or, anyway, be blindly applied to any data, irrespective of their nature; on the contrary, preprocessing techniques are usually chosen based on the data source and the specific artefacts. A summary of the preprocessing techniques which, across the years, have become well established in chemometrics and are most frequently used, can be found elsewhere [8]. Alongside with these, new preprocessing techniques, which are summarized in Table 1, have been developed over the past decade. The details reported in Table 1 point out how new strategies have emerged for the correction of all the types of artefacts previously discussed, i.e. for missing data imputation, noise removal, baseline correction, multiplicative effect correction and peak alignment. As far as the imputation of missing data is concerned, new methods have been proposed, considering the problems of both model exploitation and model building in the presence of missing data. In particular, to deal with model building in the scenario of missing data, methods which combine known data regression (KDR) and trimmed scores regression (TSR) with principal component regression (PCR) or partial least squares regression (PLSR) have emerged [31]. A free toolbox to implement several novel strategies for missing data imputation can be found and accessed in Ref. [22].

In the case of noise removal, the new emerging trend is the use of autoencoders [32]. An autoencoder is a type of bottleneck artificial neural network that is used to learn efficient data coding in an unsupervised manner. Indeed, an autoencoder is a multi-layer feed-forward network, with at least three hidden layers, the intermediate of which, that is the one providing the mapping, being characterized by a rather small number of neurons. The network is trained using both as inputs and as targets the original signals, so that, through the bottleneck architecture, the autoencoder extracts non-linear salient features from the data, leaving the noise unmodeled. The encoded data can then be decoded to reconstruct the original signal but without noise [32]. Another advancement in noise removal is related to the possibility of automating the task: for this purpose, automatic threshold selection for shearlet coefficients was proposed in Ref. [29].

Baseline correction methods can broadly be divided into three subgroups, i.e., polynomial fitting, wavelet transformations and derivatives. The new direction to baseline correction involves combining wavelet decomposition, signal differentiation and baseline fitting with penalized least squares [33]. The new strategy combining multiple baseline correction approaches shows that the different techniques carry complementary information and their combination leads to improved removal of baselines from Raman spectra [33]. Some new developments were related to the advancement of adaptive reweighting strategies to support polynomial weighting and penalized least squares [34,35]. The main benefit of the adaptive reweighing approach is that it does not require user intervention and prior information. The multiple spectra baseline fitting approach is also new and uses multiple signals to infer their common characteristics in order to learn slowly varying baselines [36]. The main aim of the multiple spectra baseline correction approach is to learn baselines that perform well on the corresponding spectra and then co-regularize the choices by penalizing the disagreements among the baseline corrected spectra based on asymmetrical least squares [36]. Even if it doesn't involve the actual correction of the signals, it is worth mentioning that another approach based on Tikhonov regularization was recently proposed to deal with the presence of a non-constant baseline in the framework of regression model building [37]. In particular, the approach is based on using the Tikhonov regularization extra terms in the linear calibration problem to force the regression coefficients to be orthogonal to the direction(s) of the sample space spanned by the baseline effects [37].

As far as the correction of multiplicative effects is concerned, the major development was related to the introduction of the possibility of weighting the individual variables according to how likely they are to be affected by the unwanted contribution to be removed. In this context, a new method called variable sorting for normalization (VSN), which utilizes a random consensus strategy to estimates the weights before implementing normalizations in a data-driven, hypothesis-free fashion, was recently proposed [38]. Also are emerging local preprocessing approaches, which divide the signal into parts and apply locally (i.e. individually on each of the parts) existing preprocessing methods, such as SNV [39]. Other new methods, which combine different approaches to correct the multiplicative effects, have also been reported [40]. Robust variants of traditional preprocessing methods are also gaining importance due to their capability to deal with outliers, leading to optimized preprocessing of data [16]. For peak alignment, an increasing number of approaches are appearing in the literature, adopting the concept of point matching inspired by the computer vision domain [41,42]. Another emerging trend is the combination of data decomposition methods, such as multivariate curve resolution (MCR), with wrapping approaches [43,44]. Recent studies show that nowadays the focus is on automated methods which do not require reference signals and perform global alignments to deal with data rich in features [41,42,45–47].

5. Preprocessing selection approaches and their limitations

The choice of preprocessing techniques can greatly affect the model performance and preprocessing selection by the user can very often be sub-optimal as the user cannot directly explore all preprocessing techniques and their combinations. In the beginning, the selection of the preprocessing approach was based on trial and error. However, with the advancement in the computing power and the evolution of chemometrics, in recent years, several approaches for the selection and optimization of preprocessing strategies have been proposed (Table 2). For example, a novel approach based on orthogonal PLS (OPLS) was proposed to evaluate the performances of individual preprocessing techniques and their combination. The OPLS-based approach decomposes the data into common and distinct variation, allowing to evaluate which part of the original information is retained and which new patterns emerge in the data after a preprocessing technique is applied [15]. On the other hand,

Table 1
Summary of the strategies for preprocessing developed in the past decade in chemometrics.

Artefacts type	Techniques	Background principle	Novel features	References
Missing data	• Known data regression with principal components regression (KDR-PCR) or partial least squares regression (KDR-PLSR) • Trimmed scores regression with principal components regression (TDR-PCR) or partial least squares regression (TDR-PLSR)	• A particular case of generalized regression model • The regression model is fitted and the missing part is estimated • Loadings obtained from PCA or PLSR depending on the method implemented	• Allows model development and application in presence of missing data • Both data visualization and predictive analysis in presence of missing data is possible	[22,31]
	Maximum likelihood PCA based data imputation	• Assigns high variance to the missing values prior to PCA allowing to fit PCA model by disregarding the missing points	• Allows PCA in presence of missing values	[48]
Noise removal	Autoencoders	• Bottleneck neural network for the extraction of non-linear features (non-linear data compression)	• Automatic noise removal	[32]
	Shearlet-based denoising	• Automatic thresholding for shearlet coefficients • Reconstruction of signal after thresholding allows noise removal	• Automatic preprocessing • Suited for imaging spectroscopy data	[29]
Baselines correction	Continuous wavelet transform + peak width estimation by signal to noise ratio enhancing derivative + background fitting using penalized least squares with binary masks	• A three-step process involving accurate detection of peak position, followed by peak width estimation and lastly background fitting using penalized least squares using binary masks	• Does not require any preprocessing for transformation of spectra into the wavelet space • Combines wavelet transformation, derivatives and polynomial fitting	[33]
	Adaptive reweighing scheme for polynomial fitting and penalized least squares	• iteratively weights are changes for sum squares errors (SSE) between the fitted baseline and original signals, and the weights of the SSE are obtained adaptively using the difference between the previously fitted baseline and the original signals	• Automated approach to baseline correction	[35]
	Multispectra baseline correction with asymmetric least squares	• Uses multiple signals to infer the common characteristics to learn slowly varying baselines	• Fast and outputs multiple baselines simultaneously	[36]
	Tikhonov regularization	• Uses regularization to force the regression coefficients to be orthogonal to the directions of multiplicative effects	• Direct regression modelling by removing the influence of baselines	[37]
Multiplicative effects correction	Variable sorting for normalization (VSN)	• Estimates weights for variables based on severity of multiplicative effects	• Weights are estimated automatically • Weights can be incorporated into traditional normalization techniques such as SNV, MSC etc.	[38]
	Local standard normal variate (LSNV)	• Splits the signal into intervals and performs SNV separately for each local part	• Easy to implement as it just consists in the local application of a standard preprocessing	[39]
	First derivative + simple spectral ratio (FD-SR), Linear regression correction + simple spectral ratio (LRC-SR), Orthogonal spatial projection + simple spectral ratio (OPS-SR)	• Uses two steps • The first one corrects for additive effects using either first derivative, linear regression correction or orthogonal spatial projection • The second step uses simple spectral ratio for the correction of multiplicative effects	• Can estimate separate additive and multiplicative correction factors for each signal unlike tradition methods which use a general statistics such as mean and standard deviation for correction	[40] [40] [40]
	Group aggregating normalization (GAN)	• Uses group assignment to estimate the dilution factor to perform the normalization	• Developed specifically to deal with the multiplicative effects due to dilution	[49]
	Robust standardization, SNV, MSC, Detrending, offset correction	• Incorporates median and mean absolute deviation to traditional preprocessing techniques	• Removes the effect of outliers during the corrections	[16]
Peaks alignments	Multivariate curve resolution + correlation optimized warping (MCR + COW)	• At first, reduces complexity of the signals by curve resolution and later uses COW to align the profiles.	• Minimizes the risk of aligning non-corresponding information	[43] [44]
	Automatic time shift alignment (ATSA)	Three step process: • Automatic baseline correction and peak detection to provide useful chromatographic information • Preliminary alignment through adaptive segment partition to correct alignment for the entire chromatogram	• Suited for chromatographic data • Parameters such as peak information, segment size are automatically optimized	[50]

(continued on next page)

Table 1 (*continued*)

Artefacts type	Techniques	Background principle	Novel features	References
		• Precise alignment based on test chromatographic peak information to accurately align time shifts		
	Coherent point drift peak alignment	• Uses the point matching algorithm from computer vision domain	• Useful for 2D chromatography-mass spectrometry data • Global peak alignment • Useful for dense data, i.e., rich in features	[51]
	BiPACE 2D	• Introduces a similarity measure for comparing peaks based on the closeness of their barycenters in 2D TIC images and the angle or inner product between the binned MS profiles. • All possible pairwise similarities between peaks in two (or more) 2D chromatograms are estimated to identify best matching signals	• Useful for 2D chromatography- mass spectrometry data • Suitable for chromatographic data with increasing number of peaks • Does not require user defined reference spectrum	[47]
	Smith-Waterman peak alignment	• Combines the Smith-Waterman local alignment with mass spectral similarity	• Eliminates the need for detection of landmark peaks and usage of retention time transformation • Automated	[41]
	Optimal peak alignment with mixture similarity measure	• Uses mixture similarity by employing peak distance and the spectral similarity measures	• Does not require pre-defined window to operate	[46]
	Distance and spectrum correlation optimization alignment (DISCO)	• Landmark peaks of samples are mapped to landmark peaks of the reference using Euclidean distance to calculate similarity and correlation coefficient. • Later, local interpolation is applied to non-landmark peaks to correct distortion	• Allows on the fly alignment reducing memory usage	[45]
	Global peak alignment with point matching algorithm	• Extracts feature peaks in the signal and then searches globally the matching peaks	• Global alignment • Suitable for homogeneous as well as heterogeneous data	[42]

DoE-based selection of the optimal combination of preprocessing techniques, where the effect of different preprocessing techniques is explored by considering each family of methods as a factor, whose levels are the individual techniques, also gained attention [14,52]. In particular, the DoE approach presented in Refs. [14] considers preprocessing selection as a 4 stage process involving the sequential application of baseline correction, scatter correction, noise removal and scaling. However, by doing so, it does not allow more than one preprocessing inside each category to be explored. Grid search based approaches, which explore all possible combinations of preprocessing techniques to decide on the best, were also developed [17,53]. However, they are computationally expensive compared to the DoE-based approach which performs a pre-selection of preprocessing techniques based on the design [14]. Recently, a faster genetic algorithm partial least squares (GAPLS) approach was proposed for parallel comparison of the preprocessing techniques, in order to achieve an optimal selection [54]. The GAPLS based approach has the benefit that it can explore the preprocessing in parallel but has the disadvantage compared to the DoE based approach presented in Ref. [14] that it does not allow sequential selection and optimization of different preprocessing tasks such as baseline correction, followed by scatter correction, noise removal and scaling. Recently, a new process analytical technology (PAT) based soft sensor development, assessment and comparison (SS-DAC) framework was presented for optimal preprocessing selection [55]. To select the optimal preprocessing technique or combination, the SS-DAC framework performs an exhaustive pairwise statistical comparison to score the models corresponding to the different options tested. However, the main drawback with all these preprocessing selection and optimization approaches (Table 2) is that they all aim at the selection of techniques rather than using the complementary information present in each preprocessing technique. In many cases, multiple preprocessing techniques contain complementary information which can be used for synergistic modelling [18,20,21].

6. Ensemble approaches to preprocessing fusion

Preprocessing techniques can enhance the quality of analytical signals by removing artefacts. However, due to the limitations of the different techniques and the complexity of the artefacts, a single preprocessing technique may often not remove the artefacts completely whereas, if used in a complementary way, a combination of different preprocessings may be more effective. In the present section, three main approaches to ensemble preprocessing selection and fusion are presented.

6.1. Ensemble based on stacked regression

The stacked regression approach involves training multiple regression models with different preprocessings and then combining them with approaches such as cross validation or model averaging. Such a strategy was presented in Xu, Zhou et al. [18], where PLS models built on differently preprocessed data are combined by Monte Carlo cross validation (MCCV) stacked regression. The results show that the fusion of complementary information obtained by different preprocessing techniques often leads to a more stable and accurate calibration model. A summary of the stacked regression approach [18] is presented in Fig. 2. At first, some common preprocessing methods are used to transform the data. Secondly, optimized calibration models are built on differently preprocessed data. The third step is to combine the

Table 2
A summary of spectral preprocessing selection and optimization.

Selection approach	Background idea	Key features	References
OPLS	• Identities the joint and unique variation between raw and preprocessed data to evaluate the preprocessing technique	• Allows exploration of individual preprocessing techniques	[15]
DoE based approach	• Considers different artefact correction techniques as different factors and uses Design of Experiments (D-optimal, full factorial) to search for the optimal combination	• Based on the design, model performance of a few preprocessing methods and combination are evaluated	[14,52,56]
Regularized MANOVA-based preprocessing optimization	• Assesses the goodness of preprocessing techniques by using the ratio of the between-group to within group variance on the first canonical variate derived from regularized MANOVA • Grid search is used to find the best preprocessing strategy	• Useful for preprocessing selection in case of classification task	[53]
Soft sensor development, assessment and comparison (SS-DAC)	• Uses exhaustive statistical pairwise comparison of multiple models developed based on different preprocessings and their combination	• Uses the standardized process analytical technology framework for soft sensor development	[55]
Grid search of all possible combinations	• Allows exploration of all preprocessing possibilities and combinations • Defines the grid to select the best	• Python based open source toolbox available with a collection of preprocessing techniques	[17]
Faster GA-PLS based optimal preprocessing identification	• Faster genetic algorithm - partial least squares modified to efficiently operate on multiple versions of same data	• Can be used to explore preprocessing techniques in parallel • Can be used for faster exploration of tall datasets	[54]

models to achieve the final model for ensemble calibration and prediction. The combination coefficients vector **W** is estimated by MCCV stacked regression. MCCV stacked regression deduces **W** by non-negative least squares (NNLS) according to the following equation:

$$\boldsymbol{Y} = \left[\widehat{y}_1,\ \widehat{y}_2,\ \widehat{y}_3,\ \ldots\widehat{y}_n\right]\mathbf{W} \tag{1}$$

where **Y** contains the reference concentration values for the left-out samples during MCCV resampling and $\widehat{\mathbf{y}}_\mathbf{i}$ contains the corresponding concentration value estimated by the *i*-th model; *n* is the number of models and corresponds to the number of different preprocessings tested. The stacking approach is easy to use and can be implemented using standard chemometric toolboxes which offer PLS modelling. However, a drawback of the stacked approach is that it does not provide insight into what new information a specific preprocess carries and to what extent it improves the overall performances. Also, there are a large number of models to be trained and optimized which might become computationally expensive.

6.2. Ensemble based on design of experiments

The ensemble based on DoE [19] takes its inspiration from the DoE based preprocessing selection and optimization approach [14]. Similarly to preprocessing selection approach, the ensemble approach conceives preprocessing as a four-stage process. However, instead of selection and optimization of the preprocessing as in the traditional DoE-based approach, the ensemble approach explores all the models corresponding to individual preprocessing techniques and all their possible combinations according to a full-factorial design. An example of DoE based ensemble approach is reported in Refs. [19] where PLS models based on numerous preprocessing methods and all their combinations obtainable by a full factorial design (sequentially applying baseline correction, scattering correction, smoothing and scaling) are calculated (Fig. 3). Finally, to perform the ensemble learning, the models which give better predictions than PLS on the raw data are selected and their predictions are averaged as the final prediction. A benefit of DoE-based approaches is that, if a full factorial design is selected, it allows exploration of all possible combinations of preprocessing techniques. Furthermore, compared to the stacked regression approach where all the model outputs were combined in the end using the weight vector, the DoE-based approach only combines the models which have better performance than the PLSR model on the raw data. However, similarly to the stacked regression approach, the DoE-based approach also has the disadvantage that is does not provide insight into how combining different techniques improves the model.

6.3. Methods inspired from multi-block data analysis

Multi-block data analysis is emerging as a key tool for performing multi-sensor data integration and fusion [57,58]. An application of multi-block data analysis can be understood as the ensemble fusion of multiple preprocessing techniques. A recent article demonstrated the benefits of using the multi-block approach for spectral data preprocessing, where several pretreatments were combined using sequential and orthogonalized partial least squares (SO-PLS), thus leading to a boosting procedure [20,21]. The corresponding strategy was given the name SPORT, as the acronym of sequential preprocessing through orthogonalization. The results showed that not only the multi-block approach allowed fusion of multiple preprocessings, but it also helped in identifying the best preprocessing techniques and their combinations. A schema of the SPORT approach is presented in Fig. 4. The SPORT approach involves a series of PLS and orthogonalization steps to extract the unique complementary information from each differently preprocessed data block. As shown in Fig. 4, SPORT starts by calculating a PLS model between the first data block (1st preprocessing) and Y, giving the PLS scores for the first data block and a first (partial) prediction of the response. Successively, both the second data block (2nd preprocessing) and the Y are orthogonalized with respect to the scores of the first regression. This allows to remove from the second block of predictors any redundant information (i.e., any source of variation already modeled by the first block). A second PLS model is then built between the

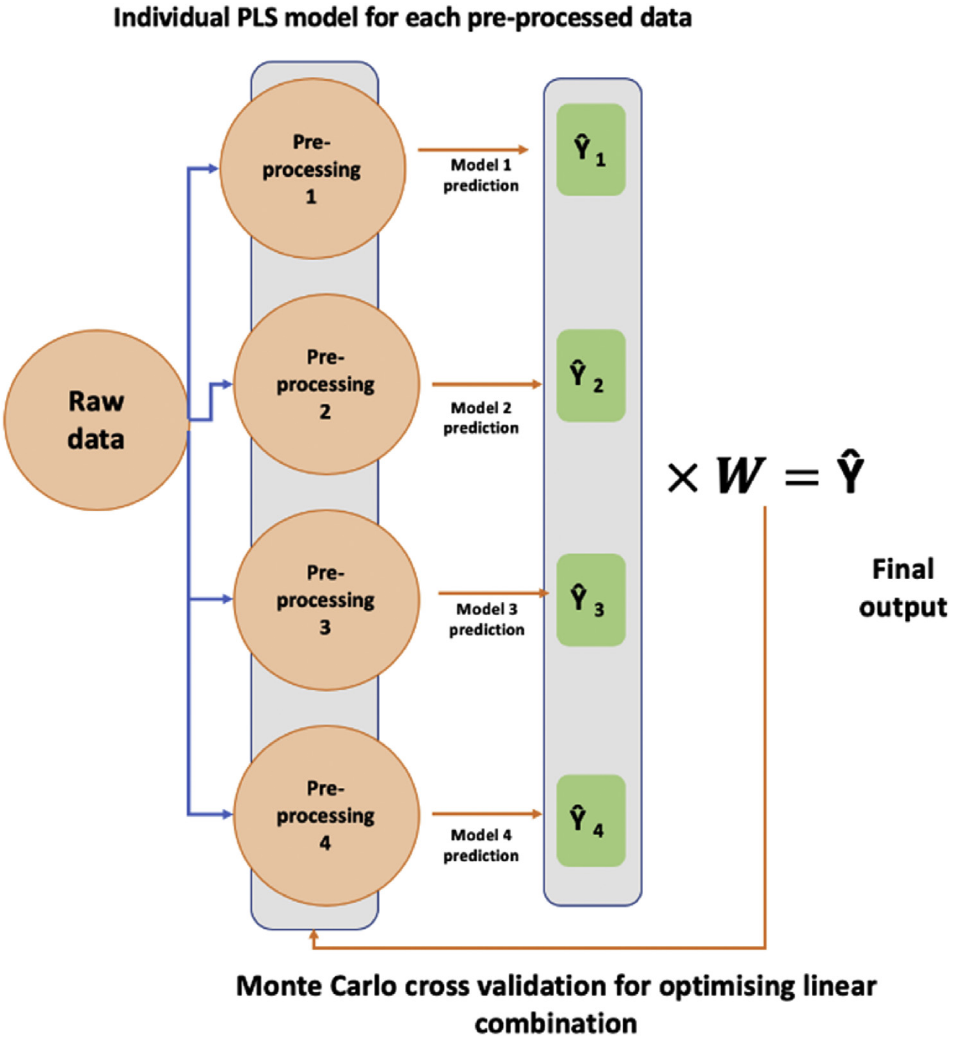

Fig. 2. A summary of the stacking approach to ensemble of preprocessing techniques. At first, some standard preprocessing methods are used to transform the data. Secondly, optimized calibration models are built on differently preprocessed data. The third step is to combine the sub-models to obtain the final global model for ensemble calibration and prediction.

orthogonalized second block and Y residuals of the first regression, giving the PLS scores for the second data block and an updated prediction of the response. The procedure continues until all the desired blocks (corresponding to individual preprocessings) have been used for modelling. Finally, the PLS scores from each block are concatenated and used to predict the response through ordinary least squares regression. The main advantage of SPORT compared to stacked regression and DoE-based ensemble approaches is that it allows exploring the unique information present in each data block (individual preprocessings) with the help of scores and loading plots [20]. However, a key issue of the SPORT approach is that, being a sequential method, it requires the user to define the order of the different preprocessings to be tested. This is a fundamental aspect to be considered and, at the same time, it is not necessarily a drawback, but it can represent a potential source of additional information, since models built on differently ordered blocks could highlight different aspects of the same data set. Indeed, if, on one hand, the need of defining the order of the different preprocessings to be tested could add a further meta-parameter to be optimized, it is also true that the literature on SO-PLS, which constitutes the modeling engine of SPORT suggests that the predictive performances are hardly affected by the sequence of the blocks [21,58]. Therefore, where the focus is only on the predictive accuracy, the challenge of defining the order of preprocessings can easily be dealt with, e.g., by placing all the faster, easier and model-free preprocessing techniques at the start and reserving complex, time-consuming and model based preprocessings for the end. On the other hand, the sequential procedure based on successive orthogonalization steps is such that a block will contribute to the model (with a non-zero number of latent variables) only if it carries unique information (i.e., information not already accounted for by the previously modeled matrices). This means that, by building different models on the same data set but just changing the order in which the individual matrices are presented, can provide a deeper insight into the relationship between the various preprocessing, in terms of which information is shared among all the blocks and which is unique of specific pretreatments. Similarly, one could in principle identify the minimum number of pretreatments carrying non-redundant information. A recent application of the SPORT approach for NIR modelling of highly scattering materials showed that multiple scatter correction techniques carry complementary information and that the model accuracies improved when they were combined through SPORT [21].

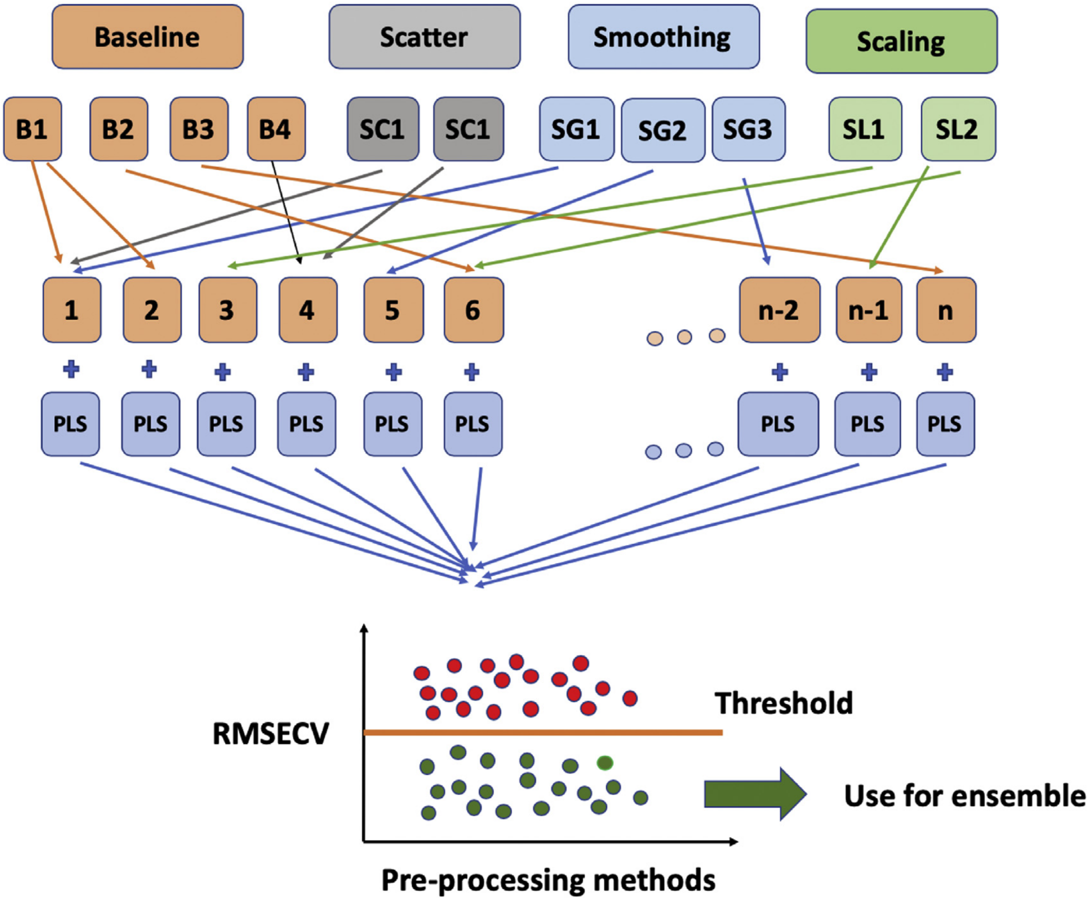

Fig. 3. A summary of the DoE-based ensemble approach. Several PLS models based on numerous preprocessing methods and their combinations (sequentially applying baseline correction, scattering correction, smoothing and scaling) are obtained by full factorial design. Finally, to perform the ensemble learning, the models which give better predictions than PLS on the raw data are selected and their predictions are averaged as the final prediction.

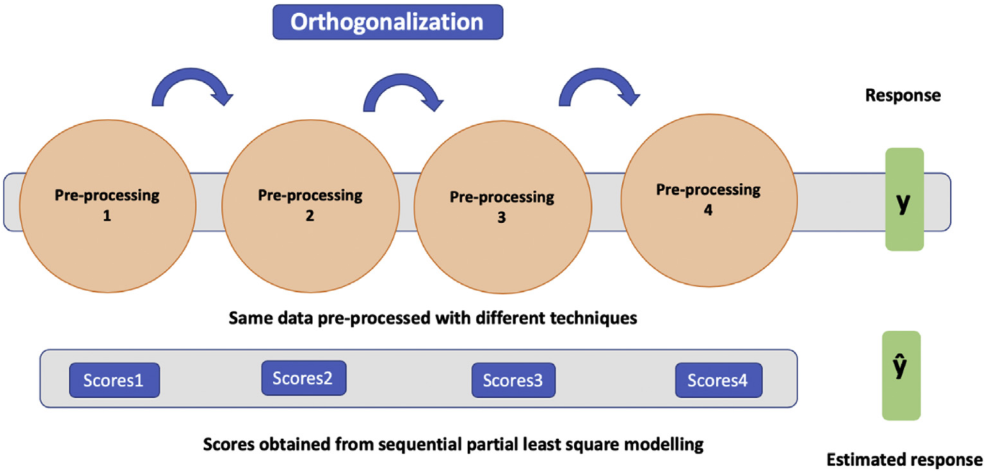

Fig. 4. A summary of the SPORT approach to ensemble fusion of preprocessing techniques. The SPORT approach develops a PLS model between the first data block (1st preprocessing) and Y giving the PLS scores for the first data block. The scores from the first data block are used to orthogonalize the second data block (2nd preprocessing) and the Y. Then a new PLS model is built between the orthogonalized second block and the orthogonalized Y giving the PLS scores for the second data block. This continues until scores are extracted from all the desired blocks (individual preprocessings and their combinations). The final model is obtained by relating the concatenated PLS scores from the different blocks to the response by means of ordinary least squares regression.

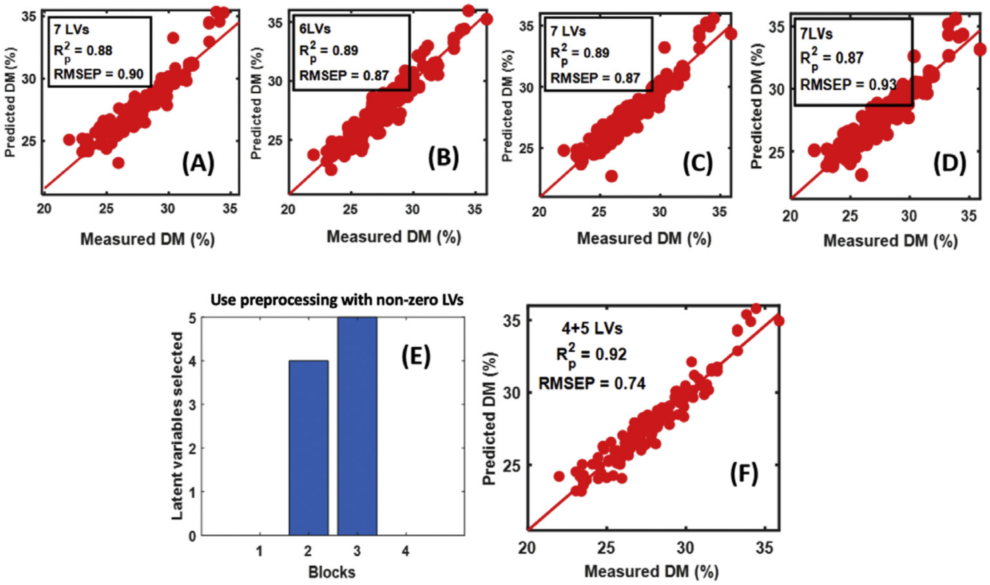

Fig. 5. Summary of SPORT preprocessing fusion for the olive fruit dataset. (A–D) Prediction performances of the PLS models built on data preprocessed by individual scatter correction techniques: (A) SNV, (B) VSN, (C) MSC, (D) RNV. (E) Optimal complexity of the SPORT model: only the VSN and MSC blocks contribute with a non-zero number of latent variables. and (F) Prediction performances of the optimal SPORT model.

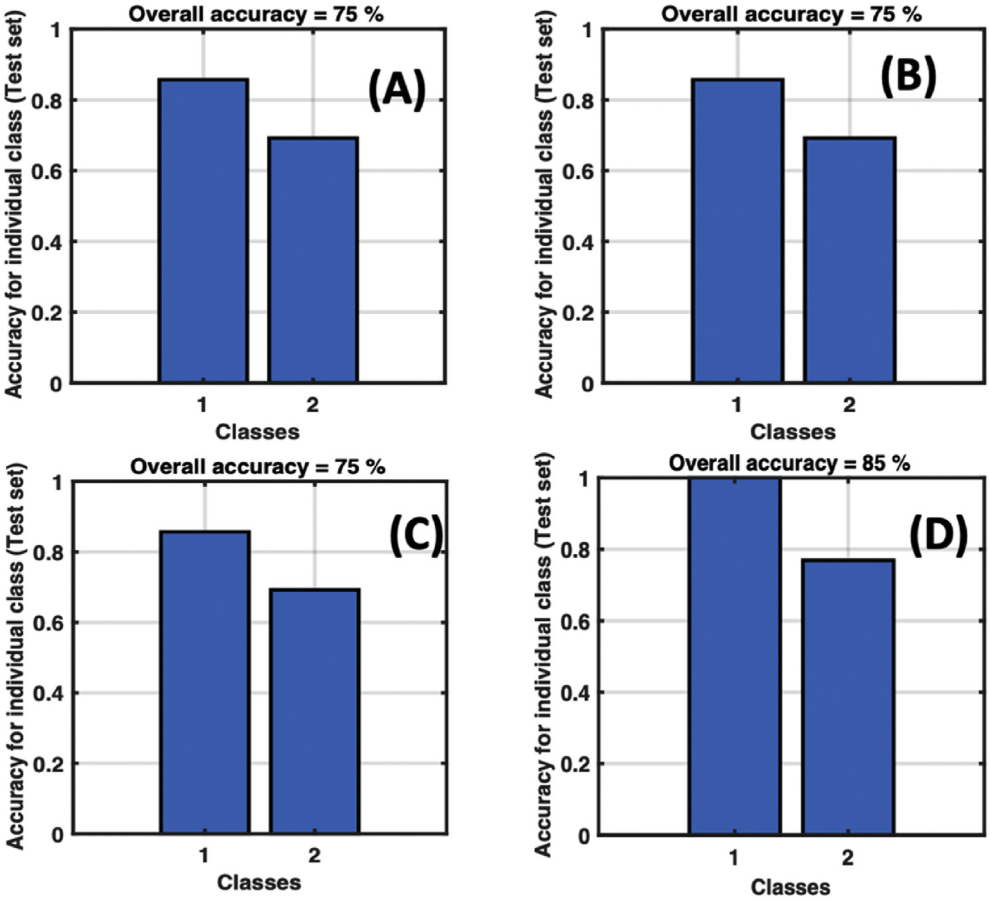

Fig. 6. Summary of SPORT preprocessing fusion for the beer classification dataset. (A–C) Prediction performances of the PLS-DA models built on data preprocessed by individual scatter correction techniques: (A) SNV (2LVs), (B) MSC (2LVs), (C) 2nd derivative (2LVs); and (D) Prediction performances of the optimal SPORT model, fusing the information from SNV (2 LVs), MSC (8 LVs) and 2nd derivative (2 LVs) preprocessed data.

7. An example of sequential preprocessing through orthogonalization

7.1. Olive dry matter regression case

To give the reader a better understanding of the ensemble approach to spectral preprocessing fusion, a demonstration case involving the use of the sequential preprocessing through orthogonalization (SPORT) approach is provided. In particular, the example demonstrates the use of SPORT in the context of the characterization of fresh fruit by NIR spectroscopy, as the signals suffer from scattering effects due to the interaction of light with the physical structure of the fruit peel. The dataset used was related to the possibility of predicting dry matter (DM) in olive fruits with a handheld NIR sensor [59]. A total of 494 individual NIR spectra and corresponding DM measurements were used for the modelling. The dataset was divided into calibration (70%) and test set (30%) using the Kennard-Stone algorithm [60]. Four different scatter correction techniques, i.e. SNV (Block 1), VSN (Block 2), MSC (Block 3) and RNV (Block 4), were used in sequential order. SPORT was implemented as explained in Ref. [20] with the freely available multi-block analysis (MBA) toolbox [61]. The results of SPORT on the olive fruits dataset are shown in Fig. 5. Fig. 5A–D shows the results of standard PLSR calibration for SNV, VSN, MSC and RNV preprocessed data, respectively. Fig. 5E shows the number of complementary LVs (4 from VSN preprocessed data and 5 from MSC preprocessed data) selected by the SPORT approach. Fig. 5F shows the results of the SPORT model fusing the information from VSN and MSC preprocessed data. The SPORT fusion reduced the prediction error by 14.5% and increased the prediction R^2 by 3.3% compared to the best performing model calculated on individually preprocessed data. In summary, fusion of multiple scatter correction techniques improved the model performance.

7.2. Beer classification case

The sequential preprocessing fusion approach SPORT can also be implemented in the context of classification problems. In Fig. 6, an example of Italian craft beer classification using NIR spectroscopy data is presented. The data used in this example are a subset of those used in a previous research to authenticate an Italian craft beer and differentiate it from other similar products [62]. Accordingly, the classification problem involves two categories, Reale beer (the craft beer of interest) and other beers. The calibration set contains a total of 40 samples and the test set consists of 20 samples. The spectral range of the data is 4000–10000 cm^{-1} with a nominal resolution of 4 cm^{-1} [62]. At first three separate PLS-DA models, each based on data individually pretreated with either SNV, MSC or 2nd derivative were developed and tested. Successively, the data pretreated with the three different preprocessings were jointly used for sequential multi-block fusion by the SPORT approach. The sequential modelling was performed using the SO-PLS-LDA option in the freely available multi-block analysis (MBA) toolbox [61]. It can be noted that the data pretreated with any of the individual preprocessing techniques (Fig. 6) resulted in a classification accuracy of 75%. However, with the sequential fusion approach the classification accuracy on the same data was improved to 85%, thus indicating the benefit of combining the useful information from different preprocessings for enhanced model performances. Some other recent applications of SPORT can be found in [21,63].

8. Concluding remarks

The sub-optimal selection of preprocessing has long been a problem in chemometrics and that is why, in recent years, several approaches for a more rational selection of the best preprocessings have been proposed. However, all the preprocessing selection strategies have the disadvantage that they are focused on selecting the preprocessing rather exploring the complementary information present in differently preprocessed data for synergistic modelling. In this study, three approaches to ensemble preprocessing fusion were identified. The three approaches were stacked regression, full factorial DoE-based ensemble and sequential multi-block data analysis by treating differently preprocessed data as separate blocks. Of the three approaches, the stacking and DoE-based ones were based on weighting or averaging the final outputs using either the weight vector or mean estimation, respectively. Both these approaches lack insight into what new information a new preprocessing technique carries. On the other hand, the multi-block data analysis-based approach (SPORT) has the main advantage that it extracts the unique information from differently preprocessed data. Furthermore, since it is based on the sequential calculation of PLS regression models, the set of scores and loadings can be used to explore and interpret what incremental (new) information is brought by the different blocks, i.e. by the different preprocessing techniques applied to the same data. Moreover, the elimination of a block not contributing to the final model with any latent variables directly shows that the application of the corresponding preprocessing technique does not add any new information: accordingly, some techniques can be discarded, leading to a preprocessing selection. Therefore, it can be concluded that, to have a better understanding of the effect of each individual preprocessing in conjunction to model improvement, multi-block analysis-based preprocessing fusion is superior to model averaging-based methods.

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] A. Biancolillo, F. Marini, Chemometric methods for spectroscopy-based pharmaceutical analysis, Front. Chem. 6 (2018) 576.

[2] E. Gorrochategui, et al., Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow, Trac. Trends Anal. Chem. 82 (2016) 425–442.

[3] H.I. Kademi, B.H. Ulusoy, C. Hecer, Applications of miniaturized and portable near infrared spectroscopy (NIRS) for inspection and control of meat and meat products, Food Rev. Int. 35 (3) (2019) 201–220.

[4] R. Deidda, et al., Vibrational spectroscopy in analysis of pharmaceuticals: critical review of innovative portable and handheld NIR and Raman spectrophotometers, Trac. Trends Anal. Chem. 114 (2019) 251–259.

[5] P. Lucci, J. Saurina, O. Nunez, Trends in LC-MS and LC-HRMS analysis and characterization of polyphenols in food, Trac. Trends Anal. Chem. 88 (2017) 1–24.

[6] M. Zoccali, P.Q. Tranchida, L. Mondello, Fast gas chromatography-mass spectrometry: a review of the last decade, Trac. Trends Anal. Chem. 118 (2019) 444–452.

[7] J. Engel, et al., Breaking with trends in pre-processing? Trac. Trends Anal. Chem. 50 (2013) 96–106.

[8] J.-M. Roger, et al., Pre-processing Methods, in: Steven D. Brown, Romà Tauler, Beata Walczak (Editors), Comprehensive chemometrics. Chemical and biochemical data analysis, second ed., Elsevier, Oxford, 2020, pp. 1–75.

[9] R. Lu, et al., Measurement of optical properties of fruits and vegetables: a review, Postharvest Biol. Technol. 159 (2020) 111003.

[10] Å. Rinnan, F.v.d. Berg, S.B. Engelsen, Review of the most common preprocessing techniques for near-infrared spectra, Trac. Trends Anal. Chem. 28 (10) (2009) 1201–1222.

[11] W. Saeys, et al., Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review, Postharvest Biol. Technol. (2019) 158.

[12] L.C. Lee, C.Y. Liong, A.A. Jemain, A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum, Chemometr. Intell. Lab. Syst. 163 (2017) 64–75.

[13] P. Oliveri, et al., The impact of signal pre-processing on the final interpretation of analytical outcomes – a tutorial, Anal. Chim. Acta 1058 (2019) 9–17.

[14] J. Gerretzen, et al., Simple and effective way for data preprocessing selection based on design of Experiments, Anal. Chem. 87 (24) (2015) 12096–12103.

[15] J. Gabrielsson, et al., OPLS methodology for analysis of pre-processing effects on spectroscopic data, Chemometr. Intell. Lab. Syst. 84 (1–2) (2006) 153–158.

[16] S. Verboven, M. Hubert, P. Goos, Robust preprocessing and model selection for spectral data, J. Chemometr. 26 (6) (2012) 282–289.

[17] J. Torniainen, et al., Open-source python module for automated preprocessing of near infrared spectroscopic data, Anal. Chim. Acta 1108 (2020) 1–9.

[18] L. Xu, et al., Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta 616 (2) (2008) 138–143.

[19] X. Bian, et al., A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples, Chemometr. Intell. Lab. Syst. 197 (2020) 103916.

[20] 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.

[21] P. Mishra, et al., SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials, Postharvest Biol. Technol. 168 (2020) 111271.

[22] A. Folch-Fortuny, F. Arteaga, A. Ferrer, Missing data imputation toolbox for MATLAB, Chemometr. Intell. Lab. Syst. 154 (2016) 93–100.

[23] F. Arteaga, A. Folch-Fortuny, A. Ferrer, 2.29 - Missing Data, in: S. Brown, R. Tauler, B. Walczak (Editors), Comprehensive Chemometrics, second ed., Elsevier, Oxford, 2020, pp. 615–639.

[24] J. Trygg, J. Gabrielsson, T. Lundstedt, 3.06 - background estimation, denoising, and preprocessing, in: S. Brown, R. Tauler, B. Walczak (Editors), Comprehensive Chemometrics, second ed., Elsevier, Oxford, 2020, pp. 137–141.

[25] P. Filzmoser, B. Walczak, What can go wrong at the data normalization step for identification of biomarkers? J. Chromatogr. A 1362 (2014) 194–205.

[26] R.H. Jellema, A. Folch-Fortuny, M.M.W.B. Hendriks, 3.05 - variable shift and Alignment☆, in: S. Brown, R. Tauler, B. Walczak (Editors), Comprehensive Chemometrics, second ed., Elsevier, Oxford, 2020, pp. 115–136.

[27] D.F. Thekkudan, S.C. Rutan, 3.07 - denoising and signal-to-noise ratio enhancement: classical filtering, in: S. Brown, R. Tauler, B. Walczak (Editors), Comprehensive Chemometrics, second ed., Elsevier, Oxford, 2020, pp. 143–155.

[28] V.M. Taavitsainen, 3.09 - denoising and signal-to-noise ratio enhancement: splines, in: S. Brown, R. Tauler, B. Walczak (Editors), Comprehensive Chemometrics, second ed., Elsevier, Oxford, 2009, pp. 165–177.

[29] P. Mishra, et al., Automatic de-noising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method, Sensor. Actuator. B Chem. 281 (2019) 1034–1044.

[30] A. Kohler, et al., 3.03 - model-based pre-processing in vibrational spectroscopy, in: S. Brown, R. Tauler, B. Walczak (Editors), Comprehensive Chemometrics, second ed., Elsevier, Oxford, 2020, pp. 83–100.

[31] A. Folch-Fortuny, F. Arteaga, A. Ferrer, PCA model building with missing data: new proposals and a comparative study, Chemometr. Intell. Lab. Syst. 146 (2015) 77–88.

[32] C. Zhang, et al., Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods, Chemometr. Intell. Lab. Syst. 203 (2020) 104063.

[33] Z.M. Zhang, et al., An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy, J. Raman Spectrosc. 41 (6) (2010) 659–669.

[34] P.J. Cadusch, et al., Improved methods for fluorescence background subtraction from Raman spectra, J. Raman Spectrosc. 44 (11) (2013) 1587–1595.

[35] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (5) (2010) 1138–1146.

[36] J.T. Peng, et al., Asymmetric least squares for multiple spectra baseline correction, Anal. Chim. Acta 683 (1) (2010) 63–68.

[37] J. Skogholt, K.H. Liland, U.G. Indahl, Baseline and interferent correction by the Tikhonov regularization framework for linear least squares modeling, J. Chemometr. 32 (3) (2018).

[38] G. Rabatel, et al., VSN: variable sorting for normalization, J. Chemometr. 34 (2) (2020) e3164.

[39] Y. Bi, et al., A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation, Anal. Chim. Acta 909 (2016) 30–40.

[40] L. Li, et al., A new scattering correction method of different spectroscopic analysis for assessing complex mixtures, Anal. Chim. Acta 1087 (2019) 20–28.

[41] S. Kim, et al., Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry, BMC Bioinf. 12 (2011).

[42] B.C. Deng, et al., Global peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using point matching algorithms, J. Bioinf. Comput. Biol. 14 (6) (2016).

[43] C. Tistaert, Y. Vander Heyden, Bilinear decomposition based alignment of chromatographic profiles, Anal. Chem. 84 (13) (2012) 5653–5660.

[44] H. Parastar, N. Akvan, Multivariate curve resolution based chromatographic peak alignment combined with parallel factor analysis to exploit second-order advantage in complex chromatographic measurements, Anal. Chim. Acta 816 (2014) 18–27.

[45] B. Wang, et al., DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics, Anal. Chem. 82 (12) (2010) 5069–5081.

[46] S. Kim, et al., An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure, Bioinformatics 27 (12) (2011) 1660–1666.

[47] N. Hoffmann, et al., BiPACE 2D-graph-based multiple alignment for comprehensive 2D gas chromatography-mass spectrometry, Bioinformatics 30 (7) (2014) 988–995.

[48] A. Folch-Fortuny, F. Arteaga, A. Ferrer, Assessment of maximum likelihood PCA missing data imputation, J. Chemometr. 30 (7) (2016) 386–393.

[49] J.Y. Dong, et al., Group aggregating normalization method for the pre-processing of NMR-based metabolomic data, Chemometr. Intell. Lab. Syst. 108 (2) (2011) 123–132.

[50] Q.X. Zheng, et al., Automatic time-shift alignment method for chromatographic data analysis, Sci. Rep. 7 (2017).

[51] Z.Y. Li, et al., Coherent point drift peak alignment algorithms using distance and similarity measures for two-dimensional gas chromatography mass spectrometry data, J. Chemometr. 34 (8) (August 2020).

[52] H. Zheng, et al., Optimal preprocessing of serum and urine metabolomic data fusion for staging prostate cancer through design of experiment, Anal. Chim. Acta 991 (2017) 68–75.

[53] A. Martyna, et al., Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components, Chemometr. Intell. Lab. Syst. 202 (2020) 104029.

[54] P. Stefansson, et al., Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocessing technique for datasets with many observations, J. Chemometr. 34 (3) (2020).

[55] T.J. Rato, M.S. Reis, SS-DAC: a systematic framework for selecting the best modeling approach and pre-processing for spectroscopic data, Comput. Chem. Eng. 128 (2019) 437–449.

[56] J. Gerretzen, et al., Boosting model performance and interpretation by entangling preprocessing selection and variable selection, Anal. Chim. Acta 938 (2016) 44–52.

[57] A.K. Smilde, et al., Common and distinct components in data fusion, J. Chemometr. 31 (7) (2017), e2900.

[58] Alessandra Biancolillo, Tormod Næs, The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions, in: Marina Cocchi (Editor), Data Fusion Methodology and Applications, Data Handling in Science and Technology 31, Elsevier, Amsterdam, 2019, pp. 157–177.

[59] X. Sun, et al., 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.

[60] R.W. Kennard, L.A. Stone, Computer aided design of Experiments, Technometrics 11 (1) (1969) 137–148.

[61] P. Mishra, et al., 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, p. 104139.

[62] A. Biancolillo, et al., Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31.

[63] Puneet Mishra, et al., Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques, Talanta (2020), 121693. https://doi.org/10.1016/j.talanta.2020.121693. In press.