Recent trends in multi-block data analysis in chemometrics for multi-source data integration

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

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

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

[3]*ITAP, INRAE Montpellier, Institut Agro, University Montpellier, Montpellier, France*

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

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

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

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

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

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

Corresponding author: puneet.mishra@wur.nl

Keywords

2D: 2 dimensional

3D: 3 dimensional

CCSWA: Common component and specific weight analysis

ComDim: Common Dimensions

CT: Calibration transfer

DISCO-SCA: Distinct and common simultaneous component analysis

GCA: Generalized canonical analysis

GSVD: Generalized singular value decomposition

GUI: Graphical user interface

H-PCA: Hierarchal principal component analysis

H-PLS: Hierarchal partial least-squares

JIVE: Joint and individual variances explained

JUMBA: Joint and unique multi-block analysis

MB-PCA: Multi-block principal component analysis

MB-PLS: Multi-block partial least-squares

MB-VIOP: Multi-block variable importance in projection

MBA-GUI: Multi-block analysis graphical user interface

MCR: Multivariate curve resolution

MIR: Mid-infrared

MOCA: Multiple co-inertia analysis

MVP: Multi-block variable partitioning

NIR: Near-infrared

O2PLS: Orthogonal 2-block partial least-squares

OnPLS: Orthogonal n-block partial least-squares

P-EASCA: Penalized exponential analysis of variance simultaneous component analysis

P-ESCA: Penalized exponential simultaneous component analysis

PARAFAC: Parallel factor analysis

PAT: Process analytical technologies

PCA-GCA: Principal component analysis generalized canonical analysis

PCA: Principal component analysis

PLS: Partial least-squares

PO-PLS: Parallel orthogonalized partial-least squares regression

PORTO: Parallel pre-processing through orthogonalization

SCA: Simultaneous component analysis

SCD-PCovR: Sparse common and distinct principal covariate regression

SLIDE: structured learning and integrative decomposition

SO-CovSel: Sequential orthogonalized covariate selection

SO-N-PLS: Sequential orthogonalized n-way partial least-squares

SO-PLS: Sequential orthogonalized partial-least squares regression

SPORT: Sequential pre-processing through orthogonalization

SR: Selectivity ratio

VIP: Variable importance in projection

1. Abstract

In recent years, multi-modal measurements of process and product properties have become widely popular. Sometimes classical chemometric methods such as principal component analysis (PCA) and partial least squares regression (PLS) are not adequate to analyze this kind of data. In recent years, several multi-block methods have emerged for this purpose; however, their use is largely limited to chemometricians, and non-experts have little experience with such methods. In order to deal with this, the present review provides a brief overview of the multi-block data analysis concept, the various tasks that can be performed with it and the advantages and disadvantages of different techniques. Moreover, basic tasks ranging from multi-block data visualization to advanced innovative applications such as calibration transfer will be briefly highlighted. Finally, a summary of software resources available for multi-block data analysis is provided.

Keywords: pre-processing fusion; incremental learning; data fusion; chemometrics

2. Introduction

In analytical chemistry, data obtained by multiple sources is frequently encountered [1-3]. A *multi-block* data set can either come from a multi-platform analysis of the same samples (e.g., to reach a better understanding of the physico-chemical properties of the analyzed objects which is not possible with a single technique [1, 4]), or by the combination of chemical measurements with non-analytical data generated from sensory or consumer sciences [5]. In

both cases, the data is not simply multivariate but is *multi-modal*, i.e., multivariate and multi-source. An example of this would be data generated by two different spectroscopic techniques such as mid-infrared spectroscopy (MIR) and Raman spectroscopy. In this case, spectral profiles are multivariate (as the responses are acquired at several wavenumbers), and the modes are represented by the two different spectroscopic techniques. Furthermore, multi-modal data can also be obtained when working under different conditions, for instance, when multiple batches of an industrial process produce data under different processing conditions [6-9].

Chemometrics has been developed to handle multivariate data generated from analytical techniques [10, 11]. The foundation of chemometrics lies on the identification of the underlying latent spaces using bilinear or trilinear multivariate data decomposition techniques. These explorations of latent spaces are specifically targeted to find any structured variation and/or correlation with the key property of interest. Once identified, latent spaces can be used to perform several data processing tasks, such as transforming high-dimensional data to a lower dimensional representation for data visualization purposes [12], or developing regression models for predictive analysis [13] and identification of key variables of interest [14-16]. Traditional chemometric methods (single-block chemometric techniques), such as principal component analysis (PCA) [12], partial least squares (PLS) regression [13] and their variants, only work properly when the data is single-mode, i.e., generated by only one source of variability, such as a single analytical technique. In the case of multi-block data, the standard single-block-techniques extract only a limited part of the information present in the data [3, 17]. It is only by using multi-block data analysis techniques that it is possible to extract the complementary information from data generated in multiple modes [3, 18].

Multi-block data analysis accomplishes similar tasks to single-block chemometric techniques, i.e., enhancing data visualization [3, 19] improving predictive performances [1, 4], and identifying the key variables that influence the models [3, 19-21]. Furthermore, multi-block

analysis can achieve an enhanced understanding of the common and the distinct information present in the data coming from different platforms [3, 19-21]. The present review provides a brief overview of the multi-block data analysis concept, the various tasks that can be performed with it and the advantages and disadvantages of different techniques. Moreover, basic tasks ranging from multi-block data visualization to advanced innovative applications such as calibration transfer will be briefly highlighted. Special attention has been paid to simplify the explanation of complex concepts and methods related to multi-block data analysis so that users with minimal experience in chemometrics can understand and use advanced multi-block methods in their daily tasks. And finally, a summary of software resources available for multi-block data analysis is provided.

3. When is the multi-block data generated and what are its characteristics?

An example of an experimental that produces a multi-block data set is presented in Fig. 1, where four different spectroscopic techniques (near-infrared spectroscopy, mid-infrared spectroscopy, Raman spectroscopy and fluorescence spectroscopy) are combined to monitor the chemical process taking place in a glass vessel. The data from all the four techniques are acquired simultaneously as presented in Fig. 2. This is just an illustration, but such acquisitions of multi-source data are becoming popular in analytical chemistry [22-25]. A similar example is the non-destructive quantification of (bio-) chemical components in an aqueous process containing fluorescent compounds.

Fig. 1: Scheme of multispectral fiber system (figure courtesy of Art Photonics GmbH, Germany [26]). Raman system (1); FTIR absorption System (2); NIR reflection System (3); fluorescence detector (4); chemical reactor (5); and fiber optic probes (6).

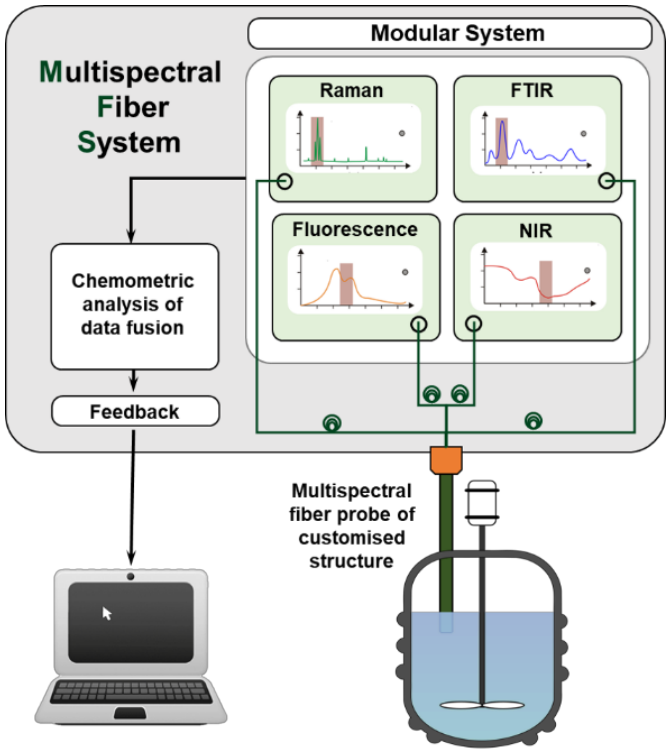


Fig. 2: A schematic of the multiblock data generated in a four-blocks scenario (figure courtesy

of Art Photonics GmbH, Germany) [26].

The main characteristics of multi-block data is that it either consists of multiple matrices corresponding to different analytical platforms generated from measurement on the same sample (Fig. 3A), a combination of matrices and higher order tensors (Fig. 3B), or different independent batch processes (Fig. 3C).

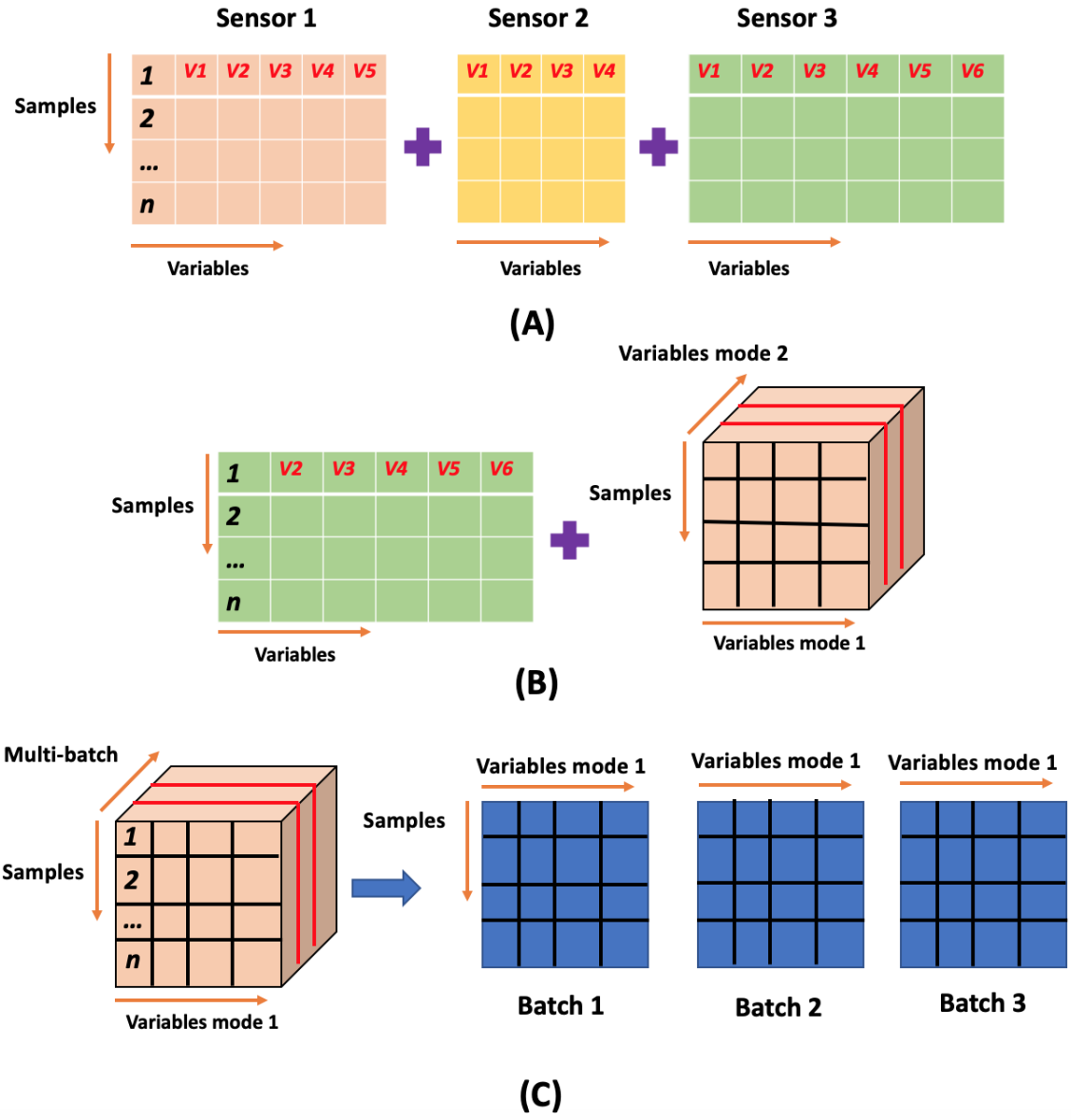


Fig. 3: A summary of typical multi-block data configurations in analytical chemistry. (A) data

generated by multiple analytical platforms in the form of 2D matrices, (B) data generated by multiple analytical platforms in the form of 2D matrices or higher order tensors, and (C) multi-set data from batch processes which can be treated as multi-block data when different batches are treated independently.

4. Multi-block data pre-processing

Data pre-processing is essential in chemometrics and, like the standard one-block methods, multi-block analysis is also influenced by the pre-processing operation. In the case of multi-block analysis, data pre-treatment can be divided into two stages, i.e., the *inter-block* and the *intra-block* pre-processing [8]. Recently, Campos and Reis provided a comprehensive systemization of multi-block data pre-processing and divided the steps into three levels [8]. In particular, the first level [8] includes the standard chemometric pre-processing operations to correct artefacts and uninteresting variations such as noise, multiplicative effects, scaling, baseline drift, peak shift and variations related to external factors [27]. At the second level, the aim is to equalize the contribution of all variables within each block and this can be achieved by classical methods such as mean-centering and unit variance scaling [8]. The third (and final) level aims at equalizing the inter-block systemic effects such as the differences in the scales, number of variables and the pseudo rank of different blocks [8]. This level of pre-processing is necessary as some multi-block analysis methods tend to favor the blocks with larger variations, leading to model bias. However, with proper scaling or weighting of blocks, it has been proven that model interpretation and predictive accuracy can be increased [28]. The third level (inter-block) pre-processing approaches are mainly scaling, such as block scaling, block variance scaling and block rank scaling [8, 28]. Block scaling and block variance scaling balance the effect of the modelled blocks, to avoid any block dominating the model [8]. More detailed information on multi-block pre-processing can be accessed in a recently published

work [8]. Although multi-block pre-processing is important, not all approaches to multi-block data analysis require all levels of pre-processing. For example, the sequential and parallel approaches to partial least-squares regression are less sensitive to the relative scaling of the blocks and can also deal with the differences in the ranks of multi-block data [8], because these methods handle the multi-block data one block at a time, involving orthogonalization steps which do not affect the relative weighting of the blocks [8, 18].

In conclusion, the main outcome of all these considerations is that multi-block pre-processing must be carefully planned in accordance with the multi-block analysis to be performed.

5. Multi-block exploratory data analysis

Exploratory analysis is often performed to obtain low-dimensional representations of high-dimensional multivariate data, to facilitate its interpretation. The key properties of the data can thus be visualized by means of interpretable 2D or 3D plots. In chemometrics, one of the aims is to identify the latent (sub-)spaces capturing key properties of data such as highest variance/closest fit in the case of PCA, or maximum co-variance to the response variables for supervised data decomposition such as PLS. The data decomposition results in a set of scores and loadings, where the scores are the low dimensional representation of the data and the loadings are the latent vectors spanning the relevant sub-space. In standard one block chemometrics, different methods are available for latent space modelling. In fact, the identification of these latent spaces depends on data modes: for a simple 2D data matrix, bilinear decomposition methods such as PCA can be implemented, whereas when the order of the data increases to 3D or more, then higher-order extensions of PCA called Tucker and parallel factor analysis (PARAFAC) can be implemented. However, one block chemometrics methods do not provide a complete solution to deal with multi-block data.

To deal with the challenges of visualizing multi-block multivariate data, several extensions of one-block chemometric methods as well as new specific multi-block approaches have emerged in recent decades. A summary of these methods is provided in Table 1. There are different classifications of the multi-block methods; one is based on the separation into two families, depending on how these approaches handle common and distinct information in the blocks. The first family comprises approaches based on identifying the common information among different data blocks and later exploring the contribution of each block to the common components. The second family of methods is based on the simultaneous extraction of the common as well as the distinct information in the different data blocks. The methods belonging to the first family are extensions of PCA. A simple, popular extension of PCA for the multi-block scenario is SUM PCA [17], where multiple data blocks are concatenated in the variable's domain and standard PCA is performed on the joint data, leading to extraction of global principal components. More advanced extensions, called multi-block PCA (MB-PCA) or consensus PCA [29], allow extraction of global components as well as the contribution of the associated blocks. The extraction of subsequent components is performed by deflation of the matrices with respect to the global components. This is done by regressing all the variables in the different blocks with the extracted global component, and the resulting residuals are then used to extract new global components, and so on. The deflation step is performed such that each global component contains unique information. A method like MB-PCA, called multiple co-inertia analysis (MOCA), was also proposed to explore multi-block data [30]. From an algorithmic point of view, MOCA follows the same procedure as MB-PCA to extract the global components, however, in the second step, the block loadings are used for block deflation, not the global scores as used in the case of MB-PCA [31]. The deflation with the block loadings in MOCA allows the blocks to capture orthogonal information. Furthermore, an advanced version of PCA for multi-block analysis is the hierarchal PCA (H-PCA) which provides the global

components along with the weights of the blocks to reflect the importance of each block in contributing to the extracted global components [29, 32]. Like H-PCA, a method called common components and specific weight analysis (CCSWA or ComDim) has gained attention, as it also allows the extraction of global components and the specific weights for each block to have an insight into each one's importance [33-35]. Ideally, both H-PCA and CCSWA lead to similar solutions but CCSWA is more sophisticated in terms of mathematical explanations with several possibilities of method extensions [35-38].

The methods extracting common components (global components) work well when the objective is to globally explore the different blocks, however, they lack the means to highlight which information is unique in each block. A framework for extraction of common and distinct information recently summarized by Smilde et al. [3], is presented in Fig. 4, where the three circles represent three different blocks of data measured on the same samples and the letters D and C indicate the distinct and the common parts of the information, respectively. Several methods can be identified in the framework of common and distinct information extraction, and the evaluation of their performances can be found elsewhere [3, 7, 20, 21]. Some examples of these are distinct and common simultaneous component analysis (DISCO-SCA) [39], principal component analysis - generalized canonical analysis (PCA-GCA) [3], generalized singular value decomposition (GSVD) [40], orthogonal2 PLS (O2PLS) [41], joint and individual variances explained (JIVE) [42], structure revealing data fusion [43, 44] and structured learning and integrative decomposition (SLIDE) [45]. In the case of DISCO-SCA, the first step involves a SCA step to decompose the matrices to a set of scores and loadings matrices. In the second step, the loadings matrices are partitioned and orthogonally rotated to reveal the common and distinct components in the multi-block scenario. In the case of the PCA-GCA, the first step is to perform PCA followed by a GCA. PCA is performed to enhance the stability of the GCA components. The second step is the regression of each block onto its own

common components, the residuals of the regression are the distinct subspace which can then be used for subsequent PCA [3]. In the case of GSVD, after performing a preliminary SCA step to filter out the noise from the data, singular value decomposition is jointly applied to the different matrices, under the constraints that the left singular vectors (i.e., normalized scores) be the same for all blocks and that the matrices of singular values $\boldsymbol{D}_b$ obey $\sum_{b=1}^{N_{blocks}} \boldsymbol{D}_b^2 = \boldsymbol{I}$, N_{blocks} being the number of blocks and $\boldsymbol{I}$ the identity matrix of appropriate dimensionality. Then, identification of a component as common or distinct is made based on the associated singular values for the different blocks [40]. The O2PLS approach can be considered as a multi-block extension of orthogonal PLS (OPLS), with the relevant difference that no asymmetric relation among the blocks is implied, so that the method can be used also for exploratory purposes [41]. At first, the distinct components are extracted from each block, which is the deflated; accordingly, then, a PLS step is carried out to extract the common components from the deflated blocks. In the case of JIVE, the different blocks of data are directly decomposed into a set of common and distinct information by an iterative procedure involving alternating steps of SVD decomposition of the concatenated blocks for the estimation of the common component and SVD decomposition of the individual blocks after deflation of the estimated common components to account for the distinct variation [42]. In the case of structure revealing data fusion, the matrices are jointly factorized and with the help of penalty terms, the common and distinct information is extracted [43,44]. Finally, the SLIDE can be considered as an intermediate model between SUM-PCA and JIVE as it allows components to be partially shared (i.e., common only to some blocks). This is achieved by arranging the loadings in a block-dependent structure and imposing structure sparsity to reveal the common, distinct and the partially shared information [45]. In analytical chemistry, experiments are often organized by means of experimental designs (DoE), and much insight about the samples can be gained in this way. Recently, to deal with this, a new multi-block data visualization tool for exploration

of multi-block data generated by designed experiments was proposed. The method is called penalized exponential analysis of variance - simultaneous component analysis (PE-ASCA). PE-ASCA is a combination of penalized exponential - simultaneous component analysis (PE-SCA) [21] and the analysis of variance - simultaneous component analysis (ASCA) [46]. In PE-ASCA, the multi-block data is first partitioned into common and distinct information and later ASCA models are used to incorporate the design information while exploring the data using the SCA. The application of PE-ASCA was recently presented in the domain of metabolomics [19].

Another interesting family of methods is the extension of the multivariate curve resolution (MCR) bilinear decomposition technique [47] to the multi-set configuration. MCR operates a self-modeling resolution of mixed profiles into the contribution of the corresponding pure constituents, through a bilinear modeling usually incorporating chemically inspired constraints (e-g-, non-negativity, unimodality, mass-balance, selectivity, just to cite a few). The basic trick behind the use of the MCR for dealing with multiple data sources is to first concatenate the data matrices along the common direction and then analyze this augmented data array through MCR, retaining, as a sort of additional constraint, the information related to the presence of different data blocks. In this respect, the method is rather flexible, as it can easily deal with cases where the common direction is represented by the variables (e.g., in multi-batch situations), by the samples (multi-source data integration) or, even by both (e.g., with different sets of samples having all been analyzed by more than one technique). More details on the extension of the MCR for multi-block data analysis can be found elsewhere [47]. Here it should also be stressed that by introducing suitable so-called selectivity constraints, it is possible to guide the model towards the extraction of both common and distinct components. As well, through the use of other constraints (e.g., correlation), MCR can also be employed for predictive purposes.

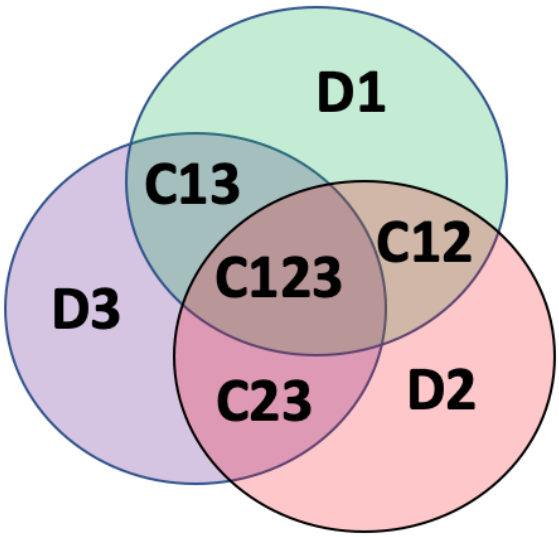


Fig. 4: A framework of common and distinct information extraction from multi-block data. Each circle represents data from a different technique. Inside each circle, D is the distinct information and C is the common information. The figure is inspired by the multi-block data concept presented in [3]).

6. Multi-block predictive modelling

In recent years, a lot of effort in analytical chemistry has been put into developing spectroscopic methodologies to replace certain complex and highly sophisticated wet chemistry routines for quantitative analysis. In chemometrics, a common method to perform this task is partial least-squares (PLS) regression [13] which decomposes the data into a set of scores and loadings. Later, the scores are used to perform the ordinary least-squares regression. In the case of PLS regression, the scores are extracted to have maximum covariance with the response variable(s). However, PLS cannot be explicitly implemented in the scenario of multi-block data, especially when the aim is to extract common and distinct information. Several approaches to do multi-block predictive analysis have recently gained attention and a summary of methods can be

found in Table 1. Unlike standard PLS regression and classification modelling, the aim of multi-block predictive methods is to extract the complementary information from multiple sources to improve the quality of the models (prediction accuracy and/or interpretability). In chemometrics, most of the methods for multi-block predictive analysis are extensions of standard PLS regression to the multi-block scenario. In this regard, one of the first methods developed was multi-block partial least-squares (MB-PLS) regression. This approach, in the formulation proposed by Qin et al., 2001 allows the extraction of global scores by maximizing the covariance with the response variable(s) [29, 48, 49]. The extracted global scores are used in ordinary least squares regression to obtain predictive models. Hierarchal PLS (H-PLS) is a more sophisticated method that allows the extraction of global and block components, giving the possibility of understanding the relative contribution of each block to the global model [29, 32, 50]. A similar method, called P-ComDim, or ComDim (k+1), extracts global scores that capture the maximum covariance with the response variable(s) [37]. This is done by maximizing the covariances between the local scores of each block and the scores of the response block. In a procedure like that of CCSWA, the loadings of the variables in each block and the weight (*salience*) of each block can be calculated for each common component. However, the MB-PLS and the ComDim (k+1) approaches do not provide a clear extraction of the common and distinct information from the different data blocks. To deal with this, orthogonal n-PLS (OnPLS) was proposed [51], which is the extension of the two-block O2PLS to the multi-block scenario. As O2PLS, OnPLS does not introduce a priori any asymmetry among the blocks, so that it could in principle be used as an exploratory technique; however, if one block contains the response(s) to be predicted, by suitably combining the scores extracted from all the other matrices, a global regression model can be calculated.

More recently, two other methods, i.e., sequential and orthogonalized-PLS (SO-PLS) and parallel and orthogonalized-PLS (PO-PLS), were also proposed as extensions of standard PLS

[52]. The SO-PLS approach involves a series of standard PLS regression and matrix orthogonalization operations to extract sequentially the complementary information from different data blocks; a generic schema of the algorithm is presented in Fig. 5. As mentioned, in SO-PLS, the extraction of information is sequential, meaning that the aim is to incorporate blocks of data one at a time and to assess their incremental contribution. A PLS regression model is calculated between the first block $\mathbf{X}_1$ and $\mathbf{Y}$, yielding scores $\mathbf{T}_1$. Then, all the remaining blocks $\mathbf{X}_2$, ... $\mathbf{X}_k$ and $\mathbf{Y}$ are orthogonalized with regards to $\mathbf{T}_1$. The process is repeated on the second block, and so on for all the blocks, taking care to orthogonalize all the following blocks with respect to the previously modelled matrices. The major advantages of SO-PLS are linked to the orthogonalization, which removes redundant information, and to its sequential nature, which allows the interpretation of the incremental contributions provided by each data block. The SO-PLS approach is particularly advantageous when the aim is to identify possible extra benefits from the inclusion of each block of information into the model [18]. On the other hand, the PO-PLS approach involves a combination of PLS regression, generalized canonical correlation analysis (GCA) and multiple orthogonalization steps [5]. PO-PLS, unlike SO-PLS, does not explore the blocks sequentially, but aims at identifying the common and the distinct information in different blocks to have a better understanding of how the combinations of blocks contribute to the improved predictive performances.

Multi-block variance partitioning (MVP), originally proposed by Skov et al. in 2008 [53], presents some similarities with both SO-PLS and PO-PLS. It was one of the first methods to specifically focus attention on identifying the unique part and the common part of the Y-related variation in the predictor blocks; this is accomplished by using PLS models between predictor blocks and a common response. For each predictor block $\mathbf{X}_k$, the total variance of the responses $\mathbf{Y}$ is partitioned into a unique part (that ascribable only to that particular predictor block), a common part (the one shared also with the other independent matrices) and an uninformative

361 part (which is the portion of Y-variance not relevant for by that particular regression model).
362 Operationally, individual PLS models are calculated between each $\mathbf{X}_k$ and $\mathbf{Y}$. For each predictor
363 block, the uninformative variance is associated to the residuals of that regression, while the
364 unique contribution is calculated after orthogonalizing the predicted responses (variable-wise)
365 with respect to the corresponding predicted responses based on all the other predictor blocks.
366 The common variation is obtained by subtracting the contribution of the unique and
367 uninformative parts from the total variance.

368

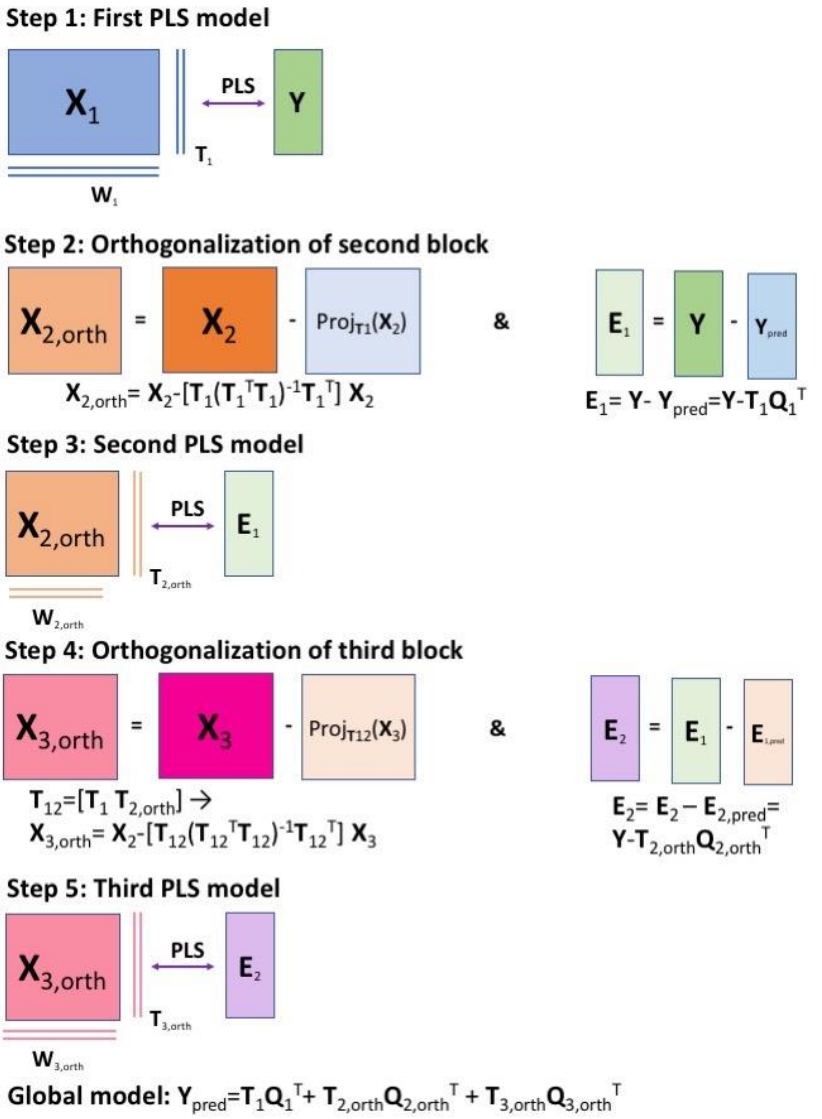


Fig. 5: A schema presenting the sequential orthogonalized partial least squares (SO-PLS) regression method [26].

7. Multi-block analysis for variable selection

Data generated from several analytical techniques, such as optical spectroscopy, mass spectrometry, nuclear magnetic resonance, and chromatography, consists of many variables, of which only a subset is informative. In fact, most of the variables are often correlated or related to some background phenomenon which is not of interest to explain the response variable(s).

Therefore, in chemometrics, variable selection is often performed to identify the most useful variables for the characterization of the response [14, 15]. To deal with the multi-block scenario, several predictive data fusion methods have also been extended to incorporate variable selection [54, 55]. A summary of the multi-block variable selection methods is presented in Table 1. The main benefit of these approaches is that they allow the selection of complementary variables from multiple sources which jointly achieve good model performance. Variable selection approaches can be divided into three main categories, i.e., filter, wrapper and embedded methods. Techniques belonging to the first family are based on ranking the predictors according to some model-based criterion, e.g., variable importance in projection (VIP) or selectivity ratio (SR), and retaining only those variables for which specific parameters exceed given thresholds. In a multi-block scenario, some PLS-inspired variable selection methods have been discussed; for instance, in the context of SO-PLS [54] or OnPLS [56, 57]. Second, wrapper approaches directly calculate multi-block models with different combinations of subsets of variables and select the one that gives the best results (usually determined by internal validation, e.g., cross-validation). Finally, embedded methods carry out the variable selection while building the model. In this context, an interesting possibility is the recently proposed multi-block method, called sequential and orthogonalized covariance selection (SO-CovSel) [55]. As the name implies, this approach is strongly related to SO-PLS, and shares some of its advantages. Nevertheless, SO-CovSel is especially suited to variable selection and the interpretation of the system under study because it directly provides information about which variables drive the model most. However, a key thing that SO-CovSel [55] and other multiblock variables selection methods [54, 56, 57] lack is the clear explanation of the variables that contribute to the common and distinct variability of the different data blocks. To deal with this, two new methods recently emerged in the chemometrics domain, the first is the multiblock variable influence on orthogonal projections (MB-VIOP) [57] and the

second is the sparse common and distinct covariate regression (SCD-CovR) [58]. The MB-VIOP utilises the variable importance in projection (VIP) approach to sort the variables based on their importance for the simplification and interpretation of the OnPLS model [57]. VIP is performed on the global, common and distinct components extracted by OnPLS, thus reflecting the key variables in the global, common and distinct fusion of multi-block data. The SCD-CovR approach on the other hand combines the sparse principal covariate regression with the simultaneous component analysis to extract the variables explaining common and distinct variations in the multi-block data [58].

8. Multi-block analysis for higher order data fusion

Higher order data in analytical chemistry is commonly encountered [59]. Let us consider a situation where N samples are analyzed by NIR spectroscopy, giving rise to spectra comprising M variables at T time points. The resulting data structure is a cube of dimensions $N \times M \times T$. If the same objects are analyzed by another platform, at only one time point, this leads to a data matrix of dimensions $N \times K$, where K is the number of variables measured by another platform. The resulting data set is a multi-block one, but the data structures present diverse dimensionalities. To handle such data sets, the most straightforward solution would be to unfold the cube into a matrix and to apply the traditional data fusion approaches. Nevertheless, it has been demonstrated that, when modelling multi-way structures, it is better to leave their natural dimensionality untouched, exploiting suitable methods for their analysis, rather than unfold them out into two-dimensional arrays. In the light of this, multi-block methods for the combination of arrays presenting different number of modes have been proposed. These multi-block methods can be used for both exploratory and predictive purposes. Currently, multi-block methods for unsupervised fusion of higher-order data are mostly based on coupled tensor and matrix factorization approaches [43, 44]. In the context of predictive analysis, both approaches

based on multi-block multi-way covariate analysis [60] and the more recently proposed extension of sequential and orthogonalized PLS regression to multi-way arrays (SO-N-PLS) [59] are available. This latter approach resembles SO-PLS presented above, with the main difference being that multi-way blocks are handled by means of N-PLS rather than PLS, to maintain their multi-way structure.

9. Innovative uses of multi-block analysis

Apart from standard chemometric tasks such as exploratory analysis, regression, classification and variable selection, other innovative applications of multi-block methods are emerging. Two such applications are pre-processing selection and fusion, and calibration transfer. Pre-processing selection is a key step in chemometric modeling, and it is largely debated since it is difficult to define an optimal strategy. Often users struggle between different pre-processing techniques to identify the best pre-processing or the best combination of pre-processing techniques. A novel application of multi-block data analysis is to perform the fusion of multiple pre-processing techniques [61-63], where the same data after pre-processing with different methods can be considered as a multi-block dataset and can then be processed by multi-block regression and classification. Recently, a technique called sequential pre-processing through orthogonalization (SPORT) was proposed [64]. SPORT is based on the SO-PLS approach to data fusion where the model learns in an incremental way the complementary information present in different data blocks. Recent applications of SPORT can be found relating to selection of pre-processing [64] and complementary fusion of scatter correction techniques [61] in NIR spectroscopy. Since the SPORT approach is sequential, it is necessary to define the order of pre-processing. The order can be decided upon, based on the complexity of pre-processing techniques so that all easy, fast, and model-free techniques are used at the start and

451 the complex, slow, model-based techniques are reserved for the end. However, to deal with the
452 decision about application, a new pre-processing fusion approach called parallel pre-processing
453 through orthogonalization (PORTO) was proposed [65]. PORTO is based on the PO-PLS
454 procedure of predictive multi-block analysis and allows different pre-processing options and
455 their combinations to be explored in parallel. The PORTO approach has the advantage over the
456 SPORT approach in that it provides a better insight into the common and distinct information
457 highlighted by different pre-processing techniques. However, it has been reported that both the
458 SPORT and PORTO approaches usually lead to the same predictive performance. The concept
459 of considering differently pre-processed versions of the same matrix as a multi-block data set
460 had already been considered in the framework of the MVP method [53]. In that context, the
461 use of MVP was advocated in order to get a deeper insight into which pre-processings could
462 carry similar information and which ones could possibly add a unique contribution,

463 The second innovative application of multi-block data analysis is related to calibration transfer
464 (CT). CT is a widely explored task in chemometrics when the aim is to use a model developed
465 using one sensor, on another similar sensor. The aim of calibration transfer is to remove the
466 differences between the two instruments so that the model developed on one instrument can be
467 transferred and used with the other sensor. Recently, methods based on multi-block techniques
468 have emerged for calibration transfer [66]. A recent method called joint and unique multi-block
469 analysis (JUMBA) was proposed for the calibration transfer of NIR models [66]. The method
470 relies on the assumption that the two instruments have a major part of information in common
471 and a minor part that is distinct. Once the common information is identified by the multi-block
472 methods, the model developed on one sensor can be applied on the other.

473

474

475 *Table 1: A summary of multi-block methods available for multi-source data integration in*

476 *chemometrics.*

Tasks	Data order	Methods	Background	Key features	References
Exploratory data analysis	2nd order data	Consensus principal component analysis (CPCA)	• Global components are extracted maximizing the variance • Individual blocks are later regressed on the global components to extract the weights for individual blocks to have an insight on the contribution of each block to the global component	• Weights of individual blocks provides importance for each block in the final model • Superweights normalized to length = 1	[50]
		Extensions of multivariate curve resolution (MCR)	• The data blocks or matrices are concatenated along the common direction (rows, columns, both) • MCR is applied to the augmented data array	• By suitable definition of the selectivity constraint, can extract both common and distinct components • Can deal with "incomplete" multi-sets (some matrices sharing the row-dimension and some others the column one • Through the use of suitable constraints, can be used also for predictive modeling	[47]
		Hierarchal principal component analysis (H-PCA)	• Like CPCA but relies on different normalization (superscores normalized to length = 1). •	• Objective function is not clear • May provide different solutions depending on initialization.	[29, 32]
		Common component and specific weight analysis (CCSWA)	• Global components are extracted sequentially by maximizing the variance of the weighted sum of cross-product matrices. • Individual blocks are later deflated on the global components, and the whole procedure is	• On each dimension, weights of individual blocks indicate the importance of each block in the construction of the corresponding component • Both global and local components for	[34]

			repeated on the deflated blocks.	each block can be obtained, as well as loadings for each block.	
		Multiple co-inertia analysis (MOCA)	• Data are preliminary transformed • Orthogonal components are extracted so to maximize the sum of the squared covariance with the scores of each block	• Provides a simultaneous ordination of measurements and variables of multiple blocks	[30]
		Orthogonal 2 partial least-squares (O2PLS)	• Preliminary estimation of common subspace by svd($\boldsymbol{X}_2^T\boldsymbol{X}_1$) • Orthogonalization of the blocks with respect to common subspace • Distinct component extracted from the orthogonalized blocks • After deflation of the distinct components, the final common components are extracted by a PLS-like step between the blocks	• Common components are different between the blocks • No asymmetric relation between the blocks is assumed	[41]
		Distinct and common simultaneous component analysis (DISCO-SCA)	• The joint subspace is extracted by SCA. • Target rotation of the block loadings is used to identify common and distinct components	• Does not allow the extraction of partially shared components	[39]
		Joint and individual variances explained (JIVE)	• Iterative extraction of common and distinct components • SVD on the concatenated data matrices to estimate the common components • Deflation of each block with respect to the common components • SVD on the deflated blocks to estimate the	• The ranks of common and distinctive matrices are determined by permutation tests.	[42]

			distinct components		
		Principal component analysis Generalized canonical analysis (PCA-GCA)	• Preliminary PCA on individual blocks to filter out noise. • Finds linear combination of the blocks which best fit to a set of orthogonal common components	• Focuses on common components • Distinctive components are obtained by PCA on the residual matrices after regressing each block on the common components.	[3]
		Generalized singular value decomposition (GSVD)	• Preliminary SCA to filter out noise. • Joint SVD of the different data matrices • Identification of common and distinct components based on the singular values	• Originally proposed for multi-set data sharing the variable dimensions.	[40]
		Structured learning and integrative decomposition (SLIDE)	• Loadings are organized in a block-dependent structure • Structure sparsity is imposed to reveal the common, distinct and the partially shared information	• Can be considered as an intermediate model between SUM-PCA and JIVE • Components common only to some blocks can be extracted	[45]
		Penalized exponential simultaneous component analysis (P-ESCA)	• Penalties are incorporated in the simultaneous component analysis for separating common and distinct information in the multi-block data	• Common and distinct variation in the data in be explored separately	[21]
		Penalized exponential analysis of variance simultaneous component analysis (P-EASCA)	• Combines the penalized exponential simultaneous component analysis with the analysis of variance simultaneous component analysis • The multi-block data is decomposed into common and distinct part and later the ASCA is used for	• Only multi-block technique available for exploration of designed experimental data	[19]

			exploratory analysis of common and distinct variation		
	Higher order data	Combined matrix and tensor factorization	• Combines the matrix factorization and tensor factorization	• Data of multiple order such as 2D, 3D etc. can be jointly explored	[43, 44]
Predictive analysis	2nd order data	Multi-block partial least-squares regression	• Global components are extracted maximizing the covariance with the response variable(s) • Individual blocks are later regressed on the global components to extract the weights for individual blocks to have an insight into the contribution of each block to the global component	• Weights of individual blocks provides importance for each block in the final model • Block weights and superweights are normalized to unit length	[48]
		Hierarchal or consensus partial least-squares regression	• A CPCA cycle is performed on the multiple X blocks. • A PLS cycle is done between the superscores and the response(s)	• Superscores are normalized to unit length	[29, 50]
		Orthogonal n partial least-squares (OnPLS) regression	• Extension of O2PLS to the multi-block scenario • A global regression model is calculated between the block containing the responses to be predicted, and the scores extracted from all the other matrices	• Can be also used for exploratory analysis, since no asymmetric relations between the blocks are assumed a priori	[51]
		ComDim (k+1), or P-ComDim	• Two sets of global components (for the predictor blocks and for the response blocks) are extracted sequentially by maximizing the variance of the sum of cross-product matrices involving both predictor and response blocks. • Individual blocks are later deflated on the global components, and the whole procedure is	• The weights (saliences) of each block on each dimension, indicates its importance in the determination of that common component	[37]

			repeated on the deflated blocks.		
		Sequential orthogonal partial least-squares (SO-PLS) regression	• Includes a combination of partial least-squares regression and sequential orthogonalization step to extract complementary latent variables from multi-block data	• Complementary unique information is extracted	[52]
		Parallel orthogonal partial least-squares (PO-PLS) regression	• Includes a combination of partial least-squares regression, canonical correlation analysis and orthogonalization step to extract common and distinct latent variables from multi-block data	• Common and distinct information can be extracted • Good for the cases when order of block is not important or all blocks are of equal importance	[52]
		Multi-block variance partitioning (MVP)	• Individual PLS models between each predictor block $\mathbf{X}_k$ and $\mathbf{Y}$ • For each block, unique Y-related variation is obtained by orthogonalizing the predicted responses with respect to the corresponding responses predicted using the other blocks • Common variation is obtained by subtracting the unique part and the residuals from the total variance	• Extracts common and distinct information • Scale invariant • Can be extended to evaluate the performances of preprocessing methods	[53]
	Higher order data	Multi-way multi-block covariates regression	• Extension of principal covariate regression • Scores are extracted so to explain the variation in their associated block and convey similarities between the blocks	• A different number of scores can be extracted from each block • The extent to which inter- and intra-block variation is accounted for is regulated by a metaparameter.	[60]

		Sequential orthogonal N-way partial least-squares regression	• Includes a combination of PLS regression or N-PLS regression (depending on the nature of the block to model) to sequentially extract complementary latent variables from multi-way multi-block data	• Complementary information is extracted	[59]
Variable selection	2nd order data	Variable importance in projection (VIP) + SO-PLS	• The method is based on estimating the variable importance on the components extracted by the sequential orthogonalized partial least-squares (SO-PLS) regression	• The variables can be extracted with simultaneous SO-PLS modelling	[54]
		Sequential orthogonalized covariate selection (SO-CovSel)	• Based on the sequential covariance maximization and orthogonalization step to select variables across multiple blocks of data	• The approach is sequential so data blocks based on their importance can be arranged by user • Discrete variables are selected which can be used for developing cheap sensors	[55]
		Multi-block variable important in orthogonal projections (MB-VIOP)	• The method is based on estimating the variable importance on the common and distinct components extracted by the orthogonal n partial least-squares (OnPLS) regression	• Allows exploration of common and distinct variables amount different blocks of data	[57]
		Sparse common and distinct covariate regression (SCD-PCovR)	• Combined the sparse principal covariate regression with the simultaneous component analysis to extract the common and distinct variables in multi-block data	• Allows exploration of common and distinct variables amount different blocks of data • Sparsity parameter can be tuned by user to command the	[58]

				variable selection	
Calibration transfer	2nd order data	The joint and unique multi-block analysis (JUMBA) for calibration transfer	• Based on the framework of Orthogonal n partial least-squares (OnPLS) to identify common and distinct information in multi-block data	• Allows multi-instrument calibration transfer • Enhanced understanding about intrinsic differences of instruments can be gained within the framework of joint and unique information in multi-block data	[66]
Pre-processing optimization and fusion	2nd order data	Sequential pre-processing through orthogonalization (SPORT)	• Based on sequential orthogonalized partial least-squares regression • Includes a combination of partial least-squares regression and sequential orthogonalization step to extract complementary latent variables from multi-block data	• Allows a sequential fusion of pre-processing techniques • Complementary information is modelled • Pre-processing techniques selection can be performed • Any pre-processing techniques having zero LVs can be discarded, thus, leading to pre-processing selection	[64]
		Parallel pre-processing through orthogonalization (PORTO)	• Based on parallel orthogonalized partial least-squares regression • Includes a combination of partial least-squares regression, canonical correlation analysis and orthogonalization step to extract common and distinct latent variables from multi-block data	• Allows parallel fusion of pre-processing technique without the need of defining the order • Allows insight to the common and distinct information present in different pre-processing techniques which help in selecting a subset of pre-processing techniques	[65]

10. Free software resources available for multi-block data analysis

Multi-block data analysis is a relatively new domain in chemometrics and software for performing the multi-block analysis are scarce. However, several groups around the world have published freely available codes so that the community can benefit from them. The first publicly available MATLAB-based toolbox is from the University of Copenhagen, Denmark (http://www.models.life.ku.dk/~courses/MBtoolbox/mbtmain.htm), which, having been last revised in 2001, focuses only on the two data fusion approaches that were most popular at that time, i.e., multi-block principal component analysis and multi-block partial least squares regression. The second toolbox is that by NOFIMA for multi-block regression by parallel and sequential partial least-squares regression [52]. Both toolboxes provide command line functionalities (within the MATLAB environment) and consist of a limited number of tools. There is also a basic graphical user interface (GUI) available for performing multi-block component analysis in the domain of behavioural research [67]. However, this GUI only proposes principal component analysis on each data block separately, simultaneous component analysis, and cluster-wise simultaneous component analysis for data exploration. Recently, a new graphical user interface, the MBA-GUI (freely available at: https://github.com/puneetmishra2/Multi-block.git) has been made available which integrates several advanced multi-block techniques related to data exploration, regression, variable selection, pre-processing selection and fusion [26]. A python library called ‘mbpls 1.0.4” was also recently developed for performing multi-block PLS and is available at https://pypi.org/project/mbpls/.

11. Some comparative examples for predictive multi-block modelling

Thanks to the chemometrics developments in recent years, multi-block techniques are now available to perform both exploratory and predictive data modelling. However, there are so many new tools, as well as extensions of standard single-block chemometric techniques (such as, MCR, PLS), that it is becoming difficult to find the best solutions to start with when dealing with a new problem/application. However, in chemometrics, as in other scientific fields, it is difficult if not impossible to univocally define what could be the best technique in an absolute sense. First of all, as the term "best", itself, can assume several meanings depending on the specific application (e.g., more robust, less impacted by interferents, more accurate, and so on). A well, different techniques have advantages and disadvantages for different data type. For this reason, these different tools could even be ensembled to get a better understanding of the data and solve the background challenges. Recently, several works have tried to compare the performances of different multi-block methods to achieve a deeper understanding of their characteristics, similarities and dissimilarities, in the light of the practical use of those techniques [3, 7, 20, 21, 35]. Since most of these works focus on exploratory approaches (i.e., to symmetric data fusion), while to the authors' knowledge, so far, no research literature provides a comparative overview of the predictive approaches, hence, three such practical comparisons are provided in the following section. The comparisons are based on the pear data set where a total of 231 pear fruit were measured with two complementary near-infrared (NIR) spectral sensors covering the spectral ranges of ~700-1050 nm and ~1050 to 1600 nm, respectively. The reference property was the moisture content (MC) measured with hot-air oven drying technique [68]. The samples were divided into a calibration set and an independent test set of 190 and 41 individuals, respectively. The data set used in these examples has already been published and it is used here for demonstration purposes only. More details on sampling

526 and reference analysis can be accessed in the original publication related to this data [68]. Out
527 of the three comparisons, the first is the comparative overview of PLS and two multiblock PLS
528 methods (SO-PLS and PO-PLS) [52], the second example is the comparison of two multi-block
529 pre-processing fusion approaches, i.e., SPORT [64] and PORTO [65], and the third is an
530 example of multi-block variable selection with SO-CovSel [55].

531 ### 11.1. Comparison of PLS, MB-PLS, SO-PLS and PO-PLS

532 A summary of the performances of the different approaches to model the spectral data from
533 multiple complementary NIR sensors is shown in Fig. 6. As a baseline, the results of PLS
534 modeling of the spectra from sensor 1 only (the best model on the individual matrices) was
535 added to show that this block alone is not sufficient to achieve an error as low as the one
536 obtained through the fusion of the complementary information in the multiple sensors. The
537 PLS model built on sensor 1 data only with 7 LVs (optimized using 5-fold cross-validation)
538 (Fig. 6A) attained a root mean square error (RMSEP) of 0.53 % to predict MC. Furthermore,
539 when the data from two sensors was concatenated along the variable direction and a new PLS
540 model was developed (Fig. 6B), the RMSEP was slightly increased, from 0.53 to 0.56 %, SO-
541 PLS also showed a similar RMSEP to that of PLS on concatenated spectral data (MB-PLS),
542 but by extracting complementary information from two sensors. On the other hand, in this case,
543 PO-PLS resulted in the lowest RMSEP (0.47 %). Furthermore, PO-PLS obtained this superior
544 performance by partitioning the common and unique information in the two sensors. However,
545 a key point to note is that the performance of MB-PLS was poorer than that of the SO-PLS
546 approach (i.e., a RMSEP of 0.52 %). Methods like SO-PLS and PO-PLS allow efficient
547 modelling of different data matrices and can lead to more accurate predictions than MB-PLS.
548 Advanced multi-block methods also bring added values such as a better understanding of
549 background chemistry, which can also be noted in Fig. 7, where the regression coefficients
550 corresponding to MB-PLS (dotted blue line) and SO-PLS (dashed black line) are presented. It

can be noted that calculating MB-PLS resulted in a model with higher absolute coefficients for sensor 1 data and several key features in the spectral range of sensor 2 are poorly modelled, e.g. at 1400 nm, which corresponds to the H_2O overtones directly related to the moisture [69]. A main challenge with MB-PLS is the need to perform proper scaling of the data, but that is not the case with methods like SO-PLS as they treat each data block sequentially [8], thus avoiding any negative effect of different data scales.

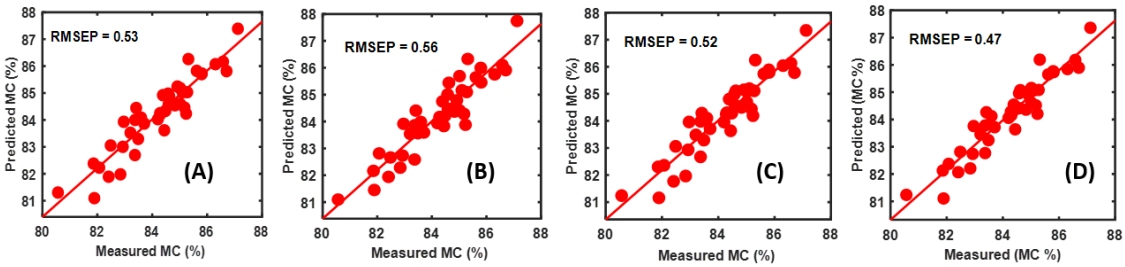


Fig. 6: Pear data set – Comparison of model performances for the prediction of moisture content (MC). Predicted vs observed MC values (test set) for: (A) the best PLS model on individual blocks; (B) MB-PLS; (C) SO PLS; and (D) PO-PLS.

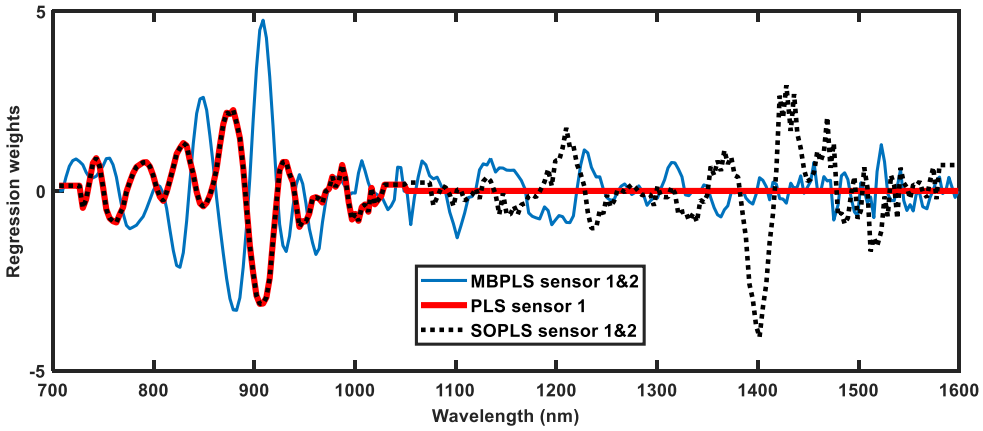


Fig. 7:. Pear data set – Comparison of models for the prediction of moisture content (MC). Regression coefficients for the PLS model built on sensor 1 data only (continuous thick red line), MB-PLS on concatenated data from sensors 1 and 2 (continuous thin blue line), and SOPLS on sensor 1 and 2 data (dotted black line; the coefficient vectors for the two separate PLS regressions involved have been concatenated, for a better visualization).

11.2. Comparison of pre-processing fusion approaches

Pre-processing selection in chemometrics is a challenging task, where a lot of time and resources are usually spent with the aim of achieving optimal pre-processing combinations [70]. However, such an approach can be considered old-fashioned due to emergence of new ensemble pre-processing fusion approaches [70] and especially the multi-block data analysis inspired methods such as SPORT [64] and PORTO [65]. A comparison of the use of SPORT and PORTO on the pear data set already described is shown in Fig. 8, as an example. The data used in this analysis are only those from sensor 1, although SPORT and PORTO can both deal with simultaneous multi-sensor multi-processing. Furthermore, only two data blocks, i.e. raw data and data pre-processed by 2^{nd} derivative, are used for the demonstration. The main thing to note is that using 2^{nd} derivative only (best individual pre-processing) results in a higher value of the RMSEP (0.53 %) (Fig. 8A), whereas modelling with both the raw and 2^{nd} derivative pre-processed data gave a lower RMSEP, 0.50 % for SPORT (Fig. 8B) and 0.49 % with PORTO (Fig. 8C).

The good performance from the combined used of raw and 2^{nd} derivative pre-processed data is not a surprise from a fruit property modelling point of view. This is because the NIR spectra of fresh fruit are a mixture of absorption and scattering profiles, the absorption can usually be related to broad peaks in the NIR data, whereas the scattering properties are expressed as the additive and multiplicative effects [71]. The chemical and physical properties of fruits are correlated to both the effect of scattering due to fruit cellular structure (which differs with the ripening stage of fruit) and absorption present in NIR data. Hence, doing a 2^{nd} derivative estimation may eliminate the global intensity differences related to scattering, and therefore, may remove some useful information related to fruit properties. The multi-block approaches compensate this loss of information by first modeling the 2^{nd} derivative and then modelling the

remaining variation within the raw data. An example of the complementary modelling performed by SO-PLS is shown in Fig. 9, where the regression vectors for the 2nd derivative pre-processed (solid blue line) and raw data (dashed red line) are shown. It can be seen that the main features of the 2nd derivative are the peaks related to overtones of -OH, -CH and -NH [69], whereas the main information captured from the raw data is the global shape of the spectrum which is an indication of the scattering information.

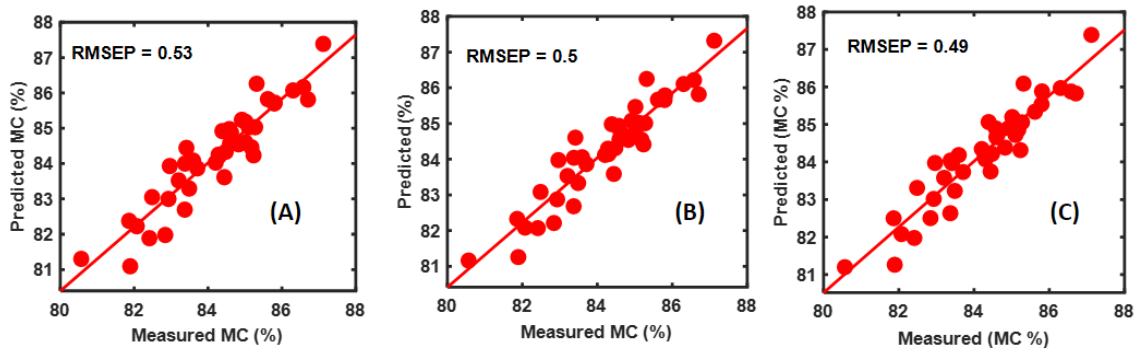


Fig. 8: Pear data set – Comparison of model performances for the prediction of moisture content (MC). Predicted vs observed MC values (test set) for: (A) PLS (on data pre-processed with 2nd derivative only), (B) SPORT, and (C) PORTO models.

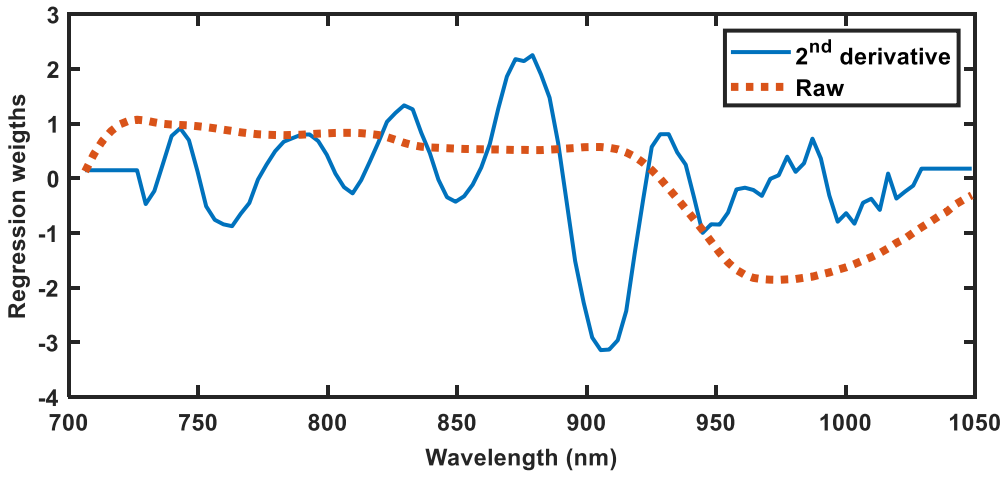


Fig. 9: Pear data set – SPORT model. Regression coefficients for the two blocks of data (raw

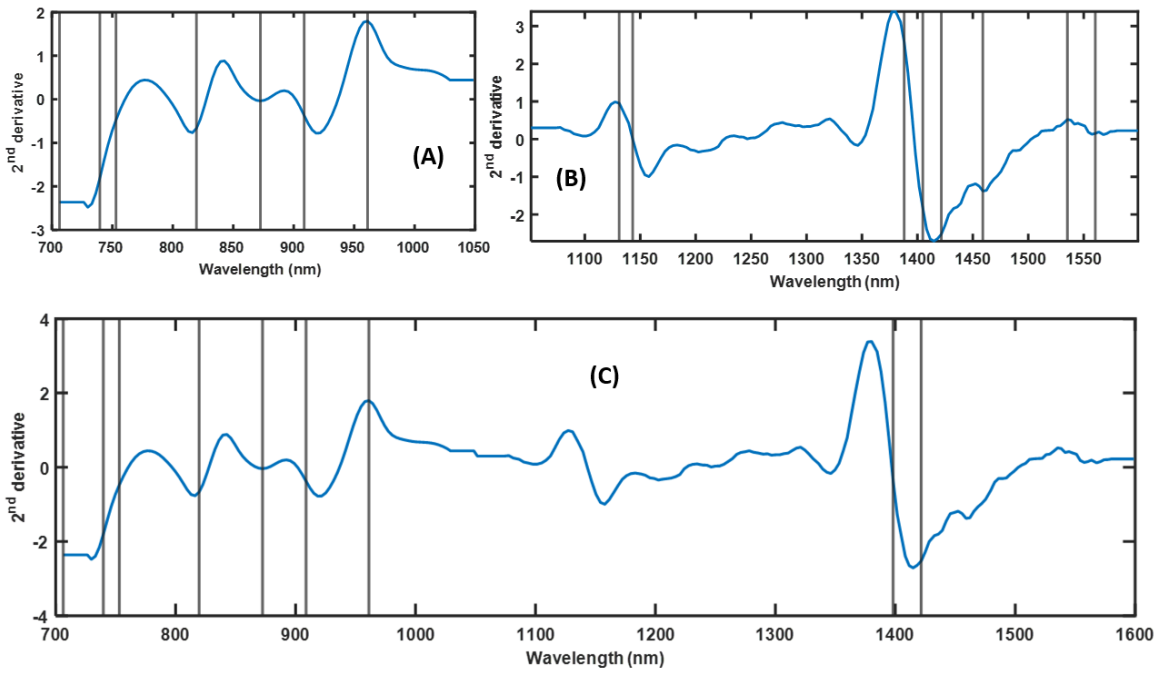
606 *and 2nd derivative pre-processed).*

607

608 ### 11.3. Selecting variables in multi-block scenario

609 Variable selection is useful in chemometrics and is even challenging when the data is multi-
610 block. Particularly, the challenge arises when the redundant information is present in multiple
611 data blocks and the aim is to just use the complementary information that improves the
612 predictive performances of the model. In such a case, new multi-block methods such as SO-
613 CovSel can be used efficiently. Fig. 10 C shows the results of performing SO-CovSel on the
614 two-sensors pear data set. For the sake of comparison, CovSel analysis on individual blocks is
615 also presented and the selected variable for sensor 1 and sensor 2 are shown in Fig. 10 A and
616 B, respectively. Separate CovSel analyses on sensor 1 and sensor 2 data selected 7 and 8
617 wavelengths, respectively. However, most of the wavelengths are related to overtones of
618 similar chemical bonds and, therefore, carry redundant information. In the case of SO-CovSel,
619 due to such redundancy, only 2 bands are selected from sensor 2 and, nevertheless, the RMSEP
620 is reduced from 0.58 % (best individual model) to 0.55 % (Fig. 11).

621



622

623 *Fig. 10: Pear data set - Comparison of the results of variable selection on the individual blocks*

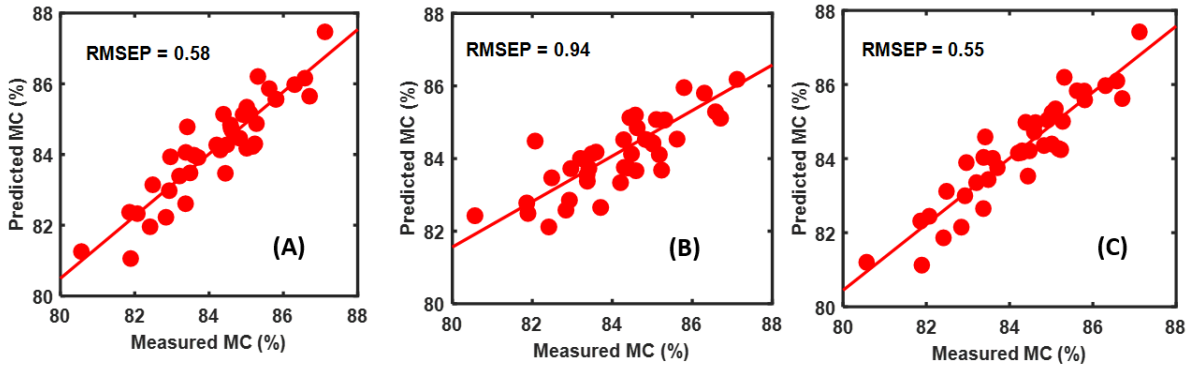
624 *and in the multi-block scenario. (A) Variables selected from sensor 1 data through single-block*

625 *CovSel analysis, (B) Variables selected from sensor 2 data through single-block CovSel*

626 *analysis, and (C) Variables jointly selected from sensor 1 and sensor 2 data through the multi-*

627 *block SO-CovSel approach.*

628



629

630 *Fig. 11: Pear data set - Comparison of the results of variable selection on the individual blocks*

631 *and in the multi-block scenario for the prediction of moisture prediction in pear fruit. Predicted*

632 *vs measured values of moisture content based on (A) Single-block CovSel analysis on sensor 1*

633 *data, (B) Single-block CovSel analysis on sensor 2 data, and (C) Multi-block SO-CovSel*

approach.

12. Concluding remarks

Multi-block data analysis in chemometrics is gaining increasing attention and the development of new methods in recent years has been rapid. Analytical chemistry can directly benefit from these new techniques to explore and combine data from multiple sources. Due to advances in sensor and computing technologies, multi-source data are now frequently encountered. Multi-block methods are available for diverse tasks such as exploratory data analysis, predictive modelling, variable selection, pre-processing optimization, and calibration transfer. There are also methods available to explore the multi-block data generated by designed experiments, which is often the case with lab-based classical analytical chemistry experiments. The main benefit of multi-block data analysis compared to the standard chemometric methods is that they allow a detailed understanding of common and distinct information present in different data-blocks or data generated from multiple sources. Recently, free software tools such as the MBA-GUI have been made available to the scientific community to explore the possibilities of multi-block data analysis. It can be expected that the future trend will be an exponential increase in the applications of multi-block data analysis methods in analytical chemistry to combine in an optimal way multiple sources of data. Another important direction that can be foreseen is the development of interactive data visualization tools [72] dedicated to multi-block data analysis, which will allow even non-experts to have a better comprehension of their data.

Declaration of Interest

None

References

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

[2] L. Zhou, C. Zhang, Z. Qiu, Y. He, Information fusion of emerging non-destructive analytical techniques for food quality authentication: A survey, TrAC Trends in Analytical Chemistry, 127 (2020) 115901.

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

[4] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, Chemometrics and Intelligent Laboratory Systems, 141 (2015) 58-67.

[5] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: Separating common and unique information in several data blocks, Food Quality and Preference, 24 (2012) 8-16.

[6] M. Ramos-Barberán, M.V. Hinojosa-Ramos, J. Ascencio-Moreno, F. Vera, O. Ruiz-Barzola, M.P. Galindo-Villardón, Batch process control and monitoring: a Dual STATIS and Parallel Coordinates (DS-PC) approach, Production & Manufacturing Research, 6 (2018) 470-493.

[7] R. Vitale, O.E. de Noord, J.A. Westerhuis, A.K. Smilde, A. Ferrer, Divide et impera: How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding, Journal of Chemometrics, n/a (2020) e3266.

[8] M.P. Campos, M.S. Reis, Data preprocessing for multiblock modelling – A systematization with new methods, Chemometrics and Intelligent Laboratory Systems, 199 (2020) 103959.

[9] Z. Ge, Review on data-driven modeling and monitoring for plant-wide industrial processes, Chemometrics and Intelligent Laboratory Systems, 171 (2017) 16-25.

[10] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, Analytical and Bioanalytical Chemistry, 409 (2017) 5891-5899.

[11] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part II: modeling, validation, and applications, Analytical and Bioanalytical Chemistry, 410 (2018) 6691-6704.

[12] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods, 6 (2014) 2812-2831.

[13] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta, 185 (1986) 1-17.

[14] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemometrics and Intelligent Laboratory Systems, 118 (2012) 62-69.

[15] T. Mehmood, S. Sæbø, K.H. Liland, Comparison of variable selection methods in partial least squares regression, Journal of Chemometrics, n/a (2020) e3226.

[16] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, CovSel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy, Chemometrics and Intelligent Laboratory Systems, 106 (2011) 216-223.

[17] A.K. Smilde, J.A. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, Journal of Chemometrics, 17 (2003) 323-337.

[18] A. Biancolillo, T. Næs, The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions. In M. Cocchi (Ed.), Data Fusion Methodology and Applications, Elsevier, 2019, pp. 157-177.

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

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

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

[22] S. Zhu, Z. Song, S. Shi, M. Wang, G. Jin, Fusion of Near-Infrared and Raman Spectroscopy for In-Line Measurement of Component Content of Molten Polymer Blends, Sensors (Basel, Switzerland), 19 (2019) 3463.
[23] S.E. Barnes, E.C. Brown, M.G. Sibley, H.G.M. Edwards, I.J. Scowen, P.D. Coates, Vibrational Spectroscopic and Ultrasound Analysis for In-Process Characterization of High-Density Polyethylene/Polypropylene Blends during Melt Extrusion, Applied Spectroscopy, 59 (2005) 611-619.
[24] K. Haroon, A. Arafeh, S. Cunliffe, P. Martin, T. Rodgers, Ć. Mendoza, M. Baker, Comparison of Individual and Integrated Inline Raman, Near-Infrared, and Mid-Infrared Spectroscopic Models to Predict the Viscosity of Micellar Liquids, Applied Spectroscopy, 74 (2020) 819-831.
[25] C. Assis, H.V. Pereira, V.S. Amador, R. Augusti, L.S. de Oliveira, M.M. Sena, Combining mid infrared spectroscopy and paper spray mass spectrometry in a data fusion model to predict the composition of coffee blends, Food Chemistry, 281 (2019) 71-77.
[26] 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.
[27] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, Pre-processing Methods. n: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics (Second Edition), vol. 3, Elsevier, Oxford, 2020, pp.1-75.
[28] M.P. Campos, R. Sousa, A.C. Pereira, M.S. Reis, Advanced predictive methods for wine age prediction: Part II – A comparison study of multiblock regression approaches, Talanta, 171 (2017) 132-142.
[29] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, Journal of Chemometrics, 12 (1998) 301-321.
[30] M. Hanafi, A. Kohler, E.-M. Qannari, Connections between multiple co-inertia analysis and consensus principal component analysis, Chemometrics and Intelligent Laboratory Systems, 106 (2011) 37-40.
[31] M. Hanafi, E.M. Qannari, B. Jaillais, Multi-Block and Three-Way Data Analysis. In: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics (Second Edition), vol. 3, Elsevier, Oxford, 2020, pp. 341-358.
[32] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, Journal of Chemometrics, 10 (1996) 463-482.
[33] E.M. Qannari, I. Wakeling, P. Courcoux, H.J.H. MacFie, Defining the underlying sensory dimensions, Food Quality and Preference, 11 (2000) 151-154.
[34] M. Hanafi, G. Mazerolles, E. Dufour, E.M. Qannari, Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques, Journal of Chemometrics, 20 (2006) 172-183.
[35] V. Cariou, D. Jouan-Rimbaud Bouveresse, E.M. Qannari, D.N. Rutledge, ComDim Methods for the Analysis of Multiblock Data in a Data Fusion Perspective. In M. Cocchi (Ed.), Data Fusion Methodology and Applications, Elsevier, 2019, pp. 179-204.
[36] D. Jouan-Rimbaud Bouveresse, R.C. Pinto, L.M. Schmidtke, N. Locquet, D.N. Rutledge, Identification of significant factors by an extension of ANOVA–PCA based on multi-block analysis, Chemometrics and Intelligent Laboratory Systems, 106 (2011) 173-182.
[37] A. El Ghaziri, V. Cariou, D.N. Rutledge, E.M. Qannari, Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K + 1) datasets, Journal of Chemometrics, 30 (2016) 420-429.
[38] V. Cariou, E.M. Qannari, D.N. Rutledge, E. Vigneau, ComDim: From multiblock data analysis to path modeling, Food Quality and Preference, 67 (2018) 27-34.
[39] M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behavior Research Methods, 45 (2013) 822-833.

[40] K. Van Deun, I. Van Mechelen, L. Thorrez, M. Schouteden, B. De Moor, M.J. van der Werf, L. De Lathauwer, A.K. Smilde, H.A.L. Kiers, DISCO-SCA and Properly Applied GSVD as Swinging Methods to Find Common and Distinctive Processes, PLOS ONE, 7 (2012) e37840.
[41] J. Trygg, O2-PLS for qualitative and quantitative analysis in multivariate calibration, Journal of Chemometrics, 16 (2002) 283-293.
[42] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES, The annals of applied statistics, 7 (2013) 523-542.
[43] E. Acar, M.A. Rasmussen, F. Savorani, T. Næs, R. Bro, Understanding data fusion within the framework of coupled matrix and tensor factorizations, Chemometrics and Intelligent Laboratory Systems, 129 (2013) 53-63.
[44] E. Acar, E.E. Papalexakis, G. Gürdeniz, M.A. Rasmussen, A.J. Lawaetz, M. Nilsson, R. Bro, Structure-revealing data fusion, BMC Bioinformatics, 15 (2014) 239.
[45] I. Gaynanova, G. Li, Structural learning and integrative decomposition of multi-view data, Biometrics, 75 (2019) 1121-1132.
[46] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.-J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics, 21 (2005) 3043-3048.
[47] R. Tauler, M. Maeder, A. de Juan, Multiset Data Analysis: Extended Multivariate Curve Resolution. In: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics (Second Edition), vol.2, Elsevier, Oxford, 2020, pp. 305-336.
[48] L.E. Wangen, B.R. Kowalski, A multiblock partial least squares algorithm for investigating complex chemical systems, Journal of Chemometrics, 3 (1989) 3-20.
[49] S.J. Qin, S. Valle, M.J. Piovoso, On unifying multiblock analysis with application to decentralized process monitoring, Journal of Chemometrics, 15 (2001) 715-742.
[50] S. Wold, PLS Modeling with Latent Variables in Two Or More Dimensions, 1987.
[51] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, Journal of Chemometrics, 25 (2011) 441-455.
[52] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, Chemometrics and Intelligent Laboratory Systems, 124 (2013) 32-42.
[53] T. Skov, D. Ballabio, R. Bro, Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks, Analytica Chimica Acta, 615 (2008) 18-29.
[54] A. Biancolillo, K.H. Liland, I. Måge, T. Næs, R. Bro, Variable selection in multi-block regression, Chemometrics and Intelligent Laboratory Systems, 156 (2016) 89-101.
[55] A. Biancolillo, F. Marini, J.-M. Roger, SO-CovSel: A novel method for variable selection in a multiblock framework, Journal of Chemometrics, 34 (2020) e3120.
[56] B. Galindo-Prieto, J. Trygg, P. Geladi, A new approach for variable influence on projection (VIP) in O2PLS models, Chemometrics and Intelligent Laboratory Systems, 160 (2017) 110-124.
[57] B. Galindo-Prieto, P. Geladi, J. Trygg, Multiblock variable influence on orthogonal projections (MB-VIOP) for enhanced interpretation of total, global, local and unique variations in OnPLS models, arXiv preprint arXiv:2001.06530, (2020).
[58] S. Park, E. Ceulemans, K. Van Deun, Sparse common and distinctive covariates regression, Journal of Chemometrics, n/a (2020) e3270.
[59] A. Biancolillo, T. Næs, R. Bro, I. Måge, Extension of SO-PLS to multi-way arrays: SO-N-PLS, Chemometrics and Intelligent Laboratory Systems, 164 (2017) 113-126.
[60] A.K. Smilde, J.A. Westerhuis, R. Boqué, Multiway multiblock component and covariates regression models, Journal of Chemometrics, 14 (2000) 301-331.
[61] 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 Biology and Technology, 168 (2020) 111271.

[62] 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.
[63] P. Mishra, A. Nordon, J.-M. Roger, Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques, Journal of Pharmaceutical and Biomedical Analysis, (2020) 113684.
[64] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometrics and Intelligent Laboratory Systems, 199 (2020) 103975.
[65] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy, Chemometrics and Intelligent Laboratory Systems, (2020) 104190.
[66] T. Skotare, D. Nilsson, S. Xiong, P. Geladi, J. Trygg, Joint and Unique Multiblock Analysis for Integration and Calibration Transfer of NIR Instruments, Analytical Chemistry, 91 (2019) 3516-3524.
[67] K. De Roover, E. Ceulemans, M.E. Timmerman, How to perform multiblock component analysis in practice, Behavior Research Methods, 44 (2012) 41-56.
[68] 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.
[69] K.B. Walsh, J. Blasco, M. Zude-Sasse, X. Sun, Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use, Postharvest Biology and Technology, 168 (2020) 111246.
[70] 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 in Analytical Chemistry, (2020) 116045.
[71] R. Lu, R. Van Beers, W. Saeys, C. Li, H. Cen, Measurement of optical properties of fruits and vegetables: A review, Postharvest Biology and Technology, 159 (2020) 111003.
[72] T. Skotare, R. Sjögren, I. Surowiec, D. Nilsson, J. Trygg, Visualization of descriptive multiblock analysis, Journal of Chemometrics, 34 (2020) e3071.

Declaration of interests

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