Response oriented covariates selection (ROCS) for fast block order- and scale-independent variable selection in multi-block scenarios

Puneet Mishra, Maxime Metz, Federico Marini, Alessandra Biancolillo, Douglas N. D.N. Rutledge

Chemometrics and Intelligent Laboratory Systems 224 (2022) 104551

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Response oriented covariates selection (ROCS) for fast block order- and scale-independent variable selection in multi-block scenarios

Puneet Mishra [a,*], Maxime Metz [b,c], Federico Marini [d], Alessandra Biancolillo [e], Douglas N. Rutledge [c,f]

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

ARTICLE INFO

Keywords:
Multi-block data analysis
Data fusion
Variable selection
Covariance selection (CovSel)
Response-oriented sequential alternation (ROSA)

ABSTRACT

Multi-block datasets are widely met in the chemometrics domain, and several data fusion approaches have recently been proposed to treat them. Apart from exploratory and predictive modelling, a key task in this context is feature selection which involves finding key complementary variables across multiple data blocks that jointly provide a good explanation of the response variables, revealing the key variables of the system. In that direction, a new method called response-oriented covariate selection (ROCS) is proposed here. ROCS is a direct extension of the covariance selection (CovSel) approach to multi-block scenarios, where the choice is based on a competition between variables in different blocks, as is done in the response-oriented sequential alternation (ROSA) method. The uniqueness of the ROCS method is its simplicity, fast execution speed, insensitivity to block order and scale-invariance. The evaluation of ROCS is presented using several multi-block modelling cases and by comparison with other variable selection methods.

1. Introduction

In the domain of analytical chemistry, there is now a growing interest in combining different sensor modalities to better understand the samples or to achieve better predictive models, as in many cases a single modality carries only partial information about the response variables [1–3]. For example, jointly using different spectral sensing modes such as near-infrared, mid-infrared, and Raman spectroscopy [4] leads to multiple data matrices, that is multi-block datasets. In fact, multi-block datasets can have data from any analytical technique which generates multivariate signals and meta-data coming from any other knowledge about the samples, as variety, origin, physical structure, etc. More details on multi-block data and modelling techniques in the chemometric domain can be found in Refs. [2,3].

To analyse multi-block datasets, several latent space based approaches are available; for example, the multi-block variants of traditional principal component analysis (MB-PCA) and partial least-squares regression (MB-PLS) can be used for exploratory and predictive data modelling [5–7], respectively. However, multi-block extensions such as MB-PCA and MB-PLS are highly scale-dependent and may not be of use in cases where the scales of data blocks are different [8–11]. Several approaches to pre-process the datasets in order to "equalize" their range of variability have been proposed [12], but it is still a challenge to find the best normalization approaches for data with different scales. Data with different scales here means data with either different signal intensity ranges or with differences in total amount of variability. To cope with these and other limitations, a growing number of methods for the analysis of multi-block data have been put forward. For instance, in the regression context, advanced chemometric approaches such as sequential- or parallel- orthogonalized partial least-squares regression (SO-PLS and PO-PLS) have been proposed. The former (SO-PLS) [10] is based on sequential PLS models between individual blocks and the response(s), interbedded with orthogonalization steps, allowing to evaluate whether the addition of further blocks provides a significant improvement in modelling. On the other hand, the latter (PO-PLS) [11] extracts components which are either common to all or part of the blocks, or specific to a

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

https://doi.org/10.1016/j.chemolab.2022.104551
Received 9 November 2021; Received in revised form 20 February 2022; Accepted 22 March 2022
Available online 5 April 2022

single block (distinctive). In both cases, the algorithmic structure is such that, besides providing a higher amount of information, especially about the contribution of individual blocks, the models are scale-invariant, though, for SO-PLS, the results can be dependent on the order of the blocks. A wide range of applications can be found in the scientific literature on the use of SO-PLS and PO-PLS, ranging from predictive modelling such as classification [13] and regression [10], as well to their extensions for pre-processing ensembles such as sequential [14] and parallel [15] pre-processing through orthogonalization. The main benefit of the SO-PLS model is when the data block order is meaningful and known so that the important data blocks can be modelled first and the less useful ones later [10]. The data block order is not a problem for PO-PLS approaches in terms of extracting common and distinct information from different data blocks [10,11]. Although both the sequential and parallel PLS based approaches have shown promising potential to analyse multi-block datasets, there is still a big challenge related to the optimisation of such models. By optimisation, we mean the identification of the optimal number of latent variables from each data block to avoid model under- or over-fitting. Furthermore, although the challenge is limited when there are a small number of data blocks such as two or three [16], it does become a problem when there is a substantial number of data blocks since all combinations of latent variables from the different data blocks need to be explored [14], exponentially increasing the computational cost [16]. To deal with the challenges of current sequential and parallel PLS approaches, a new method called response-oriented sequential alternation (ROSA) [16] has recently been proposed. ROSA is a direct extension of PLS modelling to multi-block scenarios and is a scale- and order-independent technique capable of handling many data blocks for the prediction of a single response [16]. However, due to the heuristic it uses, the ROSA algorithm can fall into a local minimum and provide solutions less efficient than SO- and PO-PLS [17].

Apart from predictive modelling based on multi-block techniques, one of the key tasks of multi-block modelling is variable selection [3]. Variable selection serves many purposes in chemometric data analysis, ranging from finding the key variables that are of importance to understand the background chemistry involved in the models or to develop application-oriented low-cost multi-spectral sensors [18,19]. In traditional single block chemometric modelling, a wide range of methods are available for variable selection involving wrapper, filter, embedded, and hybrid approaches [18,19], while in the multi-block domain, there is currently only a limited number of methods available. Two main approaches for variable selection in the multi-block scenario are the use of filters on indices such as variable importance on the projection (VIP) or selectivity ratio (SR), calculated on multi-block models, such as SO-PLS [20] or orthogonal n-block partial least-square (OnPLS) [21], and the extension of the single block covariance selection (CovSel) method [22] to the multi-block scenario, called sequential and orthogonalized covariance selection (SO-CovSel) [23]. SO-CovSel has the advantage over filter approaches based on VIP or SR, that it does not need post-processing/thresholding, since it is a hybrid method that directly extracts individual variables stepwise [23]. However, a key point to note is that since CovSel [22] is a special case of PLS regression relying on covariance maximisation, the SO-CovSel [23] approach is also a special case of SO-PLS [10] which means that SO-CovSel inherits both the advantages and disadvantages of SO-PLS. For example, SO-CovSel is powerful when the block order is important, data blocks are in different scales and only a small number of data blocks are to be processed. However, when there is not a "natural" order of the blocks and there are many data blocks, then the optimisation of the SO-CovSel model becomes even more challenging than for the SO-PLS models. Therefore, to have block order independence and to facilitate model optimisation for many data blocks, the present study proposes a new method called response-oriented covariate selection (ROCS), inspired by both the SO-CovSel [23] and the ROSA [16] methods. The ROCS method inherits from ROSA the advantages of being insensitive to the order of the blocks and being able to easily handle many data blocks [16], while at the same time keeping the scale-invariance resulting from the modelling of each data block independently.

The ROCS method has the advantages of its simplicity, its fast execution speed when processing many blocks, its block order-insensitivity and scale-independence, but it does not guarantee to give the best solution. Examples of ROCS analysis are shown using several multi-block modelling cases. Some out of the box applications of the multi-block variables selection are also shown. Just like CovSel, the ROCS analysis only ranks variables and later a separate cross-validation must be performed using multi-linear regression (MLR) or PLS modelling to select the optimal number of variables to develop predictive models.

2. Theory

A typical CovSel model aims to extract variables from a predictor data block $\boldsymbol{X}$ $(n \times p)$ concerning a response $\boldsymbol{y}$ $(n \times 1)$ by the process of repeated covariance maximisation and orthogonalization. ROCS is a modification of the CovSel approach inspired by ROSA to include $B \geq 1$ mean-centered data blocks $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_B$ of sizes $(n \times p_1), (n \times p_2), \ldots, (n \times p_B)$, respectively. ROCS is yet only defined for a single response $\boldsymbol{y}$, as is ROSA. The selection of the variables is organised as a competition between blocks, where in the first step, the variable carrying maximum covariance with the response variable is selected from each data block; successively, in a competition to minimise the residuals, the block variable giving the smallest residuals is declared the winner and the selected variable is used to orthogonalize all the data blocks and the process is repeated until the desired number of variables is selected or a model optimisation criterion is achieved e.g., minimum RMSECV (Root Mean Squared Error of Cross Validation). The orthogonalization step ensures that all the extracted variables are complementary and carry unique information. The ROCS algorithm on mean-centered data blocks is as follow:

The first step of ROCS i.e., $argmax(\boldsymbol{X}^T\boldsymbol{y}\,\boldsymbol{y}^T\boldsymbol{X})$ is a close estimate of the first latent variable of PLS. To reach this aim, PLS allows any linear combinations of the columns of$\boldsymbol{X}$, while the CovSel engine inside ROCS aims at performing a similar optimisation by allowing only linear combinations of the columns of $\boldsymbol{X}$ in the form $[0, 0, ..1.., 0]$, to carry out variable selection. Finally, the orthogonal projections in step 4 ensure that variances of $\boldsymbol{X}$ and $\boldsymbol{y}$ are captured cumulatively by each step of the algorithm. In a predictive modelling case, the selected variables can also be used for performing tasks such as regression and classification. The codes of the technique will be made available at: https://github.com/puneetmishra2.

3. Datasets

3.1. Prediction of moisture content in pear fruit by fusing information from two portable spectrometers

The dataset consisted of spectral measurements performed on 240 pear fruit samples using two portable spectrometers. The pear dataset was used to show the capability of ROCS to extract variables from spectral fingerprints. More details on samples and the experimental setup can be found in an earlier study [24]. Spectral measurements were performed on intact fruit first using a portable visible and near-infrared (Vis-NIR) spectrometer Felix F-750 (Camas, WA, USA) and then using a short-wave infrared (SWIR) dynamic light projection (DLP) spectrometer NIR Scan Nano (Texas Instrument, USA). After the spectral measurements, a cylindrical disc was cut out at the position of the largest diameter of the fruit and divided into four equal quadrants. One quadrant was used to estimate moisture content (MC) by weighing (XS10001L, Mettler-Toledo GmbH, Giessen, Germany) the samples before and after hot-air oven drying (at 80 °C for 24 h with FP 720, Binder GmbH, Tuttlingen, Germany). The spectral ranges used for ROCS modelling were 720–1000 nm for the Vis-NIR and 1000–1700 nm for the SWIR spectrometer.

Algorithm for response-oriented covariance selection

For $a = 1 : A$, where A is the total number of variables to be extracted

1. In each block $k, k = 1 \ldots B$ select the variable I_k which maximizes the squared covariance with $\boldsymbol{y}$ as $I_k = argmax\left((diag\left(\boldsymbol{X_k}^T \boldsymbol{y}\, \boldsymbol{y}^T \boldsymbol{X}\right)\right)$.
2. $\boldsymbol{y}$ residuals $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_B$ are computed after projecting $\boldsymbol{y}$ orthogonally to each of the selected variables $\boldsymbol{x}_{1,I_1}, \boldsymbol{x}_{2,I_2}, \ldots, \boldsymbol{x}_{B,I_B}$, as $\boldsymbol{r}_j = P^{\perp}_{x_{j,I_j}} \boldsymbol{y}$, for $j = 1, \ldots B$, where $P^{\perp}_{\boldsymbol{x}} = \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^T / \boldsymbol{x}^T \boldsymbol{x}$
3. The block yielding the smallest residual i.e., $i = argmin(\|\boldsymbol{r}_1\|, \|\boldsymbol{r}_2\|, \ldots, \|\boldsymbol{r}_B\|)$ is selected as the winner and the corresponding variable $\boldsymbol{x}_{i,I_i}$ is selected.
4. The collinear information linked to the selected variable $\boldsymbol{x}_{i,I_i}$ is then removed from all data blocks $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_B$ and from the response $\boldsymbol{y}$ by projecting them orthogonally to the selected predictor $\boldsymbol{x}_{i,I_i}$: $\boldsymbol{X}_1^{iter+1} = P^{\perp}_{x_{i,I_i}} \boldsymbol{X}_1^{iter}, \boldsymbol{X}_2^{iter+1} = P^{\perp}_{x_{i,I_i}} \boldsymbol{X}_2^{iter}, \ldots, \boldsymbol{X}_B^{iter+1} = P^{\perp}_{x_{i,I_i}} \boldsymbol{X}_B^{iter}$ and $\boldsymbol{y}^{iter+1} = P^{\perp}_{x_{i,I_i}} \boldsymbol{y}^{iter}$
5. The above process is repeated for A number of variables.

3.2. Spectral data from different sampling forms of rice

The dataset consisted of NIR spectral measurements performed on 200 rice samples, in three different forms i.e., rice with coat, rice kernel and rice flour [25]. The same samples were used for all three measurements; hence, the data becomes a three-block dataset. At first, the spectral measurements were performed on rice with coat, then on the rice kernel and then on the rice flour of the same kernels. The spectra were acquired with a multi-purpose analyzer (MPA) Fourier transform near-infrared (FT-NIR) (Bruker, Germany) spectrometer. The reference property was the protein content, which according to the primary study was measured using the Dumas combustion method [25].

3.3. Reflection and transmission measurements on milk to predict fat content

The dataset consisted of 300 milk samples from 300 cows from all over Flanders (Belgium) [26]. The dataset was used to show the potential of the ROCS method to use the complementary information in the reflection and transmission data to predict milk fat content. The dataset consisted of Vis-NIR (306.5–1710.9 nm) data measured with a Zeiss Corona 45 VISNIR 1.7 diode array spectrometer. The Vis-NIR (350–2500 nm) transmission measurements were performed with a LabSpec spectrophotometer from ASD. The reference measurement for fat content was performed within the 3 days following the spectral measurements [26].

4. Data analysis

ROCS, just like CovSel [22], extracts variables from multi-block datasets. Selected variables can later be used for any kind of analysis, for example regression or classification. The possibility of predictive modelling on the extracted variables has already been proven in earlier works [22,23]. Since ROCS is a special case of the ROSA [16] modelling approach, the performances of ROSA and ROCS models were compared using the three presented data sets. Furthermore, the performance comparison of the models using variables selected by ROCS and SO-CovSel is presented. The ROCS and ROSA optimal models were calculated with a 5-fold cross-validation to judge the optimal number of variables for ROCS and latent variables for ROSA. For SO-CovSel, the optimal number of variables was determined by exploring all combinations of variables in the range of [0-20] using a 5-fold cross-validation. The variable combination leading to the lowest root mean square error of cross-validation (RMSECV) was selected as the optimal combination. The results obtained by ROCS were also compared with the outcomes of a filter approach resulting from the combination of multi-block PLS with variable importance in projection (MBPLS + VIP) [20,23], a fast strategy which is scale-dependent but block order-independent. The MBPLS + VIP includes two main steps: first building a cross-validated MBPLS model

and ranking the variables according to the values of the resulting VIP indices and subsequently, building a new model including only the relevant predictors according to the VIP selection. All data sets were partitioned into calibration (70%) and test sets (30%) using the Kennard-Stone algorithm [27], where all model optimisation and calibration was performed with the calibration set, while the final models were evaluated using the test set. All analyses were performed in MATLAB 2018b (The Mathworks, Natick, MA, USA). The computing system used has an Intel ® Xeon ® W-2133 CPU @ 3.60 GHz with 64 GB of random-access memory.

5. Results and discussion

5.1. ROCS vs SO-CovSel

Firstly, ROCS was compared with the SO-CovSel approach. This was done since SO-CovSel and ROCS are similar in their operation as they both rely on identifying the predictors that carry high-covariance with the response variables. The primary difference in the methods is that SO-CovSel requires a pre-defined block order while ROCS does not, as it extracts variables through a competition among data blocks. But another major difference between the two methods lies in the heuristics used for achieving optimal variable selection. One of the heuristics (also used in this study) for SO-CovSel is to explore all repartitions of variables between blocks and search for the repartition that leads to the minimum RMSECV. The main benefit of this heuristic is that it allows exploring all possible combinations of variables in the defined range and can avoid local minima. However, the approach has the drawbacks that it is time consuming, particularly when the number of blocks is large, and it requires blocks to be ordered beforehand. On the contrary, ROCS uses a faster heuristic based on parallel competition between blocks, however, with drawback of being more prone to falling into a local minimum.

To demonstrate these differences, the ROCS and SO-CovSel analyses were performed on the pear dataset where the aim was to use NIR and SWIR data to predict MC in intact pear fruit. Fig. 1A shows the explained variance in the response variable obtained with ROCS and SO-CovSel. The number of variables selected by SO-CovSel and ROCS to explain up to 90% of variance in the response variable were different, where SO-CovSel selected only 9 variables from the NIR data block (Fig. 1D and E and Fig. 1G) while ROCS selected 17 variables from both the NIR and SWIR data blocks (Fig. 1B and C and Fig. 1F). The variables selected by both analyses can be seen to be of chemical significance to predict moisture in pear fruit [28]. For example, the variable selected by ROCS and SO-CovSel can be related to the overtones of the OH bond which is directly related to the H_2O present in high abundance in fresh fruit and the overtone of the CH and CH_2 bonds which can be related to the sugar molecules present in fresh fruit. To understand the predictive ability of the variables selected by ROCS and SO-CovSel, separate MLR and PLS analysis were carried out on the reduced data set containing only the selected variables (Fig. 2). MLR analysis suggests that the models based on ROCS achieved lower RMSEP compared to the SO-CovSel based models. The results of the PLS cross-validation analysis are shown in Fig. 2C. The cross-validation suggests that although the initial number of variables selected by SO-CovSel and ROCS were 9 and 17, with the PLS analysis, 8 and 11 latent variables were retained. Considering the final PLS models, the one based on ROCS selected variables achieved lower RMSEP than the SO-CovSel one. However, it is to be noted that in the cross-validation plot (Fig. 2C) for ROCS it appears that the total number of latent variables was lower compared to the total number of selected variables, indicating that it may be useful to carry out PLS analysis on the ROCS selected variables to handle the multi-collinearity in ROCS selected variables instead of direct MLR analysis. For SO-CovSel, the simple MLR models should suffice.

To build on the findings from the pear data analysis, the SO-CovSel and ROCS data analyses were also performed on the rice and the milk data sets. The results are shown in Table 1. ROCS either extracted the same number of variables as SO-CovSel or more. However, the optimal PLS models were achieved with fewer latent variables than extracted variables. The performances of the final models for both ROCS and SO-CovSel for the rice and milk data sets were similar. Some results from the ROCS and SO-CovSel analyses of the rice data set suggested that the optimal physical form of the rice samples was the rice flour as all the

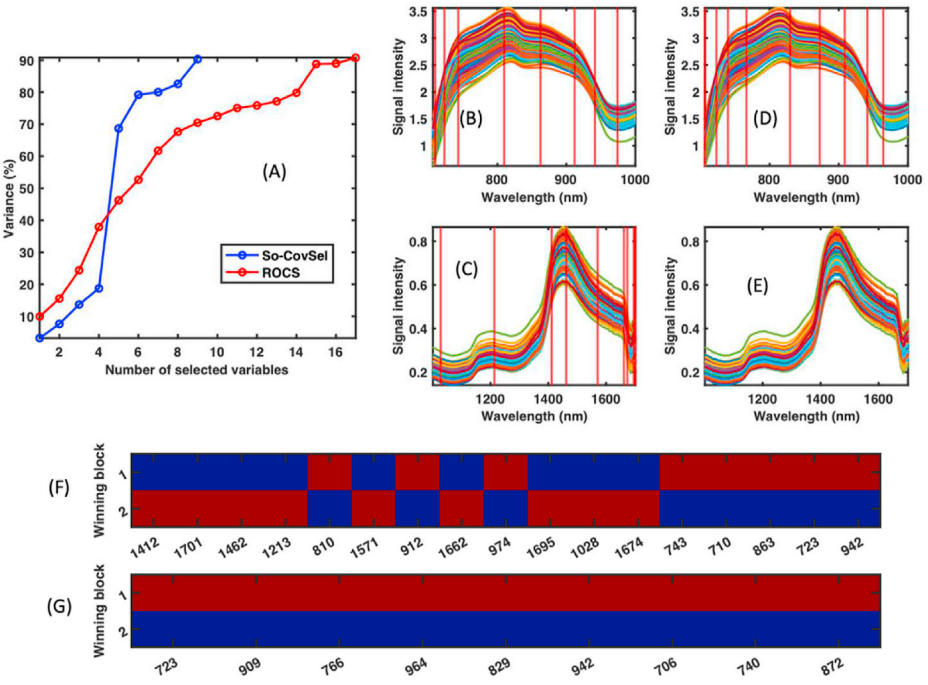

Fig. 1. SO-CovSel and ROCS analysis on Felix and DLPNIR spectrometer data to predict moisture content. (A) Explained variance plot, (B) variables selected from the NIR data block by ROCS, (C) variables selected from the SWIR data block by ROCS, (D) variables selected from the NIR data block by SO-CovSel, (E) variables selected from the SWIR data block by SO-CovSel, (F) winning order of variables and blocks for ROCS, and (G) order of variables selected by SO-CovSel.

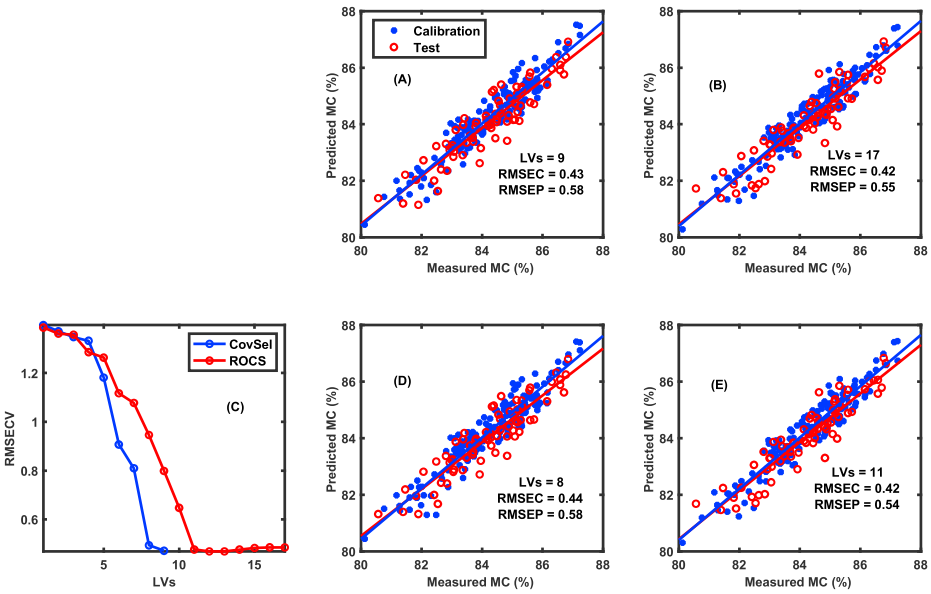

Fig. 2. A summary of MLR (top row) and PLS (bottom row) analyses carried out on the ROCS and SO-CovSel selected variables. MLR analysis: (A) SO-CovSel selected variables, and (B) ROCS selected variables. PLS analysis: (A) cross-validation plots, (D) PLS analysis on SO-CovSel selected variables, and (E) PLS analysis on ROCS selected variables.

Table 1
A summary of SO-CovSel and ROCS analysis performed on rice and milk data set.

Dataset	Variables		RMSEP MLR		PLS-LVs 5-fold CV		RMSEP PLSR	
	SO-CovSel	ROCS	SO-CovSel	ROCS	SO-CovSel	ROCS	SO-CovSel	ROCS
Rice	12 (0 + 0+12)	25 (0 + 0+25)	0.36	0.41	6	13	0.40	0.41
Milk	19 (18 + 1)	19 (7 + 12)	0.08	0.08	15	16	0.08	0.08

variables were selected from the corresponding data block. For milk data set, the selection of variables from both the reflection and transmission data blocks suggests that the two modes play a complementary role in predicting fat content in milk.

5.2. Block order independence of ROCS vs SO-CovSel

ROCS uses the parallel variable selection heuristic which gives natural independence to block order. This means that the user does not need to worry about the arrangement of the blocks prior to the variable selection. To demonstrate that ROCS is block order independent, the analysis was carried out on milk data set by changing the order of blocks. For example, at first the ROCS analysis was carried out with a block order of reflection followed by transmission and then with transmission followed by reflection. As can be noted in the results (Supplementary Fig. 1), the selected variables were the same irrespective of the block order. On the contrary, the outcomes of SO-CovSel performed on the milk data set with changing block order led to the selection of one different variable. Although the predictive performance of the model was intact (Fig. 3) as the overall objective of the SO-CovSel heuristic was to find the global minimum, the selection of a different set of variables (Fig. 3) due to changing block order shows that, unlike the ROCS, the selection

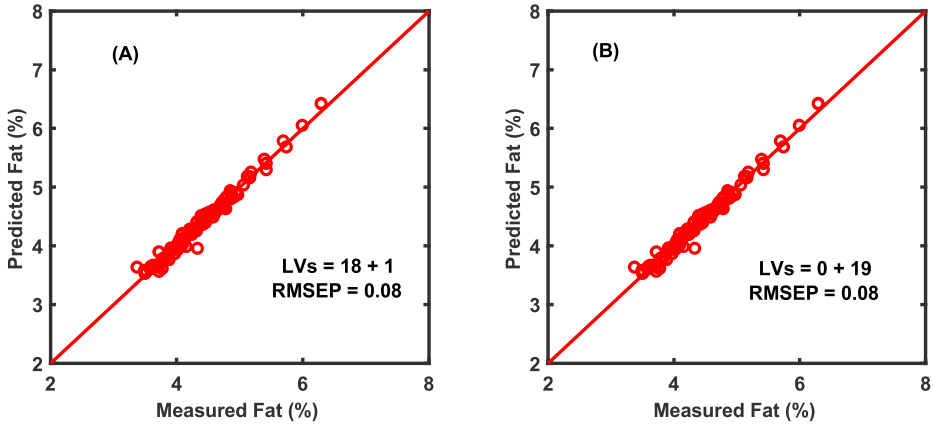

Fig. 3. Model performance on variables selected when changing block order for milk data set by SO-CovSel. (A) Prediction plot for MLR calibrated on variables selected by SO-CovSel following the block order of reflection then transmission, and (B) prediction plot for MLR calibrated on variables selected by SO-CovSel following the block order of transmission then reflection. The latent variables (LVs) are calculated from the variables selected in each block.

provided by SO-CovSel models is inherently affected by the changing block order (Fig. 4). This is not necessarily a drawback, since it was demonstrated in the original paper how such characteristics can be exploited to investigate which variables carry common, local, or unique information. However, one should note that such exploration will come with a cost of high computation time and with large number of data blocks may become impractical unless high end computers are available.

5.3. Comparison of ROCS with MBPLS + VIP

The pear data set was processed with the MBPLPS + VIP analysis for a comparison with the results of ROCS in Section 6.1. The MBPLS + VIP analysis on the pear data was performed on two versions of the pear data. The first version was the unscaled, mean-centered data and the other was the block-scaled data, where each block was normalised by its Frobenius' norm to bring the two blocks to similar data scale. This is also one of the drawbacks of MBPLS-based approaches where the model outcome is highly dependent on the scales of data blocks and requires some form of initial normalization. Conversely, both ROCS and SO-CovSel treat each block separately and so do not require any prior data scaling. The results of 5-fold cross-validation MBPLS analysis on the concatenated pear data blocks suggested different optimal LVs for unscaled (11 LVs) and scaled (16 LVs) forms of data. The different numbers of LVs can be related to the scaling effect as without scaling, the data block with higher scale will be dominant in the model as PLS models involve the step of covariance maximisation and the scale of the data is directly proportional to the covariance estimation. The effect of different LVs used for the construction of the final MBPLS models influence the shape of the VIP vectors as can also be noted in Fig. 5. Furthermore, it can be noted that for the unscaled data where the scale of the NIR data block was higher compared to the SWIR data block, the VIP vector captured finer information for the NIR part of the data (see peaks from 700 to 1000 nm)), while mainly capturing the peak around 1450 nm from the SWIR data block. After block-scaling, several new peaks can also be noted in the SWIR part of the spectrum. In comparison to the ROCS and SO-CovSel analysis presented for same data set in Fig. 1, several regions in the VIP plots are the same as the discrete variables selected by ROCS and SO-CovSel. Note that, although the MBPLS + VIP was explored as the comparative technique, however, in terms of operation, the MBPLS + VIP is very different from the stepwise technique ROCS and SO-CovSel. For example, ROCS, at each forward step of operation, selects a discrete variable and provides parsimonious information such as covariance, explained variances of predictor and responses to judge the importance of each variable. On the other hand, the use of VIP calculated by MBPLS is a filter approach and, as such, shares all the pros and cons of that family of methods. Indeed, the user is still required to use some threshold or exhaustive forward or backward approaches to select the optimal number of variables in the model.

For VIP analysis, a common approach to select the variables is by setting the threshold at 1 and selecting all the variables with VIP score of >1. In the plots presented in Fig. 5, it can be noted that setting a threshold of 1 will results in the selection of variables from both the NIR (<1000 nm) and the SWIR (>1000 nm) spectral regions. However, the selection of variables from both data block was only noted for ROCS analysis, while for SO-CovSel the selected variables were only from the NIR data block.

5.4. Speed of the ROCS method

SO-CovSel and ROCS rely on different heuristics for optimal variable selection. The cross-validation step is the most time-consuming step for SO-CovSel, and it increases exponentially as the number of blocks increases, becoming almost impractical when there are more than just 3 blocks to model. To have a practical comparison between the time requirements for SO-CovSel and ROCS to select the optimal variables, the time recording was performed for all the three data sets during the implementation of the heuristics. As expected, the ROCS heuristic is naturally faster than that of SO-CovSel as there is no pre-defined exploration of variable combinations (Table 2). For all three data sets, ROCS selected the optimal variables in <2 s. The time requirements for data sets with 2 and 3 blocks were similar for ROCS, while for SO-CovSel, the time required for data with 3 blocks was very much higher than for the two blocks data set. Such a high time requirement was due to a direct increase in the number of combinations to be explored. Since ROSA is also a direct extension of CovSel to multiblock scenario, there are no extra steps that would make ROSA slower than CovSel and MBPLS on concatenated data block.

5.5. ROCS vs. ROSA

The ROCS method is a special case of the ROSA method, just as the original CovSel method is a special case of PLS. The only difference between ROCS and the ROSA is that ROCS loading weights are of the form [0,0,0,..1, …0,0], where 1 indicates the selected variable. For data sets with a limited number of samples, as is commonly encountered in the chemometrics domain, the models based on selected variables can achieve similar predictive ability to that of models based on the full spectral range [18,19]. Hence, a comparison of the PLS models calibrated on the ROCS selected variables and the ROSA models based on full spectral ranges was carried out (Fig. 6). The results for the pear data set suggest that the models based on variables selected by ROCS had better predictive ability than the ROSA model based on the full spectral range. For the rice data set, the performance of models based on selected variables was slightly lower than the ROSA models. For the milk data set, the performance of PLS model based on variables selected by ROCS was like that of the ROSA model.

6. Extended discussion

In the domain of multi-block data analysis, one of the challenges in developing a new technique is to consider its ability to handle data of different scales [12]. Such an ability is necessary as in many real-life

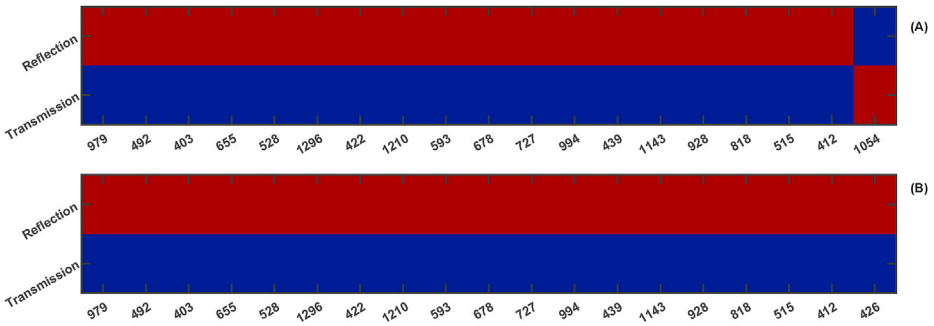

Fig. 4. Influence of changing block order on selected variables of SO-CovSel. The x-axis variables are in nanometres.

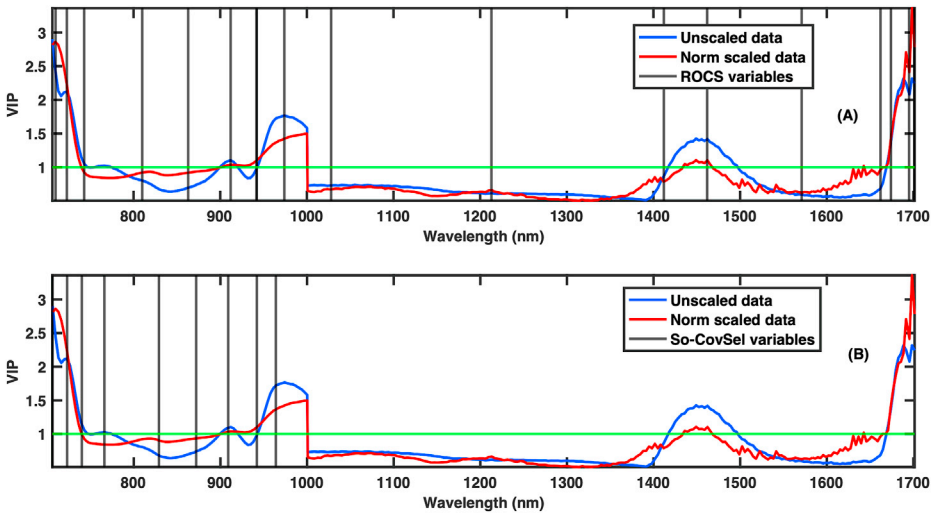

Fig. 5. Variable importance in projection vectors for MBPLS models obtained on unscaled and scaled data blocks. The data are a two blocks data set with NIR (<1000 nm) and SWIR (>1000 nm) data blocks. (A) ROCS variables are highlighted as vertical lines, and (B) SO-CovSel selected variables are highlighted as vertical lines. The green horizontal line indicates the threshold of VIP score 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
A comparison for time requirements for SO-CovSel and ROCS for variable selection on different data sets. All analyses were performed on a desktop computer with an Intel ® Xeon ® W-2133 CPU @ 3.60 GHz and 64 GB of random-access memory.

Data sets	Data matrices size	ROCS for 20 variables (Seconds)	SO-CovSel for 20 variables (Seconds)
Pear	2 blocks (161 × 90 and 161 × 203)	1.2	10.46
Rice	3 blocks (140 × 304 each)	1.35	376
Milk	2 blocks (210 × 299 and 210 × 1876)	1.66	46

situations the data generated by the different analytical instruments may have quite different scales [3,12]. In that direction, one of the most powerful methods is SO-PLS [10,29] which allows handing data of different scales by modelling each data block separately. Akin to SO-PLS, SO-CovSel [23] also allows handling data with different scales by selecting variables from each data block separately. The ROCS method proposed in this study, being like ROSA, is also a scale-independent technique treating each block separately. In ROCS, the first step of covariance maximisation is used for each block. Using the covariance criterion, as in PLS, allows ROCS to find variables which carry large variance and at the same time are related to the response. Although covariance maximisation is scale-dependent, since it is limited to the level of each block, the scale of the blocks does not affect it. To select the winning block in the second step of ROCS, the Y residuals, which are insensitive to the individual scales of the predictor blocks, are used.

Using methods such as SO-PLS [10,29] and SO-CovSel [23], it can be preferable to define the data block order when the user is aware of the importance of each data block and is capable of ordering them before doing the sequential modelling. However, if the order of the block is not known then different orderings of blocks may lead to different combinations of selected predictors and/or suboptimal accuracy [15]. To avoid the block ordering problem, the ROSA approach to multi-block modelling gives equal chances to all blocks [16]. Akin to ROSA, the ROCS approach proposed in this study also gives equal chances to all data blocks by

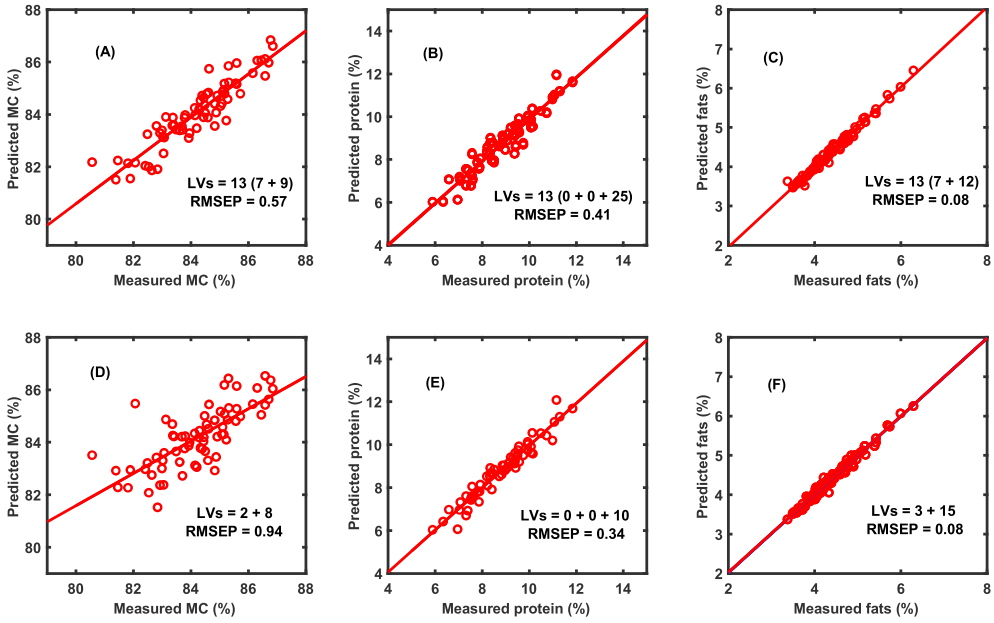

Fig. 6. Performance of ROCS and ROSA models. PLSR model based on ROCS selected variables for (A) moisture prediction in pear, (B) protein prediction in rice, and (C) fat prediction in milk. ROSA models for (D) moisture prediction in pear, (E) protein prediction in rice, and (F) fat prediction in milk.

sending their best variable to the global competition of minimising the residuals. The ROCS method gains its order independence by giving equal chances to all data blocks.

Also, just as in SO-CovSel [23], the ROCS selected variables are as little as possible correlated to each other. This low dependency is maintained both within and among data blocks. In the case of a single data block, the ROCS method converges to CovSel [22].

Optimisation of a total number of variables to be selected with the ROCS methods can be achieved with a global optimisation approach such as using ROCS cross-validation, as was done for CovSel in earlier studies and as shown in the results section of this study. For example, at first ROCS can be performed to extract several variables and then using ROCS cross-validation, the effectiveness of the selected variables can be judged (Example Fig. 1B). Note that the global cross-validation of ROCS is performed to judge the total number of variables to keep for use in the predictive modelling. Such a global optimisation approach is different from what is done in the SO-CovSel approach where the aim is to find the best combination of variables from data blocks. In the SO-CovSel, the variable combinations leading to lowest cross-validation errors are usually selected, however, the criterion to select the variables stays the same i.e., covariance maximisation.

ROCS, as well as all methods belonging to the CovSel [22] family, can handle multiple responses in the within block variable selection step. However, the second step, dedicated to the choice of the winning block, is directly inherited from ROSA [16], and does not handle multiple responses. Research is needed to adapt ROSA's block selection work to with multiple responses. The key point in this step is the definition of residuals, which can be done with many combinations of the Y columns. The interaction between how the residuals is calculated and how ROSA works needs to be studied in detail. This is however beyond the scope of this paper. Furthermore, since ROSA and ROCS are sensitive to local minima, as they use a simple stepwise heuristic, more research should be carried out to modify this heuristic to cope with this problem. In ROCS, the multiblock information is handled by selecting a winner using the minimum residuals at a particular step. One can assume that if multiple blocks at a particular step have similar residuals, then the winner selection is based on the order in which the blocks carrying similar residuals are arranged (particularly in MATLAB implementation). This is also one of the drawbacks of ROSA strategy which is the backbone of ROCS. To be cautious for such cases, as described in ROSA method, it is advised to keep the record of residuals for each block at each step of the ROCS algorithm run. Later, such record of residuals (using plots) can help in diagnosing cases where blocks may have similar residuals.

In the current demonstration of ROCS, all the datasets were spectral data sets which have comparable scales. The final predictive MLR and PLSR models were built directly by concatenating the variables selected from the different data blocks. However, one can assume that if the scales of data varies then the user need to be cautious while building the final predictive models. In that case, the user can build the final predictive models by either using autoscaling of the data or by using scale independent modelling techniques such as SO-PLS or ROSA.

7. Conclusions

A new multi-block method called response-oriented covariates selection (ROCS) was presented. The new method is a block order- and scale-independent technique which can handle many data blocks when selecting variables. The evaluation of the method on real multi-block datasets showed that the method was able to capture chemically relevant features from the different blocks. Furthermore, the selected features were highly predictive of the response variables. The selected features can then be used for a range of tasks such as regression and classification. The applications showed that multi-block variable selection can be used to answer several research questions such as finding the best physical forms of samples for spectral measurements, selecting optimal spectral sensors for dedicated applications, and combining different optical geometries of spectroscopy to predict sample properties. The applications of the method can be foreseen in many domains of sciences where multi-block datasets are encountered and where it is often necessary to find key covarying features to either understand the system or to develop predictive models. The ROCS method also inherits the main drawback of ROSA, namely that the heuristic it uses does not guarantee to find a global optimal solution.

Credit

Puneet Mishra:Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft
Maxime Metz: Conceptualization, Methodology
Federico Marini: Conceptualization, Writing - Original Draft
Alessandra Biancolillo: Conceptualization, Writing - Original Draft
Douglas N. Rutledge: Conceptualization, 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.

Acknowledgements

Authors are thankful to Prof. Jean Michel Roger for his constructive feedback during the development of the work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104551.

References

[1] M. Puneet, D. Passos, Deep multiblock predictive modelling using parallel input convolutional neural networks, Anal. Chim. Acta (2021) 338520.
[2] T. Skotare, et al., Visualization of descriptive multiblock analysis, J. Chemometr. 34 (1) (2020) e3071.
[3] P. Mishra, et al., Recent trends in multi-block data analysis in chemometrics for multi-source data integration, Trac. Trends Anal. Chem. (2021) 116206.
[4] 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.
[5] A.K. Smilde, J.A. Westerhuis, R. Boqué, Multiway multiblock component and covariates regression models, J. Chemometr. 14 (3) (2000) 301–331.
[6] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemometr. 12 (5) (1998) 301–321.
[7] 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.
[8] R. Bro, et al., Data fusion in metabolomic cancer diagnostics, Metabolomics : Official journal of the Metabolomic Society 9 (1) (2013) 3–8.
[9] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemometr. 33 (1) (2019) e3085.
[10] 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.
[11] 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.
[12] M.P. Campos, M.S. Reis, Data preprocessing for multiblock modelling – a systematization with new methods, Chemometr. Intell. Lab. Syst. 199 (2020) 103959.
[13] 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.
[14] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometr. Intell. Lab. Syst. 199 (2020) 103975.
[15] P. Mishra, et al., Parallel Pre-processing through Orthogonalization (PORTO) and its Application to Near-Infrared Spectroscopy, Chemometrics and Intelligent Laboratory Systems, 2020, p. 104190.
[16] 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.

[17] 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, p. 104497.
[18] T. Mehmood, S. Sæbø, K.H. Liland, Comparison of variable selection methods in partial least squares regression, J. Chemometr. (2020) e3226, **n/a**(n/a).
[19] T. Mehmood, et al., A review of variable selection methods in Partial Least Squares Regression, Chemometr. Intell. Lab. Syst. 118 (2012) 62–69.
[20] A. Biancolillo, et al., Variable selection in multi-block regression, Chemometr. Intell. Lab. Syst. 156 (2016) 89–101.
[21] 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.
[22] J.M. Roger, et al., CovSel: variable selection for highly multivariate and multi-response calibration Application to IR spectroscopy, Chemometr. Intell. Lab. Syst. 106 (2) (2011) 216–223.
[23] 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.
[24] 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.
[25] Z. Xu, et al., A calibration transfer optimized single kernel near-infrared spectroscopic method, Spectrochim. Acta Mol. Biomol. Spectrosc. 220 (2019) 117098.
[26] 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.
[27] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1) (1969) 137–148.
[28] B.G. Osborne, Near-infrared spectroscopy in food analysis, in: Encyclopedia of Analytical Chemistry, 2006.
[29] 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.