Analytica Chimica Acta 1250 (2023) 340957

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Sequentially orthogonalized canonical partial least squares for improved multiple responses modeling in multiblock data sets

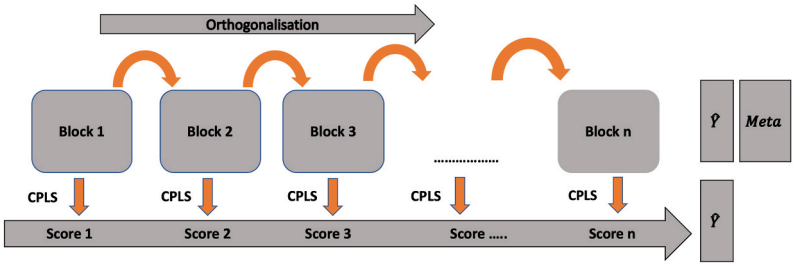
Puneet Mishra

Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands

HIGHLIGHTS

- The multiblock method combining sequential modelling and canonical PLS is evaluated.
- The method handles multiple responses more efficiently than PLS2.
- The method allows using meta information to improve subspace extraction.
- The method was tested on wide multiple responses prediction datasets.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Prof. L. Buydens

Keywords:
Data Fusion
Multiple responses
Complementary
Multivariate

ABSTRACT

Multiblock data sets and modeling techniques are widely encountered in the chemometric community. Although the currently available techniques, such as sequential orthogonalized partial least squares (SO-PLS) regression are mainly focused on the prediction of a single response and deal with the multiple response(s) case using PLS2 type approach. Recently, a new approach called canonical PLS (CPLS) was proposed for extracting the subspaces efficiently for multiple response(s) cases, supporting both regression and classification. 'Efficiently' here means more information in fewer latent variables. This work suggests a combination of SO-PLS and CPLS, sequential orthogonalized canonical partial least squares (SO-CPLS), to model multiple response(s) for multiblock data sets. The cases of SO-CPLS for modeling multiple response(s) regression and classification were demonstrated on several data sets. Also, the capability of SO-CPLS to incorporate meta-information related to samples for efficient subspace extraction is demonstrated. Furthermore, a comparison with the commonly used sequential modeling technique, called sequential orthogonalized partial least squares (SO-PLS), is also presented. The SO-CPLS approach can benefit both the multiple response(s) regression and classification modeling and can be of high importance when meta-information such as experimental design or sample classes is available.

1. Introduction

Multiblock data analysis techniques are emerging as a potential tool in chemometrics for combining data from multiple sources [1,2]. For example, data can come from multiple analytical techniques measured on same samples [3–5], data generated as a combination of analytical and sensory experiments [6], data generated by measuring the same samples in different physical forms [7], or data that has undergone several preprocessing approaches [8,9]. These are just some examples, but essentially, any data that uses more than one modality to measure

E-mail address: puneet.mishra@wur.nl.

https://doi.org/10.1016/j.aca.2023.340957
Received 15 August 2022; Received in revised form 22 December 2022; Accepted 8 February 2023
Available online 9 February 2023

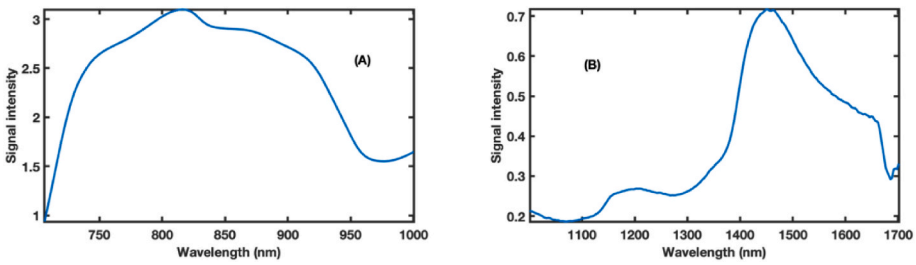


Fig. 1. The multiblock pear data set. (A) NIR spectra, and (B) SWIR spectra.

samples can be considered multiblock [2,10]. Depending on the modality, the data blocks can either be two-dimensional (2D) matrices or *n*-dimensional (*n*D) arrays [2,11].

In the domain of chemometrics, extensive developments have taken place in terms of analyzing data in the form of 2D matrices [10] as well as *n*D arrays [12]. The key focus in the chemometric domain is to extract the subspaces that are highly relevant for predicting the response variable(s). Subspace modeling techniques are of particular interest as the multivariate data generated using several analytical techniques are highly collinear. Therefore, prior transformation of data to an uncorrelated space allows for generalized conclusions and models. In that regard, some common subspace techniques applied in the chemometrics domain are inspired from the principal component analysis (PCA) [13] and partial least squares analysis (PLS) [14,15]. PCA allows for unsupervised extraction of subspaces that capture subspaces carrying high variance in data, while PLS allows supervised extraction of subspaces that carry high covariance with the response variable(s). For n-dimensional data, the extension of PCA such as Tucker3, with a special case Parallel Factor Analysis, and extensions of PLS such as *n*-way PLS are commonly used [12].

In the multiblock sub-domain of chemometrics, subspace-based approaches are commonly developed to allow information fusion at the subspace level. One of the key advantage of these approaches is that they enable the understanding of the background science of the samples or processes in terms of scores and loading plots, which are commonly extracted by bilinear decomposition methods like PCA and PLS. In the domain of supervised multiblock modeling, several extensions of PLS-based approaches have been developed, such as the multiblock PLS (MB-PLS) [16] which aims to extract global scores using hierarchal PLS approach, the sequential orthogonalized PLS (SO-PLS) [17] which sequentially extract the complementary scores from different data blocks, response-oriented sequential alternation (ROSA) [18] which extracts the scores in sequential order while competing to minimize the residuals; and parallel orthogonalized PLS (PO-PLS) [6,17], which aims to extract common and distinct scores from different data blocks.

All of the various multiblock PLS methods are valuable tools in chemometrics for multiblock data analysis, each with their own advantages and disadvantages. For example, the MB-PLS approach is highly sensitive to the scale of data blocks [19] while SO-PLS models each block individually and is not as sensitive [20]. SO-PLS only models complementary information following the sequential order of data blocks [8] and is dependent on the block order, while PO-PLS allows modeling of common and distinct information without concern for the block order [9]. Both SO-PLS and PO-PLS require significant computation costs for model optimization, whereas the new ROSA [18] technique is a computationally fast alternative. However, ROSA may get stuck in local minima due to the greedy approach to model components extraction [21]. Numerous applications of different PLS-based multiblock modeling techniques can be found elsewhere [2–5,8,9,22–26].

Despite significant efforts to develop faster multiblock techniques that are independent of scale and block order, some properties of an ideal multiblock technique remain unresolved. One of the main questions is how to develop multiblock predictive techniques that can efficiently handle multiple response(s). This is particularly relevant when extracting subspaces for discrete classes, as in classification problems. Another outstanding challenge is to develop a multiblock technique that can incorporate meta-information into the model to facilitate the efficient extraction of subspaces. Meta-information consists of one or more additional variables that describe the training data, such as experimental conditions, mixture proportions, or other analytical data collected during experiments. However, this same meta-information is often unavailable in real-world or production environments due to practical or economic limitations.

The current PLS based multiblock techniques are capable of modeling multiple response(s) by using a multiple responses PLS approach known as the PLS2 [27]. However, in single block data analysis, it has been demonstrated that the PLS2 approach may not be highly efficient for extracting subspaces for predictive modeling [28]. In the domain of chemometrics, a new technique called canonical PLS (CPLS) [28] has recently been proposed as a solution for two outstanding challenges i.e., modeling multiple response(s) and incorporating meta-information in models for improving subspace extraction. The CPLS approach is a unique combination of PLS and canonical component analysis and can be more effective than ordinary PLS2 for multiple responses scenarios [28], meaning that it can extract more information in a smaller number of latent variables (LVs).

The purpose of this study is to introduce a novel extension of the CPLS methods for sequential multiblock modeling, called SO-CPLS, which is used to model multiple responses in multiblock data sets. The effectiveness of SO-CPLS for modeling both regression and classification cases is demonstrated through several data sets. Moreover, the ability of SO-CPLS to incorporate sample meta-information to extract subspaces efficiently is also demonstrated. The study also presents a comparison of SO-CPLS and SO-PLS, which is based on PLS2 decomposition.

2. Method

The SO-CPLS method is a sequential extension of CPLS method for handling multiblock data sets. The basic framework behind the SO-CPLS is the same as the SO-PLS approaches, however, the main difference is the extraction of the scores which are extracted using the CPLS instead of PLS2. For cases with n blocks $(\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots \boldsymbol{X_n})$ used to estimate a multiple responses data matrix $\boldsymbol{Y}$, the sequential algorithm can be understood in the steps shown below. Predictor data blocks $(\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots \boldsymbol{X_n})$ and responses were mean-centered. No further preprocessing of responses was performed as the scale and variances of different responses were similar.

Algorithm. steps for SO-CPLS

1. *loop* $i = 1{:}n$
2. Multiple responses $\boldsymbol{Y}$ is fitted to the $\boldsymbol{X_i}$ using the CPLS

$$\boldsymbol{Y} = \boldsymbol{X_i B_{X_i}} + \boldsymbol{E_i} = \boldsymbol{T_{X_i} Q_{X_i}^T} + \boldsymbol{E_i} \quad (1)$$

3. $\boldsymbol{X_{i+1}}$ is orthogonalized with respect to the scores extracted from the $\boldsymbol{X_1}, \ldots \boldsymbol{X_i}$ by CPLS fitting

$$\boldsymbol{X_{i+1}^{orth}} = \left[\boldsymbol{I} - \boldsymbol{T_{X_{1:i}}}(\boldsymbol{T_{X_{1:i}}^T T_{X_{1:i}}})^{-1}\boldsymbol{T_{X_{1:i}}^T}\right]\boldsymbol{X_{i+1}} \quad (2)$$

4. The orthogonalized $\boldsymbol{X_{i+1}}$ is used to predict the $\boldsymbol{Y}$ -residuals obtained from step 1 using the CPLS fitting

$$\boldsymbol{E_i} = \boldsymbol{X}_{i+1}^{orth}\boldsymbol{B_{X_{i+1}^{orth}}} + \boldsymbol{E_{i+1}} = \boldsymbol{T_{X_{i+1}^{orth}} Q_{X_{i+1}^{orth}}^T} + \boldsymbol{E_{i+1}} \quad (3)$$

5. The full predictive model is calculated summing up results from step 1. and step

$$\hat{Y} = \boldsymbol{X_1 B_{X_1}} + \boldsymbol{X}_2^{orth}\boldsymbol{B_{X_2^{orth}}} + \boxtimes + \boldsymbol{X}_n^{orth}\boldsymbol{B_{X_n^{orth}}} = \boldsymbol{T_{X_1} Q_{X_1}^T} + \boldsymbol{T_{X_2^{orth}} Q_{X_2^{orth}}^T} + \boxtimes + \boldsymbol{T_{X_n^{orth}} Q_{X_n^{orth}}^T} \quad (4)$$

6. *loop end*

$\boldsymbol{B}, \boldsymbol{T}$ *and* $\boldsymbol{Q}$ are the regression coefficients, $\boldsymbol{X}$-scores, and $\boldsymbol{Y}$-loadings. One of the main parts of latent space modeling is extracting the optimal number of LVs to achieve generalized models. In the case of sequential orthogonalized models, a common approach is to explore all possible combinations of LVs (within a fixed maximum) using either cross-validation approaches [8] or using external validation set [29]. The optimal complexity is then defined by inspecting the root mean squared error of cross-validation (RMSECV) or root mean squared error of validation (RMSEV). Then, a SO-CPLS model using optimal LVs combination can be calibrated on the calibration set and used on the test set.

The key difference between the CPLS and a typical approach to PLS2 type of analysis is the estimation of the loading weight vector. In a typical PLS2 analysis, the covariance matrix is estimated as Eq. (5)

$$\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{Y} \quad (5)$$

As the second step, a rank 1 SVD (Singular Value Decomposition) approximation of $\boldsymbol{W}$ provides the loading weight vector (with unit length), which can later be used to estimate the scores. The scores can then be used for matrix deflation. Later the same step is repeated until desired number of components are extracted. However, performing a rank one SVD (Singular Value Decomposition) approximation on the covariance matrix $\boldsymbol{W}$, acts as unsupervised with respect to the computation of the linear combination of the $\boldsymbol{W}$-columns for loading weight estimation. On the other hand, in the case of CPLS, the loading weights are estimated by maximizing the canonical correlation between the matrix $\boldsymbol{Z}$ and the responses $\boldsymbol{Y}$, where $\boldsymbol{Z} = \boldsymbol{XW}$. Note that during the estimation of the $\boldsymbol{Z}$ matrix the covariance matrix estimated in Eq. (5) is already being used. However, during the step of canonical correlation maximization between the matrix $\boldsymbol{Z}$ and the responses $\boldsymbol{Y}$, the $\boldsymbol{Y}$ information is used one more time to have loading weights more aggressive towards the prediction of $\boldsymbol{Y}$ compared to the loading weight estimation in the PLS2 approach. In other words, the CPLS approach can be considered as estimating the loading weight as a supervised linear combination of the $\boldsymbol{W}$-columns, which is otherwise performed in an unsupervised way in the case of direct SVD on the covariance matrix. Hence, compared to the traditional PLS2 analysis, one should expect that the CPLS models requires fewer components which allows simplified model interpretation.

The above step does not include any step of including the meta-information about the samples, however, the inclusion of the meta-information is in the loading weight estimation step [28]. The meta-information can be supplied as a column vector (dummy vectors [0,0, 0 … 1, 1, 1] for discrete information or continuous response vector for a continuous meta variable) to the CPLS and the CPLS utilizes this information while estimation of the weight's vectors from the canonical correlation analysis [28]. Note that the meta-information must be supplied as the column wise concatenation with the response. For example, if $\boldsymbol{Y}_{add}$ is the meta-information available and $\boldsymbol{Y}_{prim}$ be the actual response, then the response matrix for the CPLS becomes $\boldsymbol{Y} = [\boldsymbol{Y}_{prim}\ \boldsymbol{Y}_{add}]$. Now with $\boldsymbol{Y}$ composed of both the primary and the additional information, the covariance matrix can be estimated as $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{Y}$, and the corresponding transformed data $\boldsymbol{Z} = \boldsymbol{XW}$, followed by the maximization of the canonical correlation between $\boldsymbol{Z}$ and $\boldsymbol{Y}_{prim}$. During the $\boldsymbol{W}$ matrix estimation, the presence of $\boldsymbol{Y}_{add}$ will add extra columns to the extent that the $\boldsymbol{Y}_{add}$ contains useful information for the prediction of $\boldsymbol{Y}_{prim}$ that is also present in the $\boldsymbol{X}$ predictors. Note that the meta-information ($\boldsymbol{Y}_{\text{add}}$) is only needed in the modeling phase for making a more informed choice of subspace, this information is not needed at prediction time. More details on the inclusion of the meta-information in the CPLS modeling step can be gained in Ref. [28].

In practical scenarios for multiple responses modeling, one can assume that the responses can be of varying scales and variances. This can be a challenge while modeling with the traditional SO-PLS2 modeling approach as during the covariance estimation and later the rank one SVD (Singular Value Decomposition) approximation on the covariance matrix to estimate the loading weights are directly influenced by the

variances and scales of the different responses. Particularly the responses having higher variance and larger scales will have more influence on the estimation of the loading weight vector in the case of SO-PLS2 analysis. For SO-CPLS, the problem is eliminated as the canonical analysis step used in the CPLS is not influenced by the different variances and scale of data. In summary, the SO-CPLS is the ideal approach to model multiresponses data since it is unaffected by the different data scales and variances.

In the following part, the data sets and the specific analysis performed for each data set is explained.

3. Data sets and analysis

3.1. Moisture and soluble solids content prediction in pear fruit by fusing information from two portable spectrometers

The pear data set consists of spectral measurements performed on 240 pear fruit samples using two portable spectrometers. The aim of this study is to demonstrate the effectiveness of SO-CPLS for extracting variables for continuous response variable(s) and handling multiple responses better than the traditional PLS2-based SO-PLS. More details on the samples and experimental setup can be found in an earlier study [3]. Spectral measurements of the pear samples were conducted using a Vis-NIR portable spectrometer, Felix F-750 (Camas, WA, USA), and later using a SWIR spectrometer, DLP NIR Scan Nano (Texas Instrument, USA). Moisture content (MC) was estimated by weighing the samples before and after drying, at 80 °C for 24 h with FP 720, Binder GmbH, Tuttlingen, Germany. Soluble solids content (SSC) was measured using a handheld refractometer, HI 96801, Hanna Instruments Inc, Woonsocket, RI, USA. The spectral ranges used for multiblock modeling were 720–1000 nm for NIR and 1000–1700 nm for the SWIR spectrometer. In the pear data, MC and SSC showed a high correlation. The data were divided into 66.66% calibration and 33.33% test sets using the Kennard-Stone (KS) [30] algorithm on the response variable(s). The pear dataset was used to demonstrate the ability of SO-CPLS to extract variables for continuous response variable(s) and its superior capability to handle multiple responses.

3.2. Milk data set

The milk dataset was used as an additional example to showcase the ability of SO-CPLS to model continuous response variables. Unlike the pear dataset, the milk dataset features uncorrelated responses. Specifically, the milk dataset consisted of spectral data and reference measurements of protein and fat for 296 milk samples [31]. The spectral measurements were obtained using three different NIR spectral sensors, namely NIRONE 1.4, NIRONE 2.0, and NIRONE 2.5, from Spectral Engines (Helsinki, Finland). The spectral ranges for NIRONE 1.4, NIRONE 2.0, and NIRONE 2.5 were 1100-1400 nm, 1550-1950 nm, and 2000-2450 nm, respectively. Further details on the dataset and the reference analysis protocols for protein and fat can be found in [31]. The data were partitioned into 66.66% calibration and 33.33% test sets using the KS [30] algorithm on the response variable(s). Notably, in the milk dataset, fats and protein were uncorrelated, while total solids content was correlated with fats.

3.3. Near infrared and mid-infrared data set for multiple responses modeling of apricot

The apricot data set comprised NIR (800–2770 nm) and MIR (4000-650 cm^{-1}) speof 750 apricots [32], along with their corresponding soluble solids contents (%). Further details on the reference analysis can be obtained from Ref. [32]. The samples were obtained from four different maturity stages, and this information was used as meta-information to demonstrate the potential of SO-CPLS for improved subspace extraction. The data were partitioned into a 66.66% calibration set and a 33.33% test set using the KS [30] algorithm based on the response variable(s). Some samples (sample number 568–630) had values close to 0 for certain reference properties, indicating errors in the analytical experiment. Therefore, these samples were removed prior to any modelling. To include the maturity information in the model, it was dummy-coded as a vector of zeros and ones.

3.4. Three classes classification case of meat samples

The meat data set consisted of 120 ATR-MIR spectral measurements performed on three types of minced meat samples: chicken, pork, and turkey [33]. Each meat class consisted of 40 spectra. The meat data set was used to demonstrate the superiority of SO-CPLS in handling classification tasks over SO-PLS based on PLS2. All spectra were collected on a Spectra-Tech (Applied Systems Inc.), and more details about the experimental setup can be found in the earlier work [33]. Spectra were recorded from 800 to 4000 cm−1 but were truncated to 448 data points in the region of 1000–1860 cm−1 according to a previous study [33]. To create a multiblock data set, the spectra were partitioned into two ranges: 1000–1480 cm−1 and ~1480–1860 cm−1. Additionally, since each class had 40 samples, 30 out of 40 spectra were used for the calibration set, while the remaining 10 were used for the independent test set.

3.5. Data analysis

In all cases, the capability of SO-CPLS models to handle multiple response(s) was compared with the SO-PLS [34] model based on PLS2. The optimal number of latent variables (LVs) for both SO-PLS and SO-CPLS was determined using global 5-fold cross-validation. For each data block, all possible combinations of LVs were explored in the range of [0–20]. The optimal LV combination was identified as the one with the lowest root mean squared error of cross-validation (RMSECV). Note that the analysis was carried out on raw data sets without any preprocessing to ensure a fair comparison between SO-CPLS and SO-PLS. The data analysis was performed using in-house codes for SO-CPLS and SO-PLS, programmed in MATLAB (Version 2021b, MathWorks, USA).

4. Results and discussion

4.1. Pear data set

The mean spectra of pear in the NIR and the SWIR spectra range are shown in Fig. 1. In the spectral profile, different peaks can be noted in both the NIR and the SWIR spectral ranges. In the NIR range, the valley at ~960 nm and the shoulder near 750 nm can be assigned to the 3rd overtones of the OH bonds [35] present in water which is present in high abundance in fresh fruit like a pear. The peak near 820 nm can be assigned to the 3rd overtone of the NH bond [35] present in macro components such as protein in fresh fruit. The shoulder near 900 nm can be assigned to overtones of CH and CH_2 bonds [35], which are present in macro components such as sugar, fats, and protein in fresh fruit. In the SWIR region, the peak at 1450 nm can be assigned to the 2nd overtone of the OH bond [35] present in water which is present in high abundance in fresh fruit. The peak at 1200 nm and the valley near 1700 nm can be assigned to the overtones of CH and CH_2 bonds [35].

The results of the SO-PLS and SO-CPLS analysis for modeling NIR and SWIR data for pear fruit are shown in Fig. 2. The optimal models for SO-PLS were achieved by modeling 11 LVs from the NIR data block and 8 LVs from the SWIR data block, while the optimal model for the SO-CPLS was obtained using 11 LVs from the NIR data blocks and 1 LV extracted from the SWIR data block. Furthermore, the RMSEP achieved with the SO-CPLS approach was lower than the SO-PLS model for both the MC and SSC. Achieving optimal models at almost half LVs and achieving lower prediction errors already demonstrates the efficient modeling

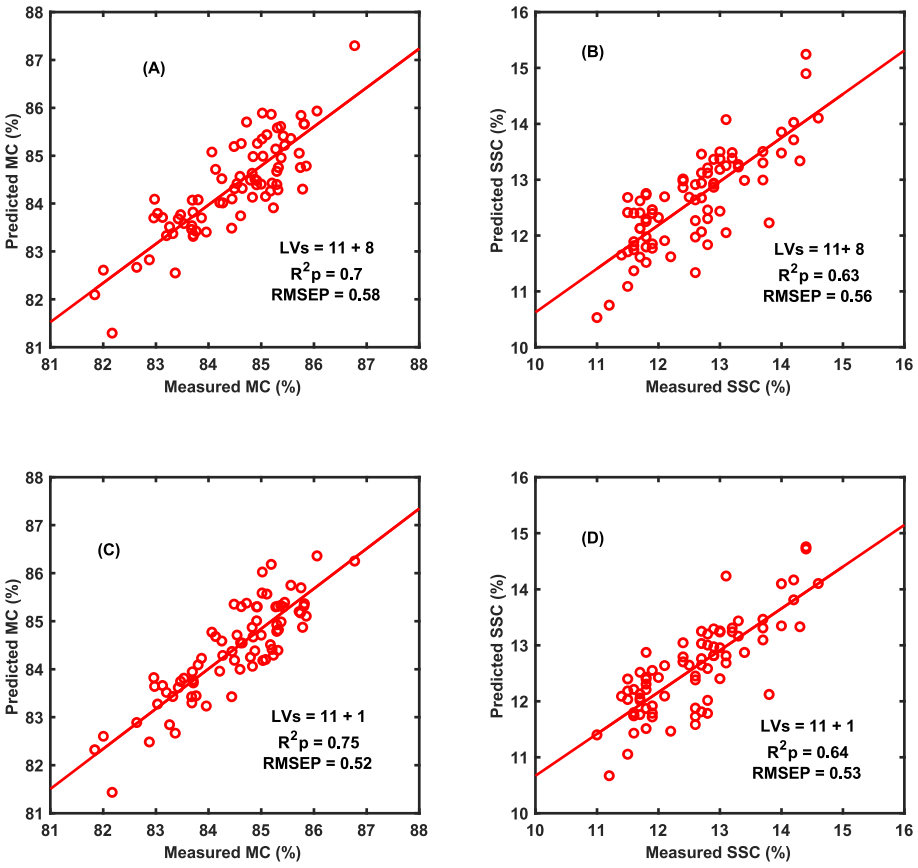


Fig. 2. SO-PLS (Top row) and SO-CPLS (bottom row) analysis for pear data set. SO-PLS based (A) moisture and (B) soluble solids prediction. SO-CPLS based (C) moisture and (D) soluble solids prediction.

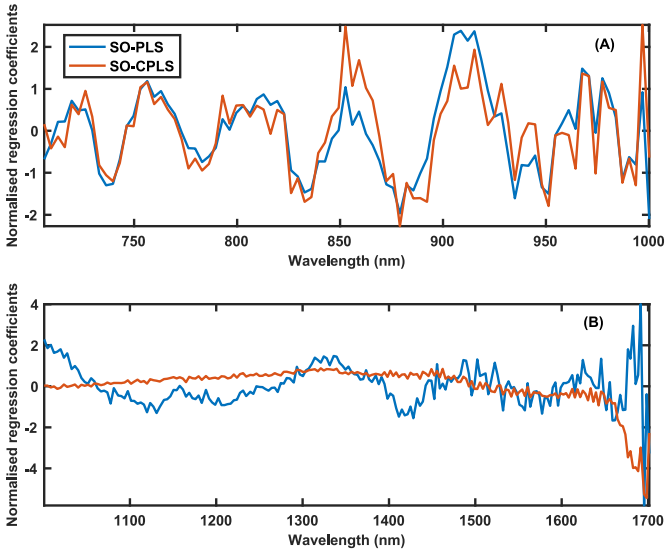


Fig. 3. Regression coefficients for the NIR (A) and the SWIR (B) data block for predicting moisture content in pear fruit.

performed by the SO-CPLS approach. To have some insights to the learning by the SO-CPLS and SO-PLS models, the regression coefficients related to MC are shown in Fig. 3. Fig. 3A are the regression coefficients for the NIR data block, while Fig. 3B are the regression coefficinets for SWIR data block. Although for NIR data block both the SO-CPLS and SO-PLS model extracted 11 LVs, the regression coefficients were not exactly the same for the NIR block. This difference already indicates that SO-CPLS and SO-PLS learned different information from the NIR data block. The regression coefficient for SO-CPLS and SO-PLS had overall similar shape however, the key difference were limited to local regions, for example, ~850 nm, 940 nm etc., which were more refined in the regression coefficients achieved with the SO-CPLS modeling (Fig. 3A). For the regression coefficients of the SWIR data block (Fig. 3B), the SO-CPLS requiring only 1 LV modelled only the information from spectral bands ~1690–1700 nm as all other bands had close to zeros regression coefficients. The SO-PLS model learned more information from the SWIR data block as many peaks can be identified in the regression coefficients (Fig. 3B). Unlike SO-CPLS, SO-PLS required more LVs to learn extra information from the SWIR data block. Even after learning a total of 8 LVs from the SWIR data block, the SO-PLS perfomed poorer than the SO-CPLS in terms of prediction error.

4.2. Milk data set

The second example of milk data set included modeling three response variables i.e., protein, fats and total solids using data from three complementary miniature spectral sensors. The results for SO-CPLS and SO-PLS models using a 5-fold cross-validation are shown in Fig. 4. The results from the analysis suggested that although the RMSEP's were similar for both the SO-CPLS and SO-PLS analysis, however, the total number of LVs and total number of blocks used in the final model were lower for the SO-CPLS than the SO-PLS. For example, in the case of the SO-CPLS only 14 LVs were required, while in case of SO-PLS 17 LVs were used in the optimal model. Apart from fewer LVs than the SO-PLS model, the SO-CPLS model optimization led to extraction of LVs from only one block of data out of three data blocks. The SO-PLS used information from two data blocks out of three. In practical

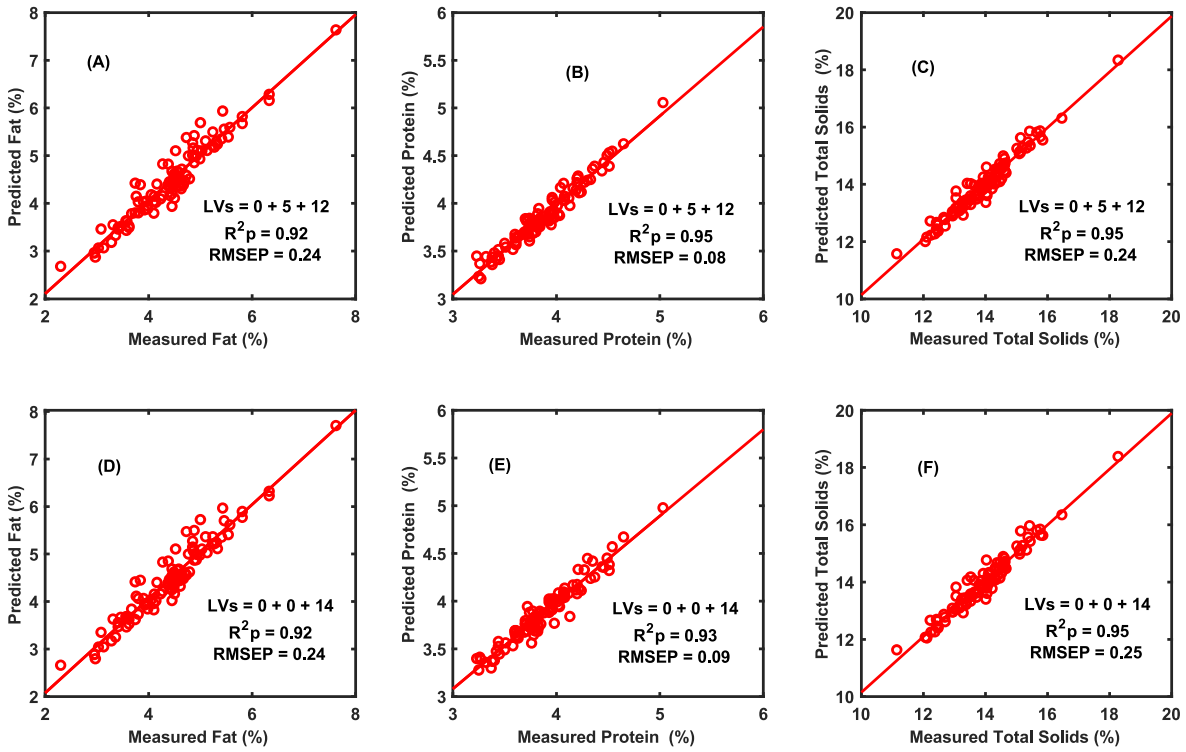


Fig. 4. A summary of SO-PLS (Top row) and SO-CPLS (Bottom row) analysis to predict fat and protein content in milk samples using multiple NIR spectral sensors. SO-PLS predictions for fat (A), protein (B), and total solids (C). SO-CPLS predictions for fat (D), protein (E), and total solids (F).

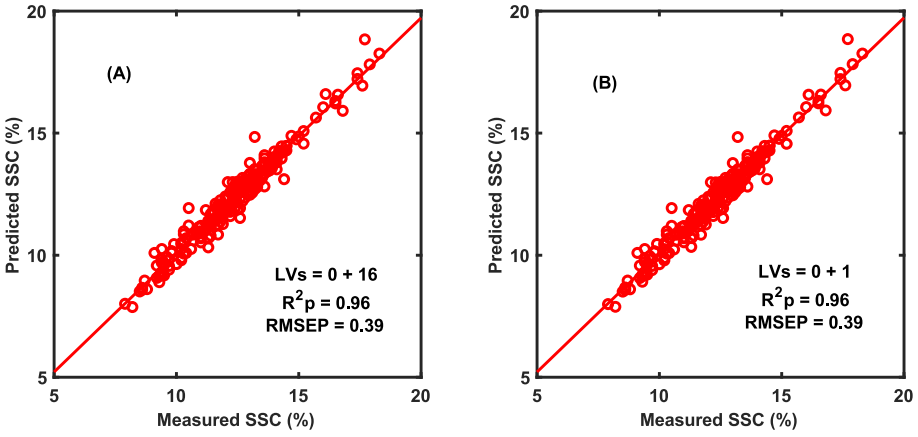


Fig. 5. A summary of SO-CPLS analysis performed on apricot dataset. (A) SO-CPLS without meta-information, and (B) SO-CPLS with meta-information.

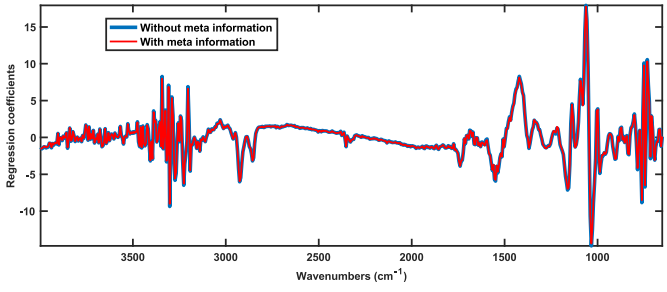


Fig. 6. Regression coefficients attained with SO-CPLS analysis with and without meta-information.

terms, the user only needs one sensor instead of multiple sensors to predict both the protein and fat contents in milk.

4.3. Apricot data set

One of the advantages of having CPLS as the backbone of the SO-CPLS is that it allows including meta-information for improving subspace extraction during the modeling process. Such meta-information is included as extra columns next to the response variable(s). An example of using the meta-information during the SO-CPLS modeling is presented using the apricot data set. In the apricot data set, the aim was to predict the soluble solids content (SSC) in fruit using the NIR and MIR data blocks. Furthermore, the meta-information was the maturity level of the fruit. Note that maturity of fruit is directly related to the SSC content of the fruit as usually more mature fruit have higher SSC content, hence, it is expected that usage of such meta-information will allow for improved

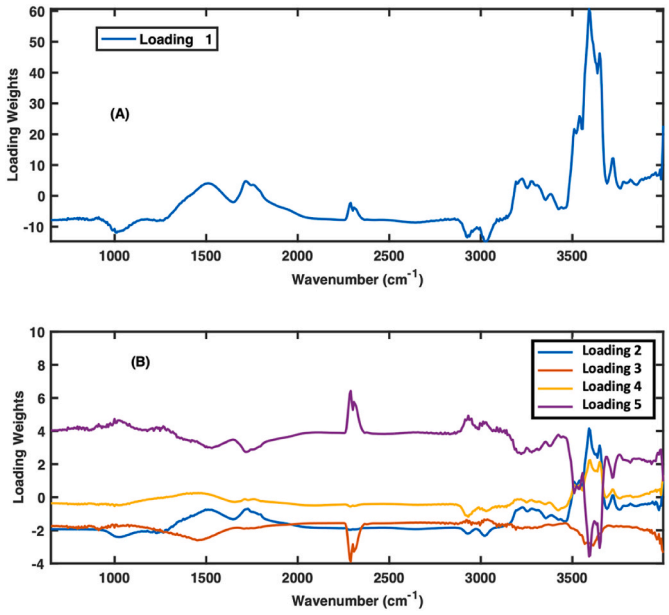


Fig. 7. The candidate loading weights for SSC (A) and the meta-information (B).

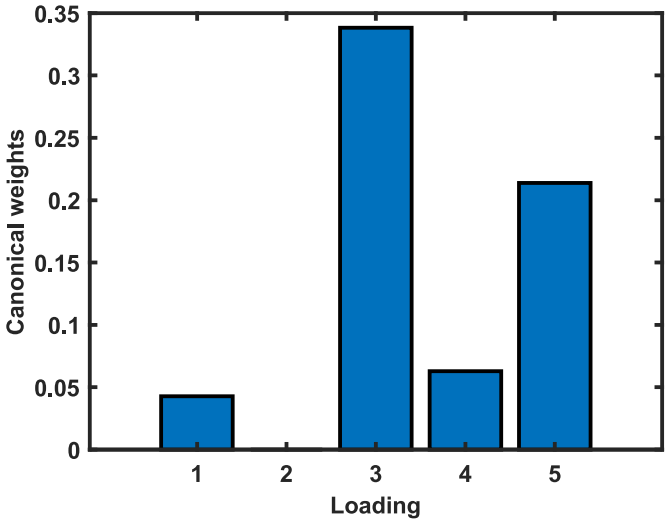


Fig. 8. The canonical weights estimated by CPLS for weighted sum of the candidate loadings to estimate the first latent variable for the case of SO-CPLS analysis with meta-information.

subspace extraction. Improved subspace extraction means more information captured in fewer LVs. The SO-CPLS without the use of meta-information achieved lower RMSEP by only using the information of the MIR data block (Fig. 5A). The SO-CPLS analysis using the meta-information about the fruit maturity achieved a similar performance as the SO-CPLS analysis without the meta-information, however, while requiring only one LV (Fig. 5B). This is a drastic decrease in the number of LVs required from the MIR data block to explain the SSC in fruit. Note that the regression vector (Fig. 6) with a 16 LVs model but without using the meta-information was almost identical to the regression vector obtained with 1 LV model using the meta-information.

To understand how the 1 LV captured such broad information which was otherwise modelled by 16 LVs, the different steps of the CPLS approach to extracting loading weights were examined. The first step of loading weight estimation, i.e., the estimation of covariance $X'Y$, resulted in five loading weights, where the first loading weight corresponded to the SSC (Fig. 7A) and the other loading weights (Fig. 7B) were for the meta-information. As a second step, the canonical weights for the loading weights were estimated which are shown in Fig. 8. Finally, the loading weights are summed up using the canonical weights to achieve the single load weight vector. Since most of the canonical weights were non-zero, this indicated that the 1 LV extracted using the meta-information was a combination of information about the SSC as well as the information present in MIR data related to fruit maturity but also related to the SSC.

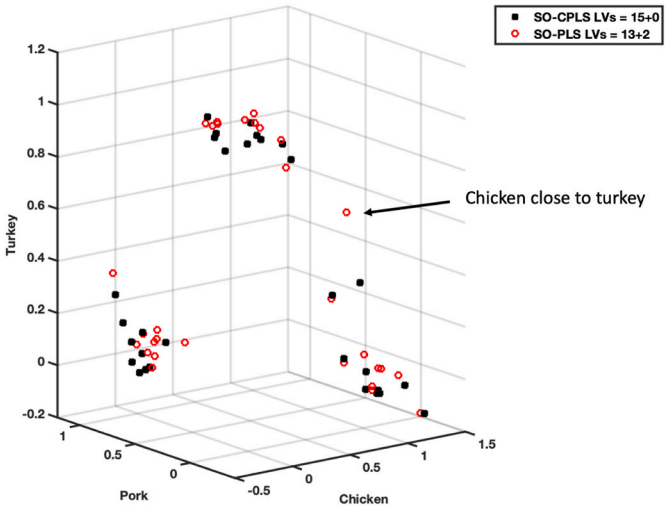


Fig. 10. Prediction from SO-PLS and SO-CPLS for different meat classes. For SO-PLS, a chicken sample was misclassified as turkey.

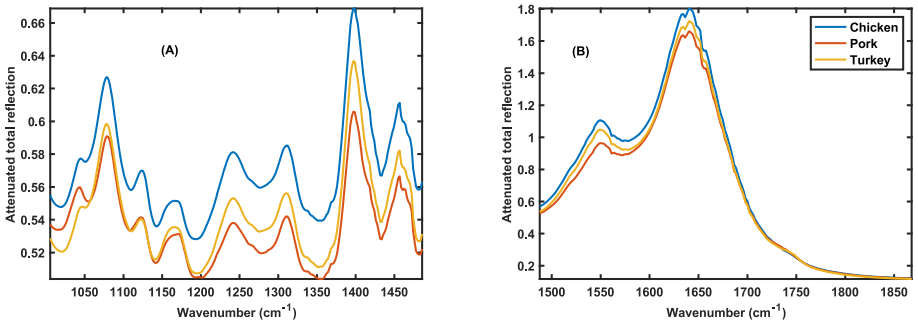


Fig. 9. The multiblock data for meat data sets. Mean spectra for each class are highlighted in different colors. (A) 1000-1480 cm^{-1}, and (B) ~1480-1860 cm^{-1}. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4.4. Meat classification

Just like the SO-CPLS can be used for multiple responses modeling, it can be used for classification modeling as well as the response vector can be supplied as the dummy class vector with the number of columns equal to the number of classes in the data. For the meat data set, there were three classes, hence, the response was a three-column matrix. The mean spectra of fresh minced chicken, pork and turkey meat are shown in Fig. 9. The spectra appear to have similar peaks but differences in intensities. According to an earlier study [33], the peak at 1650 cm^{-1} arises from water with a significant underlying contribution from protein present in high abundance in meats. According to an earlier study [33], the peak at 1550 cm^{-1} can be related to amide II absorption of protein and peak at ~1740 cm^{-1} due to fat (C=O ester).

The SO-CPLS and SO-PLS model developed on the calibration data set were tested on the interdependent test set and the prediction results are shown in Fig. 10. Both the SO-CPLS and SO-PLS model achieved optimal models with a total of 15 LVs, however, the SO-CPLS only required a single data block. The prediction showed that with the SO-PLS, a chicken sample was misclassified as the turkey samples, while with the SO-CPLS, that sample was correctly classified as the chicken sample. Like the multiple responses regression modeling, the results of the classification modeling demonstrate the efficient modeling achievable with the SO-CPLS.

5. Conclusions

This study presents a novel multiblock extension of canonical partial least squares (CPLS) for efficient multiple response modeling and improved subspace extraction via the use of meta-information. The approach replaces the PLS2 step in the traditional sequential orthogonalised partial least squares (SO-PLS) modeling with the CPLS step to develop sequential orthogonalised canonical partial least squares (SO-CPLS) models. Results on various multiple response datasets showed that the SO-CPLS approach achieved better or similar performance in terms of root mean squared error of prediction (RMSEP) compared to SO-PLS, while using fewer latent variables (LVs). This demonstrates the efficiency of the SO-CPLS approach for subspace extraction. Additionally, the SO-CPLS approach allows the incorporation of meta-information to further enhance subspace extraction, similar to the CPLS approach. Tests on the Apricot dataset showed that the use of meta-information in the SO-CPLS resulted in lower RMSEP models using fewer LVs. Notably, the SO-CPLS approach converges to SO-PLS when used for single-response datasets. Therefore, this study concludes that SO-CPLS is a more general tool for sequential multiblock modeling than the traditional SO-PLS approach, and its unique feature of using meta-information can improve subspace extraction. Finally, it is expected that SO-CPLS will be beneficial for multiple response multiblock modeling, thanks to the developer of CPLS [28].

CRediT authorship contribution statement

Puneet Mishra: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft.

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.

Data availability

Data will be made available on request.

References

[1] A.K. Smilde, et al., Common and distinct components in data fusion, J. Chemometr. 31 (7) (2017) e2900.
[2] P. Mishra, et al., Recent trends in multi-block data analysis in chemometrics for multi-source data integration, TrAC, Trends Anal. Chem. (2021), 116206.
[3] P. Mishra, et al., Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit, Talanta 223 (2021), 121733.
[4] A. Biancolillo, et al., Extension of SO-PLS to multi-way arrays: SO-N-PLS, Chemometr. Intell. Lab. Syst. 164 (2017) 113–126.
[5] 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.
[6] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, Food Qual. Prefer. 24 (1) (2012) 8–16.
[7] Z. Xu, et al., A calibration transfer optimized single kernel near-infrared spectroscopic method, Spectrochim. Acta Mol. Biomol. Spectrosc. 220 (2019), 117098.
[8] 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.
[9] P. Mishra, et al., Parallel Pre-processing through Orthogonalization (PORTO) and its Application to Near-Infrared Spectroscopy, Chemometrics and Intelligent Laboratory Systems, 2020, 104190.
[10] 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, 104139.
[11] E. Acar, et al., Structure-revealing data fusion, BMC Bioinf. 15 (1) (2014) 239.
[12] C.A. Andersson, R. Bro, The N-way toolbox for MATLAB, Chemometr. Intell. Lab. Syst. 52 (1) (2000) 1–4.
[13] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (9) (2014) 2812–2831.
[14] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[15] S. Wold, PLS Modeling with Latent Variables in Two or More Dimensions, 1987.
[16] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemometr. 12 (5) (1998) 301–321.
[17] T. Næs, et al., Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, Chemometr. Intell. Lab. Syst. 124 (2013) 32–42.
[18] K.H. Liland, T. Næs, U.G. Indahl, ROSA—a fast extension of partial least squares regression for multiblock data analysis, J. Chemometr. 30 (11) (2016) 651–662.
[19] R. Bro, et al., Data fusion in metabolomic cancer diagnostics, Metabolomics : Official journal of the Metabolomic Society 9 (1) (2013) 3–8.
[20] M.P. Campos, M.S. Reis, Data preprocessing for multiblock modelling – a systematization with new methods, Chemometr. Intell. Lab. Syst. 199 (2020), 103959.
[21] P. Mishra, et al., Pre-processing Ensembles with Response Oriented Sequential Alternation Calibration (PROSAC): A Step towards Ending the Pre-processing Search and Optimization Quest for Near-Infrared Spectral Modelling, Chemometrics and Intelligent Laboratory Systems, 2022, 104497.
[22] A. Biancolillo, F. Marini, J.-M. Roger, So-CovSel, A novel method for variable selection in a multiblock framework, J. Chemometr. 34 (2) (2020) e3120.
[23] P. Firmani, et al., Multi-block classification of Italian semolina based on Near Infrared Spectroscopy (NIR) analysis and alveographic indices, Food Chem. 309 (2020), 125677.
[24] P. Mishra, et al., Improved Prediction of Fuel Properties with Near-Infrared Spectroscopy Using a Complementary Sequential Fusion of Scatter Correction Techniques, Talanta, 2020, 121693.
[25] A. Biancolillo, et al., Variable selection in multi-block regression, Chemometr. Intell. Lab. Syst. 156 (2016) 89–101.
[26] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, Chemometr. Intell. Lab. Syst. 141 (2015) 58–67.
[27] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (3) (1993) 251–263.
[28] U.G. Indahl, K.H. Liland, T. Næs, Canonical partial least squares—a unified PLS approach to classification and regression problems, J. Chemometr. 23 (9) (2009) 495–504.
[29] P. Mishra, et al., Chemometric Pre-processing Can Negatively Affect the Performance of Near-Infrared Spectroscopy Models for Fruit Quality Prediction, Talanta, 2021, 122303.

[30] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1) (1969) 137–148.
[31] S. Uusitalo, et al., Evaluation of MEMS NIR spectrometers for on-farm analysis of raw milk composition, Foods 10 (11) (2021).
[32] S. Bureau, et al., Application of ATR-FTIR for a rapid and simultaneous determination of sugars and organic acids in apricot fruit, Food Chem. 115 (3) (2009) 1133–1140.
[33] O. Al-Jowder, E.K. Kemsley, R.H. Wilson, Mid-infrared spectroscopy and authenticity problems in selected meats: a feasibility study, Food Chem. 59 (2) (1997) 195–201.
[34] A. Biancolillo, T. Næs, M. Cocchi, Chapter 6 - the sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions, in: Data Handling in Science and Technology, Elsevier, 2019, pp. 157–177.
[35] B.G. Osborne, Near-Infrared spectroscopy in food analysis, in: Encyclopedia of Analytical Chemistry, 2006.