Analytica Chimica Acta 1206 (2022) 339786

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Swiss knife partial least squares (SKPLS): One tool for modelling single block, multiblock, multiway, multiway multiblock including multi-responses and meta information under the ROSA framework

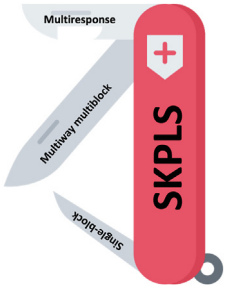
Puneet Mishra [a, *], Kristian Hovde Liland [b]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] Faculty of Science and Technology, Norwegian University of Life Sciences, 1430, Ås, Norway

HIGHLIGHTS

- A unified version of PLS is proposed.
- Method can handle wide type of data structure such as single block, multiblock, multiway.
- Method can also deal with multiple responses and include meta information in models.
- Method was demonstrated for modelling wide type of data structures.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 12 February 2022
Received in revised form
25 March 2022
Accepted 28 March 2022
Available online 30 March 2022

Keywords:
Multiway modelling
Data fusion
Multi-modal
Spectroscopy

ABSTRACT

In the domain of chemometrics and multivariate data analysis, partial least squares (PLS) modelling is a widely used technique. PLS gains its beauty by handling the high collinearity found in multivariate data by replacing highly covarying variables with common subspaces spanned by orthogonal latent variables. Furthermore, all can be achieved with simple steps of linear algebra requiring minimal computation power and time usage compared to current high-end computing and substantial hyperparameter tuning required by methods such as deep learning. PLS can be used for a wide variety of tasks, for example, single block modelling, multiblock modelling, multiway data modelling and for task such as regression and classification. Furthermore, new PLS based approaches can also incorporate meta information to improve the PLS subspace extraction. However, in the current scenario, there is a wide range of separate tools and codes available to perform different PLS tasks. Often when the user needs to perform a new PLS task, they need to start with a separate mathematical implementation of the PLS techniques. This study aims to provide a single solution, i.e., the Swiss knife PLS (SKPLS) modelling approach to enable a single mathematical implementation to perform analyses of single block, multiblock, multiway, multiblock multiway, multi-response, and incorporation of meta information in PLS modelling. It contains all that is needed for any PLS practitioner to perform both classification and regression tasks. The SKPLS backbone is the stepwise PLS strategy called response oriented sequential alternation (ROSA) which we generalize to enable all the mentioned analysis possibilities. The basic structure of the algorithm is highlighted, and

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

https://doi.org/10.1016/j.aca.2022.339786

some example cases of performing single block, multiblock, multiway, multiblock multiway, multi-response PLS modelling and the incorporation of meta information in PLS modelling are included.

1. Introduction

In the domain of analytical chemistry and chemometrics, multivariate multi-collinear data is widely encountered [1]. Such data is usually generated with a wide range of analytical instruments such as spectrometers [2]. Often, one of the main aims in analytical chemistry is to develop calibration models for analytical instruments using a few reference samples measured with wet chemistry approaches to use the data of the instrument and the calibration to predict future key properties of interest such that wet chemistry can be avoided [1,3,4]. This reduces the need for costly and labour-intensive manual analyses. There are also other benefits of using analytical instruments such as spectrometers compared to wet chemistry as the property predictions can be made in real-time and in a minimally invasive way. Since, the analytical instrument data such as from spectrometers are highly collinear, one of the most important chemometric modelling technique is the partial least squares (PLS) based modelling [3,4]. PLS is a type of bilinear latent space modelling technique with repeated steps of covariance maximisation. The latent spaces are extracted and the multi-collinearity in data is leveraged to build low rank generalizable models [5,6]. Another key benefit of PLS modelling is that being a bilinear modelling approach it provides score and loading vectors, which allow interpreting any pattern or chemical relevant background information present in the samples [6]. Furthermore, PLS can be used for both regression and classification tasks. It is no wonder that due to such many advantages, PLS is a gold standard technique in the chemometric domain [7].

Apart from using PLS modelling for single block scenarios, extensions of PLS modelling can be found for a wide range of tasks such as data fusion [8–14] as well as handling higher order arrays [15–17]. Furthermore, approaches for using meta information for improving latent space extraction as well as multi-response modelling also exist [18]. The task of data fusion can be termed as multiblock PLS modelling [12], and the handling of higher order arrays is known as multiway PLS [17]. The method allowing to incorporate meta information is the canonical PLS modelling [18]. Meta information is one or more additional variables describing the training data. This can be experimental conditions, mixture proportions or other data collected when performing the spectral measurements. The same meta information is often not available for test data or when the model is applied in a production environment because of practical or economic limitations. However, since the meta information is only used in the modelling phase for increasing the adaptability of the latent variables and making a more informed choice of subspace, this information is not needed at prediction time. These extensions were developed for handing the varying nature of data structures as presented in Fig. 1. The single block PLS case can be understood as where multivariate data measured on a set of samples is used to predict responses. The second case of PLS comes when there are multiple blocks of predictors, predicting responses using multiple multivariate data measured on the same samples. The third case of PLS arises when the data, instead of being a 2D matrix, is an n-dimensional array, i.e., a multiway data block. The fourth case is multiblock data with one or more multiway blocks where multiway data is jointly analysed with a 2D matrix or multiway block. Furthermore, the responses can be either single response or multi-response, and with or without extra meta information.

In the domain of chemometrics, there are many approaches available to perform PLS analysis, a summary of several PLS decomposition approaches can be found elsewhere [19,20]. For multiblock data modelling there are also several approaches [12,13] such as multiblock PLS based on hierarchal [11,21], sequential [15,22,23], parallel PLS [23,24] and response-oriented sequential alternation (ROSA) [25] frameworks. For multiway data analysis approaches can be found such as the multiway PLS framework [17]. Also, for including the meta information, an approach called CPLS is available [18]. However, currently what is lacking is a unified PLS version which can combine all the approaches under a single umbrella and have a flexible single-interface model suitable for handling any type of data structures mentioned in Fig. 1. Such a unified version of PLS will be of wide user attention and can increase the applications of widely scattered PLS approaches in the wide domain of sciences.

Of particular interest is that all the PLS approaches are stepwise approaches [25] where at each step the model components are extracted by maximizing covariance or canonical correlation. Furthermore, to achieve stable and generalizable models', a suitable number of components are extracted with approaches such as cross-validation [26]. Since the components are modelled in a stepwise approach it gives PLS the flexibility to model components from different data blocks in each step and finally to select the optimal data block. Winning strategies can be leverage metrics such as maximum correlation, minimum residual [25], or maximum canonical correlation. This is also the motivation behind the ROSA modelling in which several data blocks can be modelled in a stepwise strategy but at each step the winner is decided as the one resulting in minimum residual [25]. In such a way, the PLS model can be extended to work on multiple data blocks learning the complementary information available in all possible blocks. Also note that when only a single block of data is available then the ROSA modelling approach will become the standard single block PLS where in each step the model components are extracted by covariance maximisation. The presented SKPLS approach is also based on a ROSA [25] backbone and reaps the benefit of the stepwise modelling nature of the PLS approach. Publications with ROSA [25] currently only cover the case of two-way predictors and a single response. However, the R package *multiblock* [27] has been published, which extends ROSA with CPLS, enabling multiple responses and meta information. The SKPLS extends this further by enabling multiway blocks, thus providing a single tool for single- and multiblock with two- and multiway data, using single and multiple responses with meta information.

This study aims to provide a single solution, i.e., the Swiss knife PLS (SKPLS) modelling approach for all the mentioned modelling tasks (Fig. 1). The basic structure of the algorithm is highlighted, and some example cases are presented for performing single block, multiblock, multiway, multiblock multiway, multi-response PLS modelling, as well the incorporation of meta information in PLS modelling. The codes of the technique will be made available at: https://github.com/puneetmishra2.

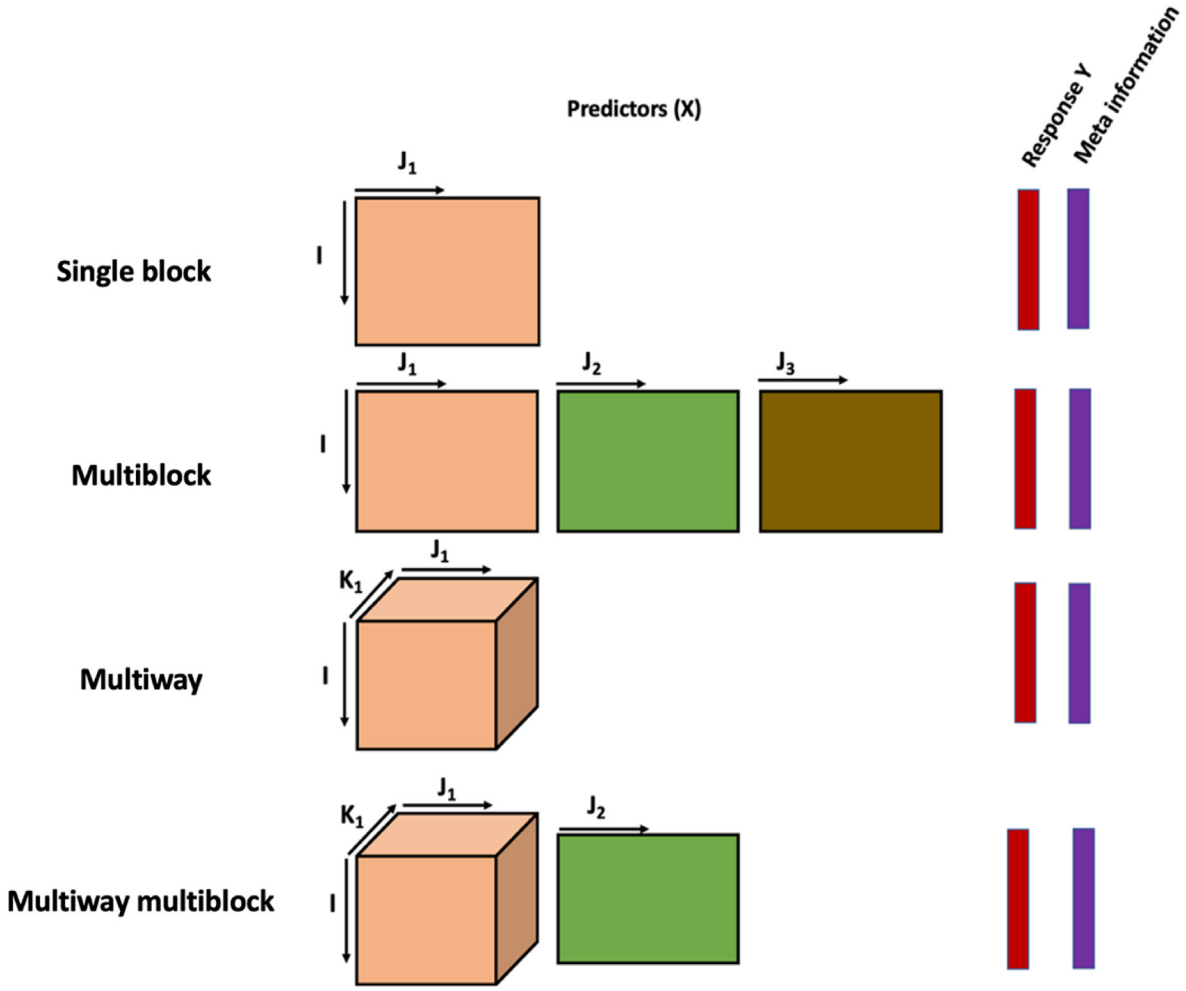


Fig. 1. Examples of data structures that can be analysed by Swiss knife PLS.

2. Method

The SKPLS is an extension of the PLS based ROSA [25] modelling strategy where each PLS model component extraction step is given the possibility to explore *B* number of data blocks to learn complementary latent variables. In each step of ROSA [25], at first, the latent spaces are extracted by maximizing the covariance with the response variable at the block level and later the scores based on the latent spaces compete to minimise the residual for the response variable *y*. The block score leading to the lowest sum of squared residuals is selected as the winner for that step, and later the same step is repeated but constrained to be orthogonal to the subspace already spanned by the earlier selected components. Although the ROSA methodology has been extended further in software [27], it still lacks the ability to handle multiway blocks and a unified description covering multiple responses and meta information. On the other hand, the SKPLS, covers all the necessary tools to perform all major tasks required for PLS modelling of single block, multiblock, and multiway data for single and multi-responses. The SKPLS approach is presented as follow.

Algorithm steps for SKPLS based on the ROSA backbone

Define $\boldsymbol{Y}$ as the response matrix, $\boldsymbol{Y_{meta}}$ as the meta information, B as be the number of (centred) data blocks $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_B$ and A as the desired number of components to be extracted. Note that data blocks can be of any dimensionality: one-way (column vector), two-way (matrix), or multiway, and they will always be mean-centred in the sample mode. $\boldsymbol{Y_{concat}} = [\boldsymbol{Y}\ \boldsymbol{Y_{meta}}]$ is the column wise concatenation of $\boldsymbol{Y}$ and $\boldsymbol{Y_{meta}}$.

for $a = 1{:}A$

Step 1: If the data is two-way data, like a matrix of size (I, J), compute the B loading weights $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_B)$ as $\boldsymbol{v}_i = \boldsymbol{X}_i{}'\boldsymbol{Y_{concat}}\boldsymbol{c}_i$, where $\boldsymbol{c}_i$ are the canonical weights obtain by the canonical correlation between $\boldsymbol{Z}_i = \boldsymbol{X}_i\boldsymbol{X}_i{}'\boldsymbol{Y_{concat}}$ and $\boldsymbol{Y}$, as explained in the CPLS algorithm [18]. The $\boldsymbol{c}_i$ vector obtained through canonical correlation maximizes the correlation between $\boldsymbol{Z}$ and the response(s). Thus, the corresponding loading weights $\boldsymbol{v}_i$ will also result in an optimal score vector $(\boldsymbol{Z}\boldsymbol{c}_i)$ for explaining the response(s) with respect to the expanded space including $\boldsymbol{Y}_{meta}$. If the data block is a multiway array, for example, a 3-way array of dimension (I, J, K) the data is first unfolded for the first mode $(I, (J \times K))$, then a singular value decomposition (SVD) is performed on reshaped $\boldsymbol{X}'\boldsymbol{y}$, a $(J \times K)$ matrix, leading to 2-way weights for which the corresponding loading weight vector can be estimated by the Kronecker product [17]. For higher mode data blocks (more than 3 modes), the SVD step can be replaced with parallel factor analysis (PARAFAC) [17] type decomposition leading to multiway weights, which are converted to the loading weight vector by estimating the Kronecker product. Note that the meta information cannot be used currently for the multiway data blocks, and the topic of how to integrate meta information in multiway modelling is currently under investigation.

Step 2: Based on weights, estimate the competing scores $\boldsymbol{\tau}_1 = \boldsymbol{X}_1\boldsymbol{v}_1, \boldsymbol{\tau}_2 = \boldsymbol{X}_2\boldsymbol{v}_2 \ldots\ldots \boldsymbol{\tau}_B = \boldsymbol{X}_B\boldsymbol{v}_B$ and if $a > 1$, modify them into their normalised projection onto the orthogonal complement of the subspace spanned by previously selected scores $\boldsymbol{T} = [\boldsymbol{t}_1 \ldots \boldsymbol{t}_{a-1}]$, i.e., $\boldsymbol{\tau}_j \Leftarrow \boldsymbol{\tau}_j - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{\tau}_j)$ and $\boldsymbol{t}_j = \frac{\boldsymbol{t}_j}{\|\boldsymbol{t}_j\|}$, where $j = 1, 2, \ldots\ldots B$. The numerical stability of the orthogonalization approach has already been confirmed, e.g., in [28].

Step 3: The score carrying the maximum canonical correlation with the response variable(s) $\boldsymbol{Y}$ is declared as the winner, $\boldsymbol{t}_a$, for that step and the block corresponding to that score will be declared as the winner for that step. Please note that unlike the single response ROSA that uses minimum residual as the winning criterion, this study uses the canonical correlation to handle the multi-response scenario as canonical correlation is unaffected by the scales of variables. Also note that, like for most multiblock methods, the scores will lie in the space spanned by all input blocks and not necessarily in the space of any single block.

Step 4: The $\boldsymbol{t}_a$ is used to update the response by estimation of residuals as $\boldsymbol{Y}_{a+1} = \boldsymbol{Y} - \boldsymbol{t}_a\boldsymbol{t}_a{}'\boldsymbol{Y}$

Step 5: Normalise the winning loading weight $\boldsymbol{\omega}_a = (0^t\boldsymbol{v}_a^t0^t)^t \in \boldsymbol{R}^P$ $(p = \sum_{k=1}^{B} p_k)$ and estimate the orthogonal complement with respect to the previous winning loadings $\boldsymbol{W} = [\boldsymbol{w}_1 \ldots \boldsymbol{w}_{a-1}]$ as $\boldsymbol{w}_a = \boldsymbol{\omega}_a - \boldsymbol{W}(\boldsymbol{W}^t\boldsymbol{\omega}_a)$ and $\boldsymbol{w}_a = \frac{\boldsymbol{w}_a}{\|\boldsymbol{w}_a\|}$

Step 7: Estimate regression coefficients with respect to the $\boldsymbol{t}_a$ as $\boldsymbol{q}_a = \boldsymbol{Y}_a^t\boldsymbol{t}_a$.

Step 8: Accumulate vectors into matrices $\boldsymbol{T} = [\boldsymbol{t}_1\boldsymbol{t}_2 \ldots.\boldsymbol{t}_a]$ *and* $W = [\boldsymbol{w}_1\boldsymbol{w}_2 \ldots.\boldsymbol{w}_a]$.

end

for $a = 1{:}A$

(***continued***).

Step 9: Concatenate all data blocks as $\boldsymbol{X}_c = [\boldsymbol{X}_1\boldsymbol{X}_2 \ldots . \boldsymbol{X}_B]$ and estimate the loadings as $\boldsymbol{P} = \boldsymbol{X}_c^t\boldsymbol{T}$

for i = 1: *number of responses*

Step 10: Compute the associated regression coefficients for concatenated data block as $\boldsymbol{b}_i = \boldsymbol{W}(\boldsymbol{P}^t\boldsymbol{W})^{-1}\boldsymbol{q}_{ai}$ and associated offset terms $\boldsymbol{b}_{oi} = \overline{\boldsymbol{Y}} - \overline{\boldsymbol{X}}_c\boldsymbol{b}_i$ for the prediction of future uncentered spectral data. Here, $\overline{\boldsymbol{Y}}$ is the mean of the responses and $\overline{\boldsymbol{X}}_c$ is the mean of the concatenated data blocks $\boldsymbol{X}_c$

end

end

(***continued***).

Please note that a key feature of the above algorithmic steps is that in the case of single matrix type data or a multiway type data block, the competition step will not be present and the method will converge to a single-block PLS or multiway PLS analysis. In the presence of multiple blocks of either matrix type or multiway array, the method will become a multiblock approach where the model components will be selected as the one maximizing the correlation with the response in the case of single response, while maximizing the canonical correlation in the case of multi-response. Furthermore, due to the use of the CPLS [18], the method can use the meta information about samples, in the case of matrix type data blocks, to achieve efficient subspace extraction. Please also note that just like ROSA [25], the method does not have any predictor matrix deflation step as commonly present in other PLS algorithms [20], thus giving the SKPLS the advantage of speed of execution compared to other PLS algorithms involving predictor matrix deflation. It should be noted that, like the original ROSA, SKPLS also generalizes sequential and orthogonalized PLS (SO-PLS) [28] since the SO-PLS solution can be obtained by constraining the components to be selected from one block at the time, e.g., 8 components from three blocks: 1,1,1, 2,2, 3,3,3. This makes separate theory for SO-PLS redundant since it is covered by SKPLS. Just like the traditional PLS approaches, the optimal number of latent variables to extract for generalised modelling can be optimised using cross-validation approaches or using validation sets in the presence of larger sample sets. In this study, a 5-fold cross-validation was implemented to optimise the total number of latent variables for all the demonstrated cases.

3. Data sets for demonstration

3.1. Protein and fat prediction in milk with three spectral sensors and two measurement modes

The milk data set was used for demonstrating the single-block and multiway PLS modelling, multiblock data fusion of matrix type data blocks and multiblock data fusion of multiway type data blocks. The milk data set consisted of spectral data and reference protein and fat measurements performed on 296 milk samples [29]. The spectral measurements were performed with three different portable spectral sensors working in complementary NIR spectral ranges: NIRONE 1.4, NIRONE 2.0 and NIRONE 2.5 from Spectral Engines (Helsinki, Finland). The NIRONE 1.4 was in 1100–1400 nm, NIRONE 2.0 was in 1550–1950 nm and NIRONE 2.5 was in 2000–2450 nm spectral ranges. For all the three spectrometers, all measurements were performed in transmission mode except for the NIRONE 2.0, for which additional measurement of the same samples were performed in reflectance mode. More information on the data set and reference protein and fat analysis protocols can be obtained in the earlier study. Data are summarized in Table 1.

3.2. Soluble solids prediction in apricot puree using near-infrared and mid-infrared sensing and using maturity level as the meta information

The apricot data set is used to demonstrate the capability of the SKPLS to use meta information for efficient subspace extraction.

Table 1
A summary of the milk data set.

	NIRONE 1.4	NIRONE 2.0	NIRONE 2.5	Protein (% w/w)	Fat (% w/w)
Spectral range (nm)	1100–1350	1550–1950	2000–2450	*	*
Data shape	296 × 126	296 × 201 × 2	296 × 226	296 × 1	296 × 1
Reference range (Average ± standard deviation)	*	*	*	3.90 ± 0.41	4.71 ± 1.10

*not relevant.

Table 2
A summary of apricot data set.

	NIR	MIR	Soluble solids content (%)
Spectral range	800–2772 nm	3996-651 cm^{-1}	*
Data shape	750 × 769	750 × 579	750 × 1
Reference range (Average ± standard deviation)	*	*	12.36 ± 2.43

*not relevant.

The apricot data set has NIR (800–2770 nm) and MIR (4000-650 cm^{-1}) spectra acquired on 750 apricots, along with the reference soluble solids content (SSC %) measurements. The raw spectral data without any pre-processing step was used for this analysis. More details on the reference analysis can be obtained in Ref. [30]. Apart from reference measurements, the data set also has information about fruit maturity from three different maturity stages: very green, ripe, over-ripe. The maturity information about the samples will be used as the meta-information to demonstrate the potential of SKPLS to improve subspace extraction for SSC prediction. A summary of the data set can be found in Table 2.

4. Results and discussion

4.1. Spectral data description

Summaries of mean spectral profiles of the milk and apricot data set are shown in Figs. 2 and 3, respectively. For the milk spectral profiles, the data were measured in three complementary NIR spectral ranges: 1100–1350 nm, 1550–1950 nm, and 2000–2450 nm. We can observe from the mean spectral profile that Block 1 (Fig. 2A) and Block 3 (Fig. 2C) of the milk data were 2-way data while Block 2 (Fig. 2B) was 3-way data where one extra mode corresponds to the optical measurement geometry of the spectrometer, i.e., transmission and reflection. In all the spectral ranges, peaks and valleys can be assigned to chemical overtones of OH, CH, and NH bonds present in abundance in milk due to macromolecules such as water, fats, and proteins [29,31,32]. For the apricot data set (Fig. 3), the peaks present in the mean spectral profiles in the NIR and MIR range can also be assigned to a wide range of chemical absorptions and overtones as identified in an earlier study [30]. What comes next are the individual demonstration analyses of SKPLS for modelling all previously mentioned combinations of block dimensionality, and incorporation of meta information for efficient subspace modelling. Please note that for both the milk and apricot data sets, the samples were partitioned into calibration (60%) and test sets (40%) using the Kennard-Stone algorithm [33], where model cross-validation and calibration was performed using the calibration set while the optimised models were tested on the test set.

4.2. Single block PLS solution from SKPLS

An example analysis of SKPLS for single block analysis was performed on one data block from the milk data set (NIRONE 1.4) to predict fat content, and the results are shown in Fig. 4. The 5-fold cross-validation allowed selecting 8 latent variables (Fig. 4A–C) and the model achieved similar calibration and prediction error.

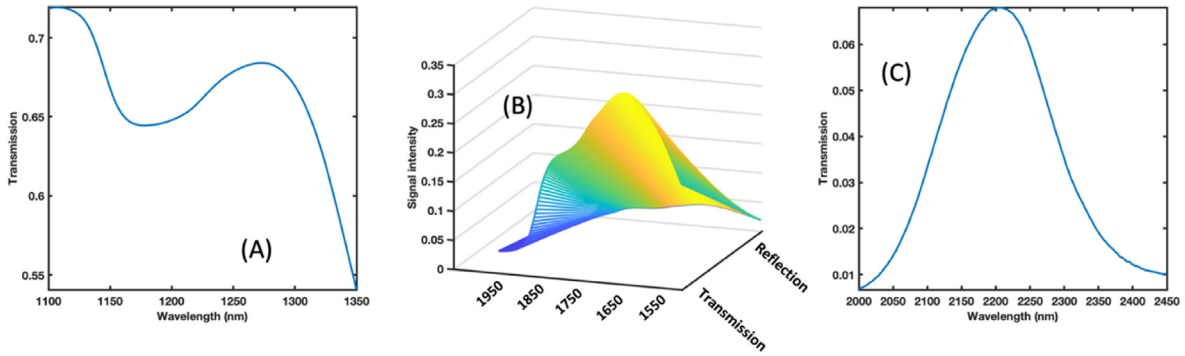


Fig. 2. Mean spectral responses for the milk data set. (A) transmission mode NIRONE 1.4, (B) transmission and reflection mode NIRONE 2.0, and (C) transmission mode NIR 2.5.

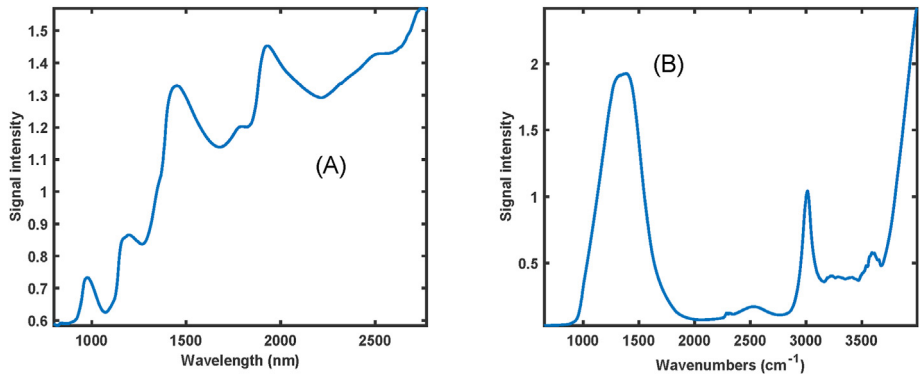


Fig. 3. Mean spectral responses for apricot data set. (A) NIR, and (B) MIR.

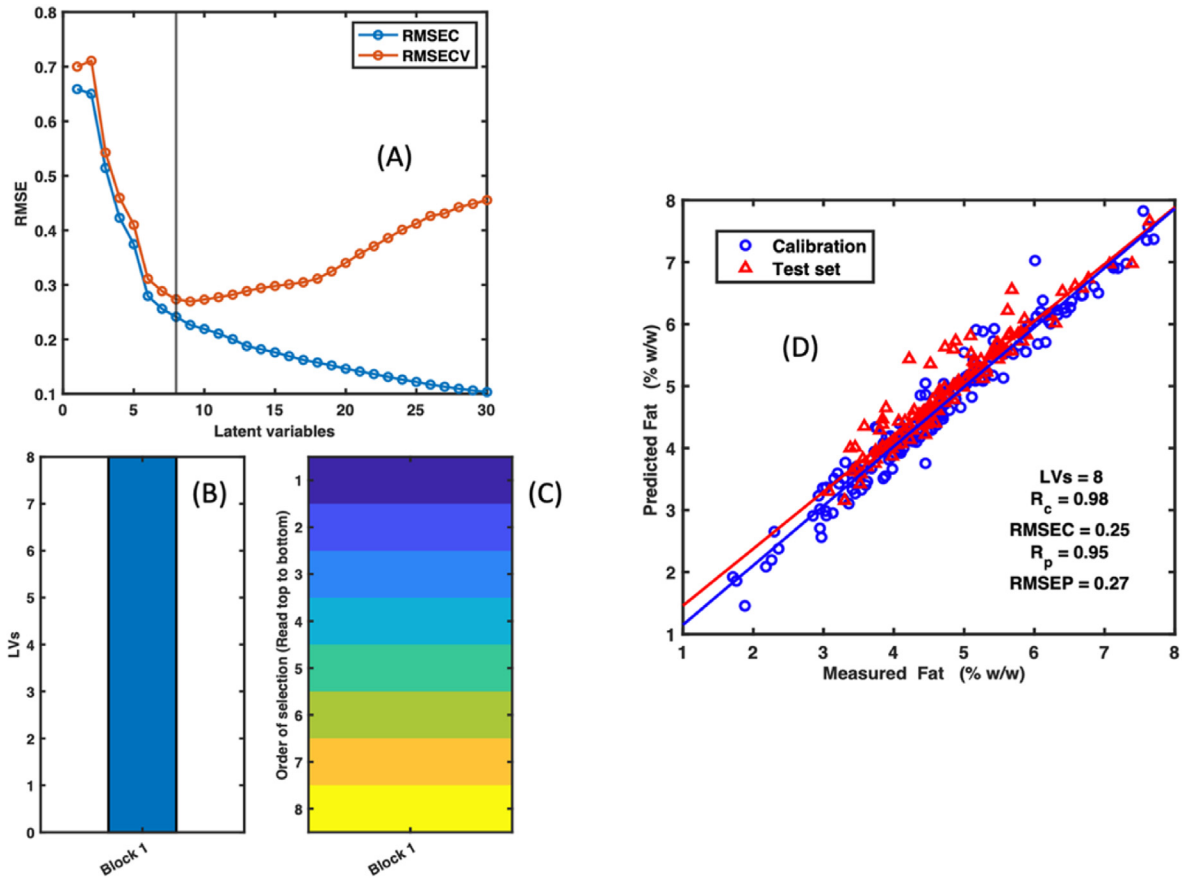


Fig. 4. Results of single block SKPLS analysis performed on NIRONE 1.4 sensor data for predicting fat content. (A) Cross-validation plot used to select latent variables, (B–C) 8 latent variables were selected, and (D) prediction plot. The changes in colour in (C) from dark blue towards yellow indicates the increasing number of latent variables. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The model errors were in similar range as what was obtained in an earlier study using traditional PLS modelling [29]. Note that plots B and C are redundant when the analysis contains only a single data block but is included for consistency.

4.3. Single-block multiway PLS solution with SKPLS

In analytical chemistry, multiway data can also be frequently encountered. For example, in this study, the spectral measurements performed on milk samples were performed in two different measurement modes (transmission and absorption) leading to a 3-way data set. The SKPLS allows modelling multiway data sets by inclusion of multiway PLS, and as an example the multiway data set was processed to predict fat content in milk samples. The results of the SKPLS for modelling 3-way data are shown in Fig. 5. It can be noted that with cross-validation 8 latent variables were selected (Fig. 5A–C) which led to a prediction error ~0.2% w/w for predicting fats content in milk.

4.4. Multiblock 2-way data analysis with SKPLS

The SKPLS, relying on the ROSA stepwise strategy, can also be used for multiblock analysis of 2-way data blocks. This is because at each step of latent variable extraction, the SKPLS gives the opportunity of latent variable extraction to all the available data blocks and selects the winning block as the one carrying maximum correlation with the response variable in case of single response and maximum canonical correlation with the response variables in case of multi-response. As an example, multiblock analysis was performed for fusing information from three data blocks corresponding to transmission spectral measurements performed using NIRONE 1.4 (Block 1), NIRONE 2.0 transmission mode (Block 2) and NIRONE 2.5 (Block 3) sensor data for predicting fat content in milk. The results are shown in Fig. 6, where the cross-validation plot (Fig. 6A) suggested extraction of 8 latent variables. The total latent variables plot (Fig. 6B) suggested extraction of 4 latent variables from Block 1, and 2 latent variables each from Block 2 and Block 3. The order of latent variables (Fig. 6C) indicates that initially the latent variables were extracted from Block 1 and later from Block 2 and Block 3. The prediction errors for calibration and test set were similar indicating optimal model fitting regarding possible overfitting.

4.5. Multiblock multiway data block analysis with SKPLS

Just like the SKPLS allows multiblock modelling of multiple 2-way data blocks, it also allows modelling when the data blocks are of multiway type. As an example, the three block milk data set, where the 2nd block was 3-way data, was modelled using the SKPLS to predict fat content in milk. The results of the multiblock multiway data modelling are shown in Fig. 7, where the cross-validation allowed selection of 8 latent variables (Fig. 7A). Out of the 8 selected latent variables, 5 belonged to the 3-way data (Block 2), while 1 latent variable belonged to Block 1 and 2 to Block 3. In terms of order (Fig. 7C), the first latent variable was selected from Block 1, while later the following latent variable were selected from

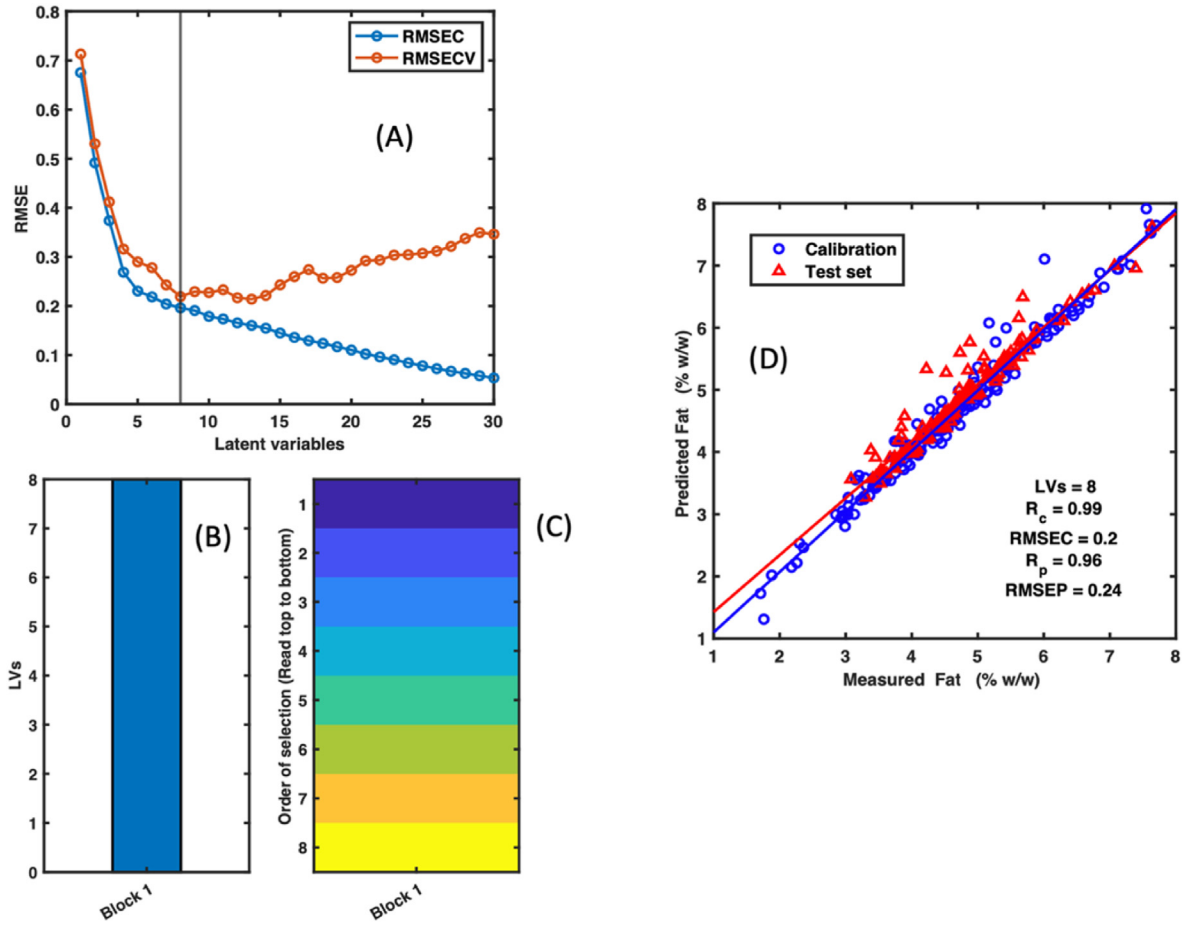


Fig. 5. Results of single block SKPLS analysis performed on 3-way NIRONE 2.0 sensor data for predicting fats content. (A) Cross-validation plot used to select latent variables, (B–C) 8 latent variables were selected, and (D) prediction plot. The change in colour in (C) from dark blue toward yellow indicates the increasing number of latent variables. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

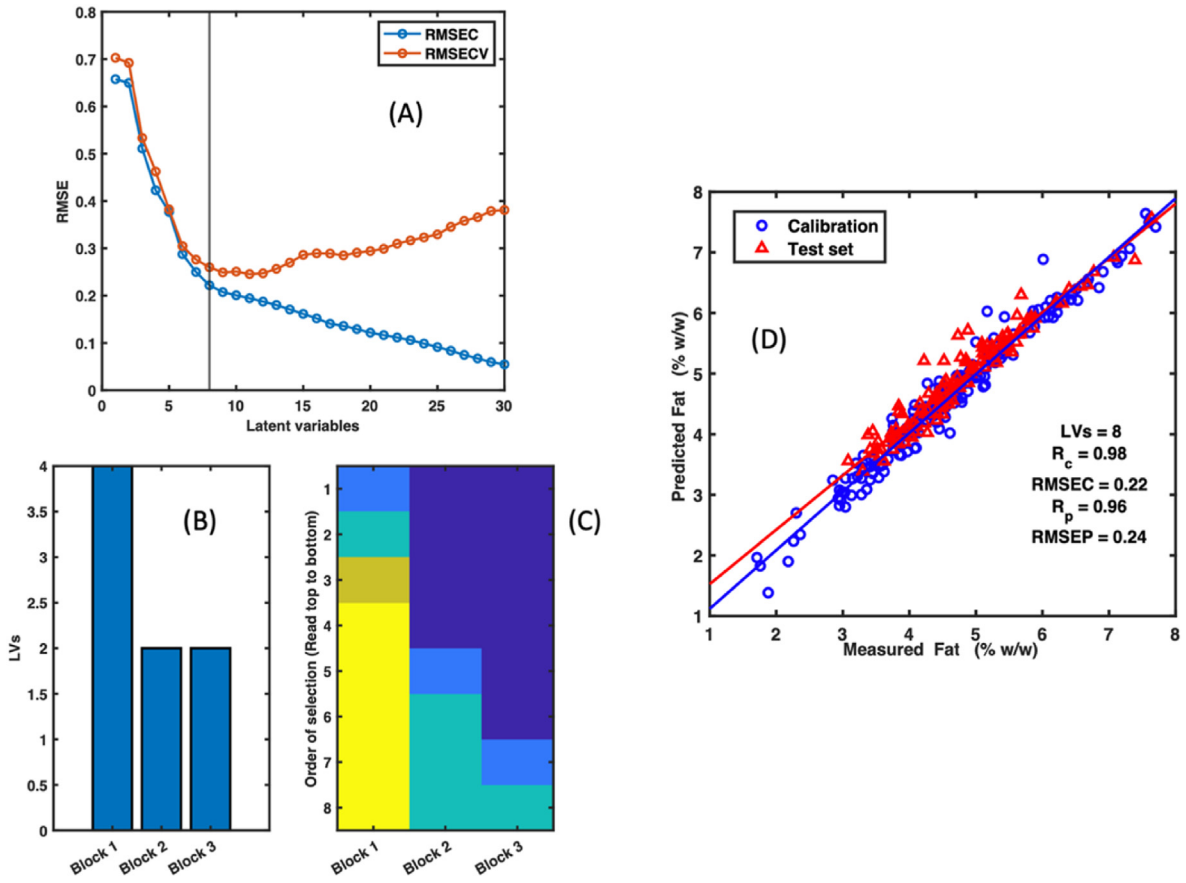


Fig. 6. Results of multiblock SKPLS analysis performed on NIRONE 1.4 (Block 1), NIRONE 2.0 transmission mode (Block 2) and NIRONE 2.5 (Block 3) sensor data for predicting fat content. (A) Cross-validation plot used to select latent variables, (B) 8 latent variables were selected where 4 were from NIRONE 1.4 and 2 each from NIRONE 2.0 and NIRONE 2.5, (C) the order of latent variables extraction, where the initial latent variables were selected from NIRONE 1.4 and later from NIRONR 2.0 and NIRONE 2.5, and (D) prediction plot. The change in colour in (C) from dark blue toward yellow indicates the increasing number of latent variables. An illustration of the component-wise correlations of block candidate scores to the winning block is shown in Supplementary Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Block 2 and Block 3. The prediction errors by using the 3-way data block were slightly lower (Fig. 7D) than not using the 3-way data structure in the analysis presented in section 5.4 (Fig. 6D). Please note that in this example the multiblock multiway analysis was performed using one 3-way data block and two 2-way data blocks, while in practise the method can also be used when there are multiple multiway data blocks available.

4.6. Multi-response modelling with SKPLS

One of the many facets of the SKPLS approach is its capability to handle multiple responses, both continuous response variables for regression cases and dummy-coded matrices of 0 and 1 for classification. The SKPLS gains it capability to handle multiple responses as SKPLS uses the CPLS approach in the backend for efficient subspace extraction, and can even be used with a simultaneous mix of continuous and categorical (dummy-coded) responses. Please note that SKPLS can handle multi-responses for all the analysis cases presented in earlier sections 5.2-5.5. However, to demonstrate with a practical example, the SKPLS was used to analyse the multiway multiblock data to predict fat and protein content in milk. The cross-validation plot allowed selection of 8 latent variables (Fig. 8A), where 7 were from the 3-way data (Block 2) and 1 from the 2-way data (Block 3) (Fig. 8B). In terms of order, the first latent variable was selected from Block 3, while all other latent variables were selected from Block 2 (none from Block 1). The prediction plots for the fat and protein content are shown in Fig. 8D and E, respectively.

4.7. Incorporation of meta information for efficient subspace extraction with SKPLS

One extra feature of the SKPLS is its capability to incorporate meta information during the subspace extraction. Such an inclusion of meta information is possible as the SKPLS uses the CPLS approach to subspace extraction. Meta information is only needed during model training to extract subspaces and not needed in future testing of the model. An example of how to use the meta information using the SKPLS is demonstrated using the apricot dataset where the information about fruit maturity level (in the form of a dummy matrix of 0/1) was used as the meta information to improve the prediction of SSC using NIR and MIR in a multiblock fusion scenario. The results with and without incorporation of meta information in the SKPLS model are shown in Fig. 9. Without inclusion of the meta information, the model cross-validation allowed selection of 3 latent variables (Fig. 9A), where first 1 latent variable was modelled from the NIR data block and then 2 latent variables were from the MIR data block (Fig. 9B and C). The model based on 3 latent variables achieved a prediction error of 0.73% (Fig. 9D). After the incorporation of meta information, only 1 latent variable was selected (Fig. 9E), which was from the MIR data block (Fig. 9F and G), and the model led to lower prediction error of

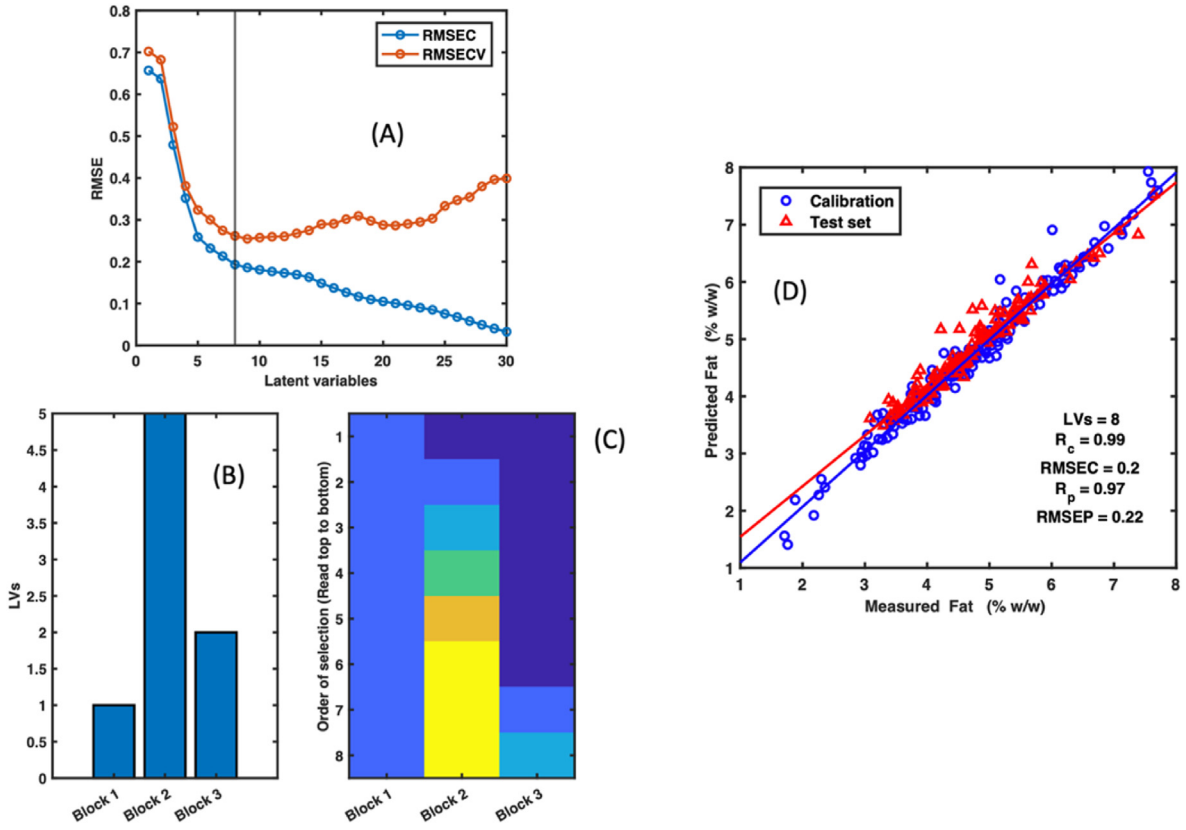


Fig. 7. Results of multiblock multiway SKPLS analysis performed on NIRONE 1.4 (Block 1), NIRONE 2.0 transmission and reflection mode (Block 2) and NIRONE 2.5 (Block 3) sensor data for predicting fats content. (A) Cross-validation plot used to select latent variables, (B) 8 latent variables were selected where 5 were from NIRONE 2.0 and 1 from NIRONE 1.4 and 2 from NIRONE 2.5, (C) the order of latent variables extraction, where the initial latent variables were selected from NIRONE 2.0 and later from NIRONR 2.5 and NIRONE 1.4, and (D) prediction plot. The change in colour in (C) from dark blue toward yellow indicates the increasing number of latent variables. An illustration of the component-wise correlations of block candidate scores to the winning block is shown in Supplementary Fig. 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

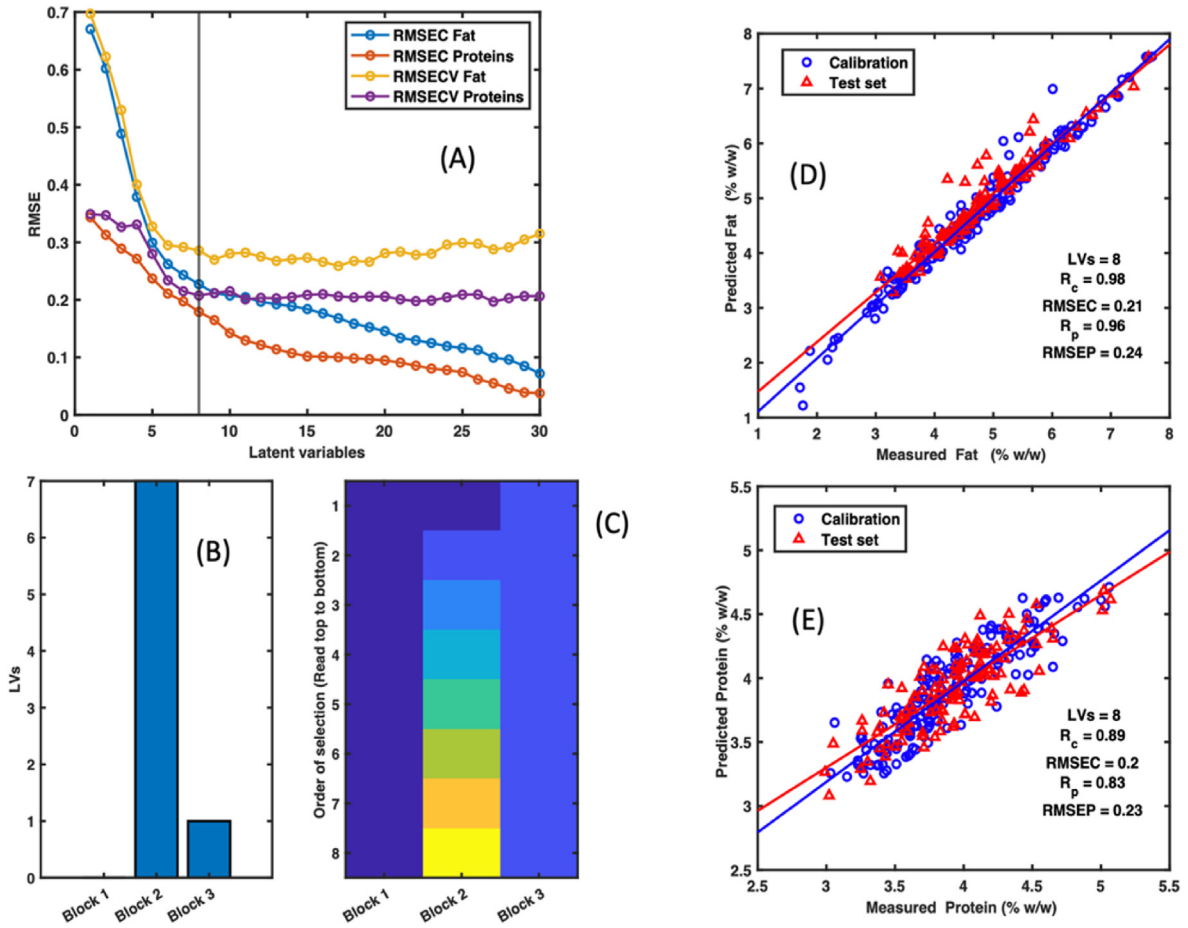


Fig. 8. Results of multi-response multiblock multiway SKPLS analysis performed on NIRONE 1.4 (Block 1), NIRONE 2.0 transmission and reflection mode (Block 2) and NIRONE 2.5 (Block 3) sensors data for simultaneously predicting fats and protein content. (A) Cross-validation plot used to select latent variables, (B) 8 latent variables were selected where 7 were from NIRONE 2.0 and 1 from NIRONE 2.5, (C) the order of latent variables extraction, where the first latent variable was selected from NIRONE 2.5, while later were selected from NIRONE 2.0, (D) prediction plot for fats content, and (E) prediction plot for protein content. The change in colour in (C) from dark blue toward yellow indicates the increasing number of latent variables. An illustration of the component-wise correlations of block candidate scores to the winning block is shown in Supplementary Fig. 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

0.69% (Fig. 9H) for the exact same test set. Reduction of latent variables from 3 to 1 alongside the reduction of prediction error from 0.73 to 0.69% indicates the highly efficient modelling performed by the SKPLS. An additional aspect that can become more prominent when the efficient CPLS subspaces are used in a multiblock setting is that some blocks may become redundant, like the NIR block in this example. Please note that in this example, the meta information was in the form of discrete class labels corresponding to maturity stage of the fruit samples, while in practice the meta information can even be continuous variables such as easy to measure chemical analytes, ambient temperatures, etc. Note also that currently the inclusion of the meta information is only possible to improve the subspace extraction of 2-way type data matrices because of the CPLS algorithm being suitable for 2-way type data only. However, further work is needed for developing a multiway CPLS algorithm for including meta information to improve the multiway type data decomposition. This topic is out of scope for the current study and is only a minor limitation of the algorithm functionality.

5. Conclusions

This study presented a unified version for PLS modelling called Swiss knife PLS (SKPLS) which allows a wide range of PLS modelling operations inside one coherent framework. The backbone of the SKPLS is the stepwise latent variable extraction strategy for single and multiblock analyses called ROSA. The SKPLS can be used for a wide ranges of task such as single block 2-way data modelling as commonly performed by PLS modelling in the analytical chemistry and chemometric domain, single block multiway data modelling which is relevant when the predictor matrix is a multiway array, multiblock data fusion modelling when multiple data blocks of 2-way and/or multiway types are available and the aim is to learn complementary information from different data blocks, multi-response modelling which makes the method capable of both multi-response regression and classification tasks, and finally the capability to incorporate meta information for efficient subspace extraction. These are the most common predictive modelling tasks performed in the chemometric domain using PLS subspace-based approaches. The SKPLS method can be considered a general PLS subspace extraction-based. It is a Swiss knife in the domain of PLS modelling, covering a wide range of modelling scenarios. Since the method does not include deflation of the predictor matrices as in other PLS based algorithms, just like ROSA, the method is naturally expected to be faster than most other deflation based PLS modelling approaches. As SKPLS is based on ROSA, it inherits the

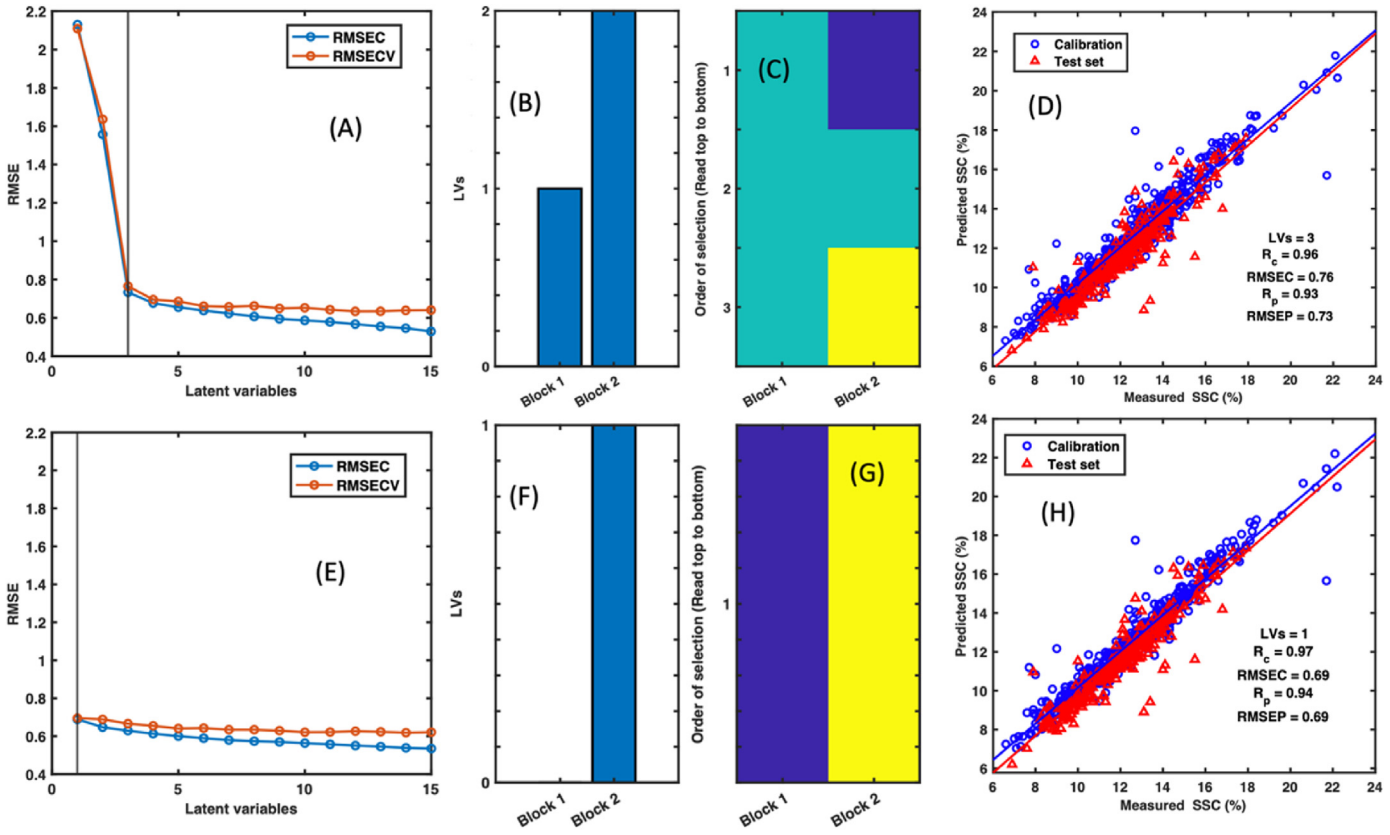


Fig. 9. Results of SKPLS analysis for including meta information about fruit maturity to predict soluble solids content using NIR and MIR data. (A) Cross-validation plot without meta information indicating selection of 3 latent variables, (B) selected latent variables, 1 for NIR and 2 for MIR, (C) order of latent variables selection, where the first latent variable was selected from NIR and later 2 from MIR, (D) prediction plot for fat content without use of meta information, (E) cross-validation plot after inclusion of meta-information, (F–G) selected latent variable, only 1 from MIR, and (H) prediction plot for fat content with use of meta information. The change in colour in (C and G) from dark blue toward yellow indicates the increasing number of latent variables. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

limitation of only using one block for each component. When two blocks are tied in the competition for winning a component competition, this leads to arbitrary block selection. However, since the tie is often due to overlapping information in the input blocks, the predictive power is typically correspondingly little affected. If the exact choice of blocks is important to the user, diagnostic tools are available like plotting the fit of the candidate scores (represented by canonical correlation to the residual response(s)) and possibly manually forcing desired block order for selections. See supplementary information as examples of such plots.

CRediT authorship contribution statement

Puneet Mishra: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Kristian Hovde Liland:** 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2022.339786.

References

[1] S. Wold, et al., Multivariate data analysis in chemistry, in: B.R. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, Springer Netherlands, Dordrecht, 1984, pp. 17–95.
[2] L.L. Simon, et al., Assessment of recent process analytical technology (PAT) trends: a multiauthor review, Org. Process Res. Dev. 19 (1) (2015) 3–62.
[3] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130.
[4] S. Wold, PLS Modeling with Latent Variables in Two or More Dimensions, 1987.
[5] P. Geladi, Chemometrics in spectroscopy. Part 1. Classical chemometrics, Spectrochim. Acta B Atom Spectrosc. 58 (5) (2003) 767–782.
[6] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[7] R. Bro, Multivariate calibration: what is in chemometrics for the analytical chemist? Anal. Chim. Acta 500 (1) (2003) 185–194.
[8] 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, 2020 arXiv preprint arXiv:2001.06530.
[9] T. Skotare, et al., Joint and unique multiblock Analysis for integration and calibration transfer of NIR instruments, Anal. Chem. 91 (5) (2019) 3516–3524.
[10] B. Galindo-Prieto, J. Trygg, P. Geladi, A new approach for variable influence on projection (VIP) in O2PLS models, Chemometr. Intell. Lab. Syst. 160 (2017) 110–124.
[11] L.E. Wangen, B.R. Kowalski, A multiblock partial least squares algorithm for investigating complex chemical systems, J. Chemometr. 3 (1) (1989) 3–20.
[12] P. Mishra, et al., Recent trends in multi-block data analysis in chemometrics for multi-source data integration, Trac. Trends Anal. Chem. (2021) 116206.
[13] P. Mishra, et al., MBA-GUI: A Chemometric Graphical User Interface for Multi-Block Data Visualisation, Regression, Classification, Variable Selection and Automated Pre-processing, Chemometrics and Intelligent Laboratory Systems, 2020, p. 104139.
[14] A.K. Smilde, et al., Common and distinct components in data fusion, J. Chemometr. 31 (7) (2017), e2900.

[15] A. Biancolillo, et al., Extension of SO-PLS to multi-way arrays: SO-N-PLS, Chemometr. Intell. Lab. Syst. 164 (2017) 113–126.
[16] T. Skov, D. Ballabio, R. Bro, Multiblock variance partitioning: a new approach for comparing variation in multiple data blocks, Anal. Chim. Acta 615 (1) (2008) 18–29.
[17] C.A. Andersson, R. Bro, The N-way toolbox for MATLAB, Chemometr. Intell. Lab. Syst. 52 (1) (2000) 1–4.
[18] 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.
[19] Å. Björck, U.G. Indahl, Fast and stable partial least squares modelling: a benchmark study with theoretical comments, J. Chemometr. 31 (8) (2017) e2898.
[20] M. Andersson, A comparison of nine PLS1 algorithms, J. Chemometr. 23 (10) (2009) 518–529.
[21] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemometr. 10 (5-6) (1996) 463–482.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] F. Westad, F. Marini, Validation of chemometric models – a tutorial, Anal. Chim. Acta 893 (2015) 14–24.
[27] K.H. Liland, multiblock: multiblock Data Fusion in Statistics and Machine Learning (https://cran.r-project.org/web/packages/multiblock/index.html) Available from: https://cran.r-project.org/web/packages/multiblock/index.html, 2021.
[28] K. Jørgensen, B.-H. Mevik, T. Næs, Combining designed experiments with several blocks of spectroscopic data, Chemometr. Intell. Lab. Syst. 88 (2) (2007) 154–166.
[29] S. Uusitalo, et al., Evaluation of MEMS NIR spectrometers for on-farm analysis of raw milk composition, in: Foods, 2021.
[30] 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.
[31] B.G. Osborne, Near-Infrared spectroscopy in food analysis, in: Encyclopedia of Analytical Chemistry, 2006.
[32] B. Aernouts, et al., Visible and near-infrared spectroscopic analysis of raw milk for cow health monitoring: reflectance or transmittance? J. Dairy Sci. 94 (11) (2011) 5315–5329.
[33] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1) (1969) 137–148.