Received: 25 June 2022 | Revised: 7 August 2022 | Accepted: 26 August 2022

DOI: 10.1002/cem.3441

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

# Swiss knife covariates selection: A unified algorithm for covariates selection in single block, multiblock, multiway, multiway multiblock cases including multiple responses

**Puneet Mishra**[1] | **Kristian Hovde Liland**[2] | **Ulf Geir Indahl**[2]

[1]Food and Biobased Research, Wageningen University and Research, Wageningen, The Netherlands

[2]Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

**Correspondence**
Puneet Mishra, Food and Biobased Research, Wageningen University and Research, Wageningen, The Netherlands.
Email: puneet.mishra@wur.nl

**Abstract**

A novel unified covariates selection algorithm called Swiss knife covariates selection (SKCovSel) is presented. It is suitable for selecting covariates in a wide range of data scenarios such as a single two-way data block, two-way multiblock, multiway, multiway multiblock, selection of covariates along different modes for multiway data blocks and for selecting covariates for all mentioned cases in multiple response scenarios. In the multiblock case, the method can be scale and data block order-independent depending on the preference of the user. For multiway scenarios, the method can be multiway mode order independent, depending on the preference of the user. The proposed SKCovSel algorithm generalises the recent speed improvements from faster CovSel to all mentioned data block cases. It also reformulates the multiway case to do proper deflation and rank one slab selections. Particularly, for modelling of multiblock data sets, the SKCovSel follows the "winner takes all" strategy of the stepwise response-oriented sequential alternation modelling. In the case of multiway data, the SKCovSel strategy considers multiway loading weights after decomposition of a high-dimensional squared covariance matrix to select features across different modes. The algorithmic steps of the methods are presented, and cases of modelling different data types such as single block, multiblock, multiway multiblock, modes selection for multiway data and multiple responses modelling are shown. The method incorporates all popular covariates selection algorithms existing in the chemometric literature.

**KEYWORDS**

feature selection, multivariate, multiway, multiblock

## 1 | INTRODUCTION

In the domain of analytical sciences and chemometrics, multivariate data sets are widely acquired and analysed.[1–3] Furthermore, multivariate data in numerous cases such as measurements with spectroscopic techniques usually result in highly collinear predictors.[4–6] In the domain of chemometrics, latent space-based approaches such as principal

component analysis (PCA)[7] and partial least squares (PLS)[8,9] are popular for deriving robust latent variable subspaces from originally highly multivariate measurement data.

PCA is typically the preferred choice when information about response variables is either not available or for some reason chosen to be ignored. In the presence of one or more response variables, PLS represents the dominant approach for latent space modelling in the analytical chemistry and chemometric domains.[10,11] Both PCA and PLS represent bilinear models that include a set of scores (the latent variables) and a corresponding set of loading vectors describing how the latent variables relate to the original multivariate measurements and vice versa. The scores and loadings are essential for both data visualisation purposes and predictive model building.

Apart from the latent space modelling for data visualisation and predictive modelling, one of the main aims behind data modelling in the domain of analytical chemistry and chemometrics is to achieve insights into the key features present in the data.[12,13] The identification of key features can serve a wide range of purposes, for example, improving insights about the system, improving model accuracy, improving model robustness and discovery of low-cost selective sensors such as multi-spectral sensing systems.[12,14–20] In the domain of chemometrics and analytical chemistry, a wide range of methods are available for performing feature selection in multivariate scenarios, which can be classified as wrapper, filter and embedded methods.[12,13,16]

Most of the feature selection methods in chemometrics literature[21] involve post-processing of the regression coefficients obtained with PLS decomposition. One can assume that it is highly important for such methods that a proper optimisation of the PLS model is performed at first-hand. However, one family of embedded methods that outperforms other feature selection methods based on the simplicity of operation and direct alignment with the subspace modelling approach PLS is the covariates selection (CovSel) approach.[17] The CovSel family of methods (Figure 1) does not fit in the framework of feature selection methods that are based on post-processing of PLS regression coefficients as CovSel is a hybrid method where feature selection and modelling goes together.

The covariates selection is a Gram–Schmidt (GS) process,[22] similar to the classical PLS modelling where at each step in the covariance maximisation, the associated weight vector is chosen as a (sparse) standard basis vector in the direction of the variable of maximum covariance with the response(s). Subsequently, just like in the NIPALS PLS algorithm,[9] the data matrix is deflated, and the process continues for extracting the desired number of variables according to minimisation of the (residual) covariance with the response(s). When considered as a GS process, the CovSel (like PLS) can be extended to handle multi-block[23] and multiway problems.[24] It should be noted that extensions of the CovSel idea to multiblock[19,25] and multiway[26] data scenarios are already addressed in the literature.

To supply a summary, Figure 2 presents all data scenarios where extensions of the CovSel modelling can be performed such as the selection in case of two-way data, multiway data and multiblock data. In the current state of the art, different covariate selection approaches are available as distinct methods that seems to miss a unified mathematical formulation to be considered as a unified tool suitable for all major types of data sets. An extended algorithm covering all the mentioned versions should be of considerable interest to the practical user and contribute to evolving applications of the CovSel approach not only in chemometrics but also making it being considered as a valuable methodology within other domains of empirical modelling.

In the covariate selection methods,[17–19,25–27] the variable selection is a stepwise process much like the selection of latent variables (the scores) in PLS. Hence, the CovSel version for multiblock data allows for selecting a variable from

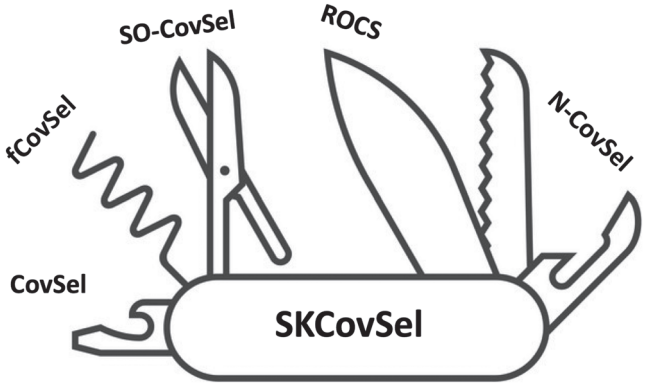


FIGURE 1 Swiss knife covariates selection algorithm spanning various covariates selection methods.

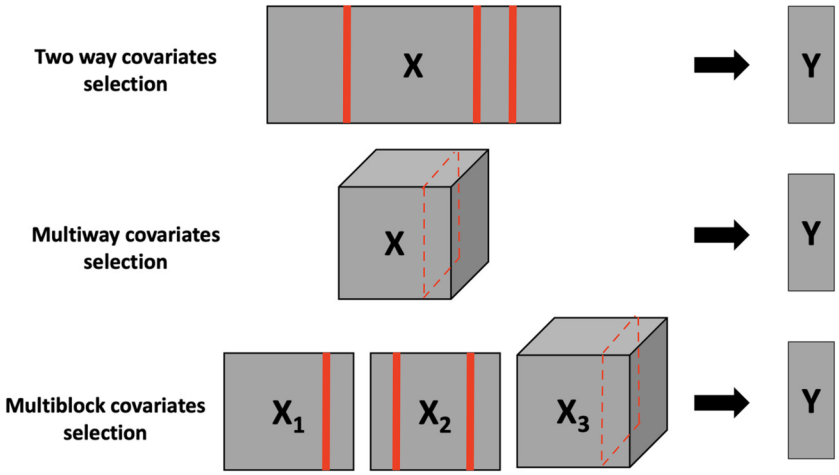


FIGURE 2 A summary of data set scenarios where covariate selection can be deployed to select highly co-varying features. The red lines depict the selected features, which can be columns of a matrix or slices of higher order arrays

any block in each step. This corresponds to the fundamental stepwise idea in the recent multiblock method known as response-oriented sequential alternation (ROSA),[28] where PLS-like latent variables (scores) are calculated for each individual data block and compared for selection in terms of how well they fit the response variable(s). The idea of the ROSA approach is also attractive for establishing a unified CovSel algorithm much similar to the algorithm of the unified PLS method called the Swiss Knife PLS (SKPLS).[29] This method is designed to cover all major PLS modelling scenarios. Furthermore, the re-orthogonalisation step speeding up both the PLS and the ROSA have also been used to obtain the fast covariates selection (fCovSel),[27] (by avoiding deflations of the predictor data matrix). In combination, these methods (the PLS, the ROSA and the fCovSel) together with the response oriented covariates selection (ROCS)[25] and SKPLS pave the way for a unified covariates selection algorithm.

The present study aims at developing and testing a unified covariates selection algorithm called Swiss knife covariates selection (SKCovSel) capable of selecting covariates in a wide range of measurement data scenarios such as a single two-way data block, two-way multiblock data, multiway data, multiway multiblock data (including covariate selection along different modes of the multiway data blocks) and for covariate selection for all of these scenarios in the case of multiple responses. In the case of multiblock data, the algorithm allows the user to choose whether to use scale independent and data block order independent calculations. For multiway scenarios, the algorithm can be specified to operate multiway mode order independently, according to the preference of the user.

The algorithmic steps of the method are presented below, and modelling with the different data types such as single block, multiblock, multiway multiblock, modes selection for multiway data and multiple responses modelling are demonstrated in a case study. The main point to be highlighted is that the proposed SKCovSel algorithm includes the model building capabilities of the various covariates selection algorithms known in the chemometric literature (Figure 1).

## 2 | SWISS KNIFE COVARIATES SELECTION ALGORITHM

The proposed swiss knife covariates selection (SKCovSel) algorithm can be considered as an extension of the recent and fast fCovSel[27] approach to faster covariate selection with multiblock and multiway data. Particularly for modelling of multiblock data sets, the SKCovSel follows the "winner takes all" strategy of the stepwise ROSA algorithm.[28] In the case of multiway data, the SKCovSel strategy considers multiway loading weights after decomposition of a high-dimensional covariance matrix (by using SVD or PARAFAC) to select features across different modes. In the following, all matrices and higher-order arrays are denoted with bold italics uppercase letters such as $\boldsymbol{X}$. All vectors are denoted with bold italics lowercase letters such as $\boldsymbol{w}$.

Define $\boldsymbol{Y}$ $(N \times K)$ as the response matrix, $B$ as the number of (centred) data blocks $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_B$ and let $A$ be the desired number of features to be extracted. Note that data blocks can be of any dimensionality; two-way or multiway, and they are assumed to be mean centred along the sample mode. In unfolded form (each sample being vectorised) the

blocks will have dimensions $(N \times J_b)$. The responses $\boldsymbol{Y}$ are also assumed to be mean centred. The tensor notation, $\underline{\mathbf{B}}$, is associated with the tensor dot product exemplified in Liland et al.[30]

**Algorithm for Swiss Knife Covariates Selection (SKCovSel)**

for a=1:A - loop over $A$ components
  for b=1:B - loop over $B$ blocks
    $\underline{\mathbf{C_b}} = \sum_k (\underline{\mathbf{X_b}}^t \mathbf{Y}_k)^2$ - sum of squared covariances (folded block $\underline{\mathbf{X_b}}$)
    if $S(b) == 1$ - select single variable (combination), where $S(b)$ indicates the feature dimension
      $(m_b, s_b) = (arg)max_{j_b}(vec(\underline{\mathbf{C_b}}))$ - maximum value and its position
      $\mathbf{t}_b = \mathbf{X_b}[:, s_b]$ - select candidate score (unfolded block $\mathbf{X_b}$)
      if $a > 1; \mathbf{t}_b = \mathbf{t}_b - \mathbf{T}\mathbf{T}^t\mathbf{t}_b$; end - orthogonalise on previous scores
      $\mathbf{t}_b = \mathbf{t}_b / \text{norm}(\mathbf{t}_b)$ - normalise score
      $r_b = \text{cancor}(\mathbf{t}_b, \mathbf{Y})$ - canonical correlation to responses
      $\mathbf{w}_b = \{0\}, \mathbf{w}_b(s_b) = 1$ - candidate loading weights
    else - select a slice
      if $\#dim(b) == 3$ - three-way block
        $\mathbf{w}_1, \mathbf{w}_2 = \text{SVD}(\underline{\mathbf{C_b}}, 1)$ - one component SVD
      else - multi-way block
        $\mathbf{w}_1, \mathbf{w}_2, ... = \text{PARAFAC}(\underline{\mathbf{C_b}}, 1)$ - one component PARAFAC
      end - end inner conditional handling
      for m = 1:M - loop over $M$ modes
        $(m_{m,b}, s_{m,b}) = (arg)max_j(\mathbf{w}_m)$ - maximum value and its position
        $\mathbf{w}_{m,b} = \{0\}, \mathbf{w}_{m,b}(s_b) = 1$ - candidate loading weights
        $\mathbf{t}_{m,b} = \underline{\mathbf{X_b}}\mathbf{w}_1 ... \mathbf{w}_{m,b} ... \mathbf{w}_l ...$ - candidate scores (one mode with selection)
        if $a > 1; \mathbf{t}_{m,b} = \mathbf{t}_{m,b} - \mathbf{T}\mathbf{T}^t\mathbf{t}_{m,b}$; end - orthogonalise on previous scores
        $\mathbf{t}_{m,b} = \mathbf{t}_{m,b} / \text{norm}(\mathbf{t}_{m,b})$ - normalise score
        $r_{m,b} = \text{cancor}(\mathbf{t}_{m,b}, \mathbf{Y})$ - canonical correlation to responses
      end - end mode loop
    end - end outer conditional handling
    $v_a = argmax_m(r_{m,b})$ - winning mode
    $\mathbf{t}_b = \mathbf{t}_{v_a,b}$ - winning mode score
    $\mathbf{w}_b = \mathbf{w}_{v_a,b}$ - winning mode loading weight
    $r_b = r_{v_a,b}$ - winning mode correlation
    $\mathbf{p}_b = \underline{\mathbf{X_b}}^t \mathbf{t}_b$ - winning mode loadings
  end - end block loop
  $v_a = argmax(r_b)$ - winning block
  $\mathbf{T}(:, a)_{v_a} = \mathbf{t}_{v_a}$ - winning block score
  $\mathbf{W}(:, a)_{v_a} = \mathbf{w}_{v_a}$ - winning block loading weights
  $\mathbf{P}(:, a)_{v_a} = \mathbf{p}_{v_a}$ - winning block loadings
  $\mathbf{Q}(:, a)_{v_a} = \mathbf{Y}^t\mathbf{T}(:, a)_{v_a}$ - winning block Y loadings
  $\mathbf{Y} \Leftarrow \mathbf{Y} - \mathbf{T}(:, a)_{v_a}\mathbf{Q}(:, a)^t_{v_a}$ - Y deflation
end - end component loop
$\underline{\mathbf{R}} = \underline{\mathbf{W}}(\underline{\mathbf{P}}^t\underline{\mathbf{W}})^{-1}$ - projections for score prediction$*$
$\underline{\mathbf{B}} = \text{cumsum}(\underline{\mathbf{R}}\mathbf{Q}^t_A)$ - regression coefficients
$\underline{\mathbf{B}}_0 = \underline{\bar{\mathbf{Y}}} - \underline{\bar{\mathbf{X}}}\underline{\mathbf{B}}$ - mean compensation

* Calculation of projection for score predictions ($\underline{\mathbf{R}}$) assumes winning loadings and loading weights stacked with matrices of zeros for the loosing blocks.

## 3 | COMMENTS ON THE SKCOVSEL ALGORITHM

The SKCovSel algorithm provides a set of selected variables/features useful for data visualisation purposes and predictive classification and regression model building including cross-validation to decide the model complexity in terms of the number of selected variables/features. As the method can handle multiway datasets, the selected features can be of any type ranging from a vector to multiway features with *n-1* modes for *n*-way data. Furthermore, in the MATLAB implementation of the algorithm (codes to be added at https://github.com/puneetmishra2), the user can also define the restriction of modes for multiway data in which the features are selected. In the case of a three-way array, the selected feature can either be a single variable or a rank one latent variable of a 2-D slab/slice defined by some variable along a particular mode. It should be noted that such slices may be selected repeatedly to produce additional latent variables.

We would like to stress that the foremost novelty of the SKCovSel algorithm is that it covers all the major cases of covariates selection approaches in the scientific literature. In the case of a single two-way data block, the algorithm will provide exactly the same solution as the standard covariates selection method[17] (just faster as we take advantage of the computationally more efficient steps in the fCovSel algorithm[27]). For problems including multiple two-way type data blocks, that is, a multiblock dataset without any predefined block order, the computationally efficient SKCovSel algorithm will provide exactly the same solution as the ROCS.[25] If the user sets the predefined order of blocks from which features need to be selected, the algorithm will provide a computationally efficient solution consistent with the sequential orthogonalised covariates selection (SO-CovSel).[19] Furthermore, when the data set is multiway the algorithm will efficiently provide a solution of the N-CovSel[26] problem computationally consistent with the other CovSel-versions. In the MATLAB implementation available online, all the major cases of covariates selection are covered by the SKCovSel algorithm.

Note: For the selection of higher order features from multiway data, the strategy implemented in the SKCovSel leads to a solution that is slightly different from the solution provided by earlier N-CovSel algorithms. For example, in the earlier strategy, the squared covariance estimation for higher order features is performed for one feature at a time, hence, requiring a loop for the complete estimate in a particular mode. As a second step, the variable carrying the maximum squared covariance is selected. The last step of the earlier N-CovSel algorithm removes the information of the selected feature by a deflation step in the feature modes. To assure the implementation of a Gram–Schmidt process producing orthogonal features, however, the deflation/orthogonalization operations should always be conducted in the sample mode of both for $X$ and $Y$. This is assured by (1) unfolding the feature modes, (2) pre-multiplying by ($\boldsymbol{I}$ - $\boldsymbol{S}$), where $\boldsymbol{I}$ is the identity and $\boldsymbol{S}$ is the desired rank one $N \times N$ projection matrix, and (3) refolding the result. These operations do not seem to be implemented correctly in the earliest version of the N-CovSel.[26]

In the corrected strategy, we estimate the squared covariance directly for all features by unfolding the multiway data before estimating $diag(\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X})$. Then, we reshape the squared covariances to have the same dimensions as the feature modes of the multiway array. Finally we perform a one factor SVD or PARAFAC depending on the number of modes of the squared covariance matrix. The results of the SVD or PARAFAC decomposition are the normalised loading weights for each mode that can be used directly for selecting features along different modes by finding the variable carrying maximum absolute loading weight. Once the variable is selected, the corresponding loading weights are used to estimate the scores.

The estimated scores can be used directly to deflate the responses. Furthermore, the re-orthogonalization approach of the SKCovSel algorithm does not require deflations of the predictor matrix which makes the algorithm faster in the fashion of fCovSel. In summary, the new strategy for solving the N-CovSel problem conducts the selection of higher order features in a non-deflating algorithm to obtain a unified and consistent SKCovSel framework. It should also be noted that the computational efficiency of the new non-deflating N-CovSel is considerably better than the original N-CovSel algorithm[17–19,25] conducting deflations in the multiway data structure.

The current CovSel approaches are in general sensitive to outliers similar to PLS modelling. This is because the first step of CovSel is the estimation of the covariance, which is estimated as $\boldsymbol{XY}$; hence, the presence of outlying samples in the data can influence the covariance estimation and the selection of the features. Currently, the ideal approach is to do some form of outlier removal before the CovSel analysis such that the estimation of covariance is minimally influenced by the outlying samples, thus, allowing to have robust selection of features.

## 4 | DATA

To show the unified SKCovSel for covariate selection a milk data sets was used. The milk dataset is a perfect example of multiblock multiway multiple response data set, hence suitable to show all the capabilities of the SKCovSel method. The milk data set has spectral, protein and fat measurements performed on 296 milk samples.[31] Three portable spectral sensors were used to collected spectral data on milk samples: NIRONE 1.4 (1100 to 1400 nm), NIRONE 2.0 (1550 to 1950 nm) and NIRONE 2.5 (2000 to 2450 nm) from Spectral Engines (Helsinki, Finland). All measurements were performed in transmission mode except for the NIRONE 2.0, for which more measurements of the same samples were performed in reflectance mode. Due to the extra measurement in the reflection mode, the data from NIRONE 2.0 can be considered as 3-way data (samples × spectral variable × measurement mode). More information on the data set and reference protein and fat analysis protocol can be obtained in the earlier study.[31] Data are summarised in Table 1. Data were partitioned into calibration (60%) and test (40%) set to show the capability of the extracted features and the regression coefficients to predict the multiple responses in the milk. All data analyses were carried out in MATLAB.[32]

In the following part of the manuscript, the capability of the SKCovSel will be shown for selecting covariates for single two-way and multiway type data blocks, jointly for multiblock two-way and multiway data blocks and covering single and multiple response cases. Please note all analyses presented are performed using the SKCovSel codes provided in GitHub, verifying that the supplied code is fully functional.

## 5 | RESULTS

### 5.1 | SKCovSel for two-way type data for single and multiple responses

In the case of two-way type data, the SKCovSel performs the standard CovSel variable selection. As an example, the SKCovSel analysis was carried out on a two-way dataset (NIRONE 2.0) to select wavelengths that are predictive of fat content and jointly fat and protein content. The results for calibration and test set as a function of features extracted are shown in Figures 3 (for predicting only fat) and in 4 (for jointly predicting fat and protein content). The explained

**TABLE 1** A summary of near-infrared milk data set

| | NIRONE 1.4 | NIRONE 2.0 | NIRONE 2.5 | Protein | Fats |
|---|---|---|---|---|---|
| Spectral range (nm) | 1100–1350 | 1550–1950 | 2000–2450 | [a] | [a] |
| Data shape | 296 × 126 | 296 × 201 × 2 | 296 × 226 | 296 × 1 | 296 × 1 |
| Reference range | [a] | [a] | [a] | 3.90 ± 0.41 | 4.71 ± 1.10 |

[a]Spectral data are associated with a "Spectral range (nm)," while reference measurements are associated with a "Reference range."

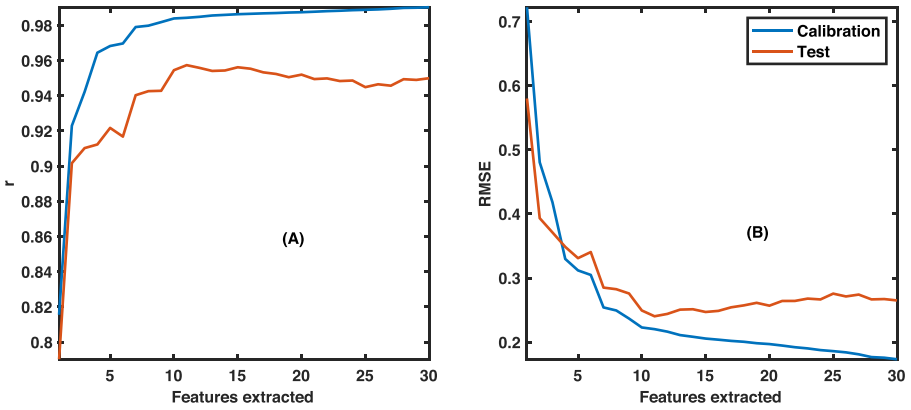


**FIGURE 3** Swiss knife covariates selection (SKCovSel) analysis for analysing two-way data to predict fat content. (A) Correlation coefficients between predicted and actual values for the models based on selected features, and (B) root-mean-squared error (RMSE) estimated with predicted and actual values for the models based on selected features

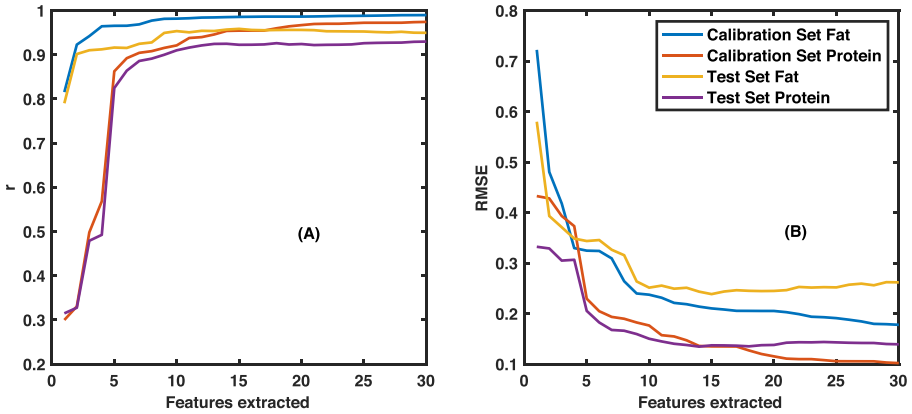


**FIGURE 4** Swiss knife covariates selection (SKCovSel) analysis for analysing two-way data to predict fat and protein content. (A) Correlation coefficients between predicted and actual values, and (B) root-mean-squared error (RMSE) estimated with predicted and actual values

**TABLE 2** A summary of features selected by SKCovSel and the CovSel algorithm published in earlier studies[19]

| | SKCovSel | CovSel |
|---|---|---|
| Fat prediction | 1696, 1728, 1844, 1554, 1774, 1872, 1944, 1652, 1716, 1818 | 1696, 1728, 1844, 1554, 1774, 1872, 1944, 1652, 1716, 1818 |
| Fat and protein prediction | 1696, 1728, 1844, 1554, 1654, 1942, 1774 | 1696, 1728, 1844, 1554, 1654, 1942, 1774 |

variance in the responses reached >95% with only 10 features. For both the cases, the correlation coefficient ($r$) at first increased and later stabilised. Similarly, the root-mean-squared error (RMSE) at first decreased then stabilised showing that most of the learning happened in the initial features extracted. That is normal as the features selected by CovSel carry decreasing amount of covariance; hence, most of the covariance is limited to the initial features. We also compared the features selected for the two-way data with the codes for CovSel available in an earlier study[19] and found that both algorithms led to the selection of the same features and in exactly the same order (Table 2). Note that the selected features (Table 2) for protein and fat content are also chemically relevant as most of the features correspond to overtones of OH, CH and NH bonds,[33] present in abundance in macro-molecules such as fat and protein.

## 5.2 | SKCovSel for single block multiway type data

The SKCovSel approach for multiway data performs an N-CovSel type analysis. For multiway data, the method allows to select features in different modes. For example, for a three-way array ($I \times J \times K$), the features can be either a 1-D column selected as *(J,K)*, or the slices for second and third modes. Note that the *(J,K)* type feature is exactly the same as selecting the feature on the unfolded multiway data. To show the capability of SKCovSel, the features were selected for jointly explaining the fat and protein contents in milk. The feature selection in all three cases (Figures 5–7) showed increasing correlation coefficients and decreasing RMSE as a function of number of features selected. A summary of the features is further presented in Table 3. The selected features of type *(J,K)*, for example, the first selected feature (1696,1) is a feature corresponding to 1696 nm for data measured in reflection mode. The second selected feature (1666, 2) indicated a feature corresponding to 1666 nm for data measured in transmission mode. For mode 2, the feature selected, for example, 1696 indicates 1696 nm measured in both transmission and reflection mode. For mode 3, the feature selected, for example 1, indicates a feature selection of the whole reflection mode of spectral data measured in the spectral range of 1550–1950 nm. In Table 3, it can be noted that for mode 3, the same feature was selected multiple times. This is since for multi-dimensional features, the SKCovSel method extracts rank one covariates each time. Although, the features are repeatedly selected, the information learned is always complementary.

Note that in the presented case the multiway data was a 3D array, hence, either columns (1D) or slices (2D) can be extracted as the features for that scenario. However, in general, the SKCovSel algorithm can select from 1D to $(n-1)$D

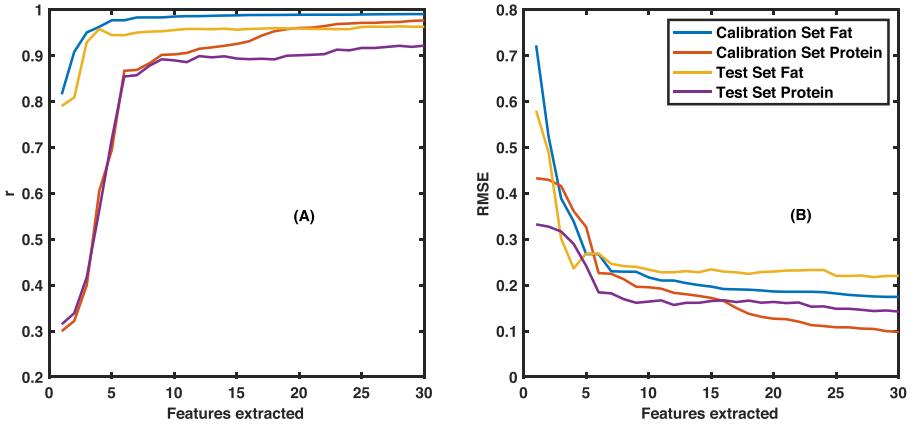


**FIGURE 5** Selection of features in mode 1 for multiway data to jointly predict fat and protein. (A) Correlation coefficients between predicted and actual values, and (B) root-mean-squared error (RMSE) estimated with predicted and actual values

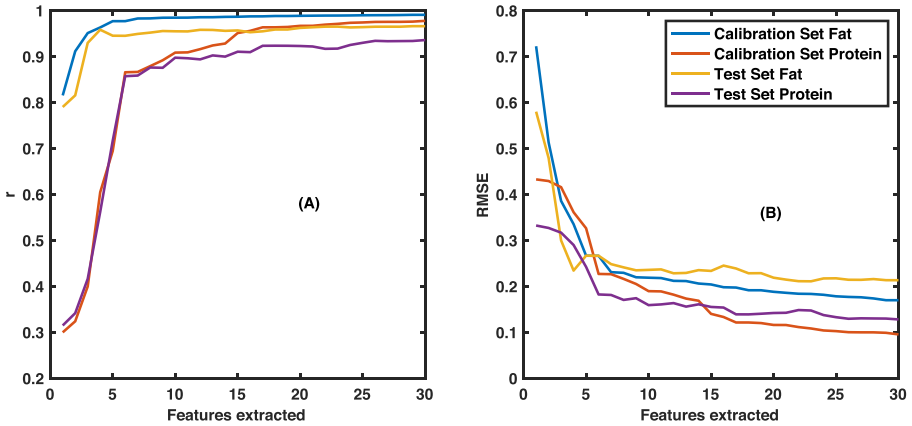


**FIGURE 6** Selection of features in mode 2 for multiway data to jointly predict fat and protein. (A) Correlation coefficients between predicted and actual values, and (B) root-mean-squared error (RMSE) estimated with predicted and actual values

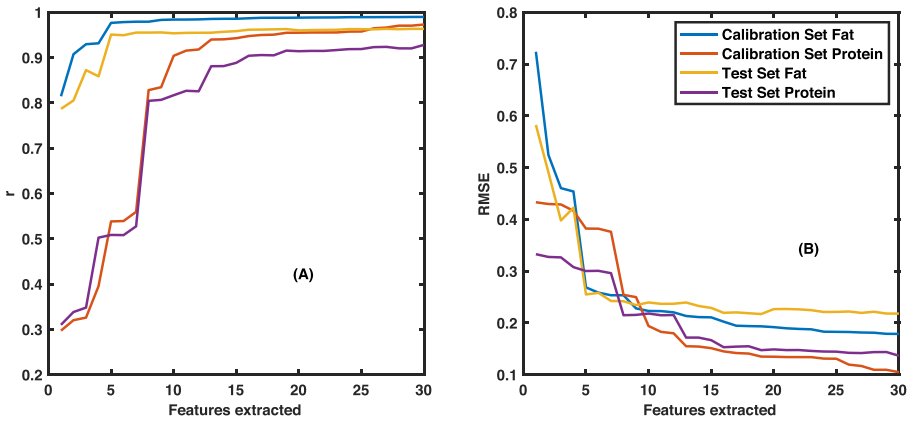


**FIGURE 7** Selection of features in mode 3 for multiway data to jointly predict fat and protein. (A) Correlation coefficient between predicted and actual values, and (B) root-mean-squared error (RMSE) estimated with predicted and actual values

type features for a *n*D multiway array, where *n* is the number of modes of the multiway array. Application of the SKCovSel algorithm with 4D data will allow for either selecting 1D type features or 3D type features. However, for 4D data, features can be of 1D, 2D and 3D type. The selection of 2D type features for 4D type data is also possible, however, will require slight modification to the algorithm. The key idea behind selecting feature of dimensions between 1D and

**TABLE 3** A summary of features selected for multiway data in different modes

| Modes | Selected features |
|---|---|
| (2,3) | (1696,1), (1666,2), (1550,1), (1842,1), (1724,1), (1654,1) |
| 2 | 1696, 1660, 1550, 1842, 1724, 1654 |
| 3 | 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1 |

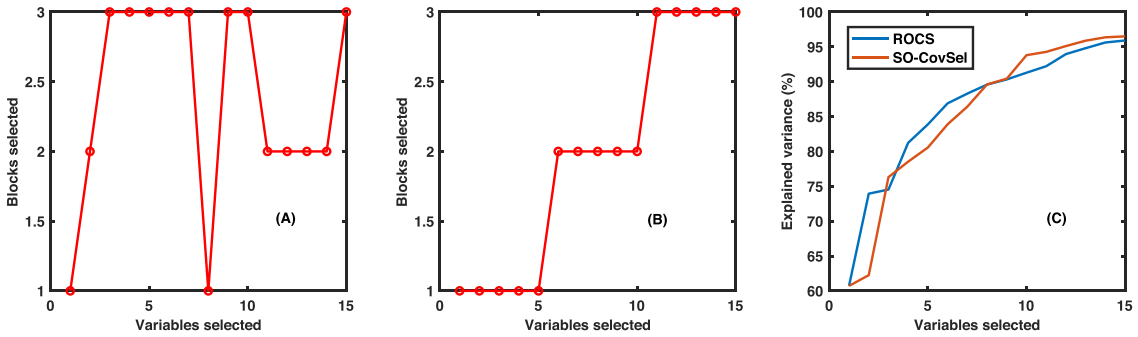


**FIGURE 8** Multiblock multiway feature selection with SKCovSel. The analysis is also known as response-oriented covariates selection (ROCS)[25] when block order is not defined and sequential orthogonalised covariates selection (SO-CovSel) when block order is defined.[19] (A) Block order of variable selection for ROCS, (B) block order of variable selection for So-CovSel and (C) explained variance in response

**TABLE 4** A summary of features selected for multiblock multiway data

| Method | First block | Second block | Third block |
|---|---|---|---|
| ROCS | 1296, 1350 | (1666,2), (1722,1), (1608,1), (1838,1), (1844,2) | 2110, 2298, 2180, 2128, 2244, 2196, 2290 |
| SO-CovSel | 1296, 1210, 1100, 1350, 1144 | (1648,1), (1728,1), (1852,1), (1550,1), (1694,1) | 2218, 2334, 2106, 2220, 2204 |

$(n-1)$D is to use jointly loading weights from multiple modes. For example, for selecting 3D features from 4D data, the user needs to perform the selection using the loading of any one mode, while for selecting a 2D feature from a 4D data, the user needs to perform the selection in loadings from two modes, and for selecting 1D features from 4D data, the user needs to perform the selection using the loading's of all the three modes.

## 5.3 | SKCovSel for multiblock multiway type data

In the presence of multiblock data sets, the method generates the solution obtained with ROCS or SO-CovSel, depending on if the block order is fixed. To show this, the three-block milk data set was processed with SKCovSel to extract the features without defining block order (Figure 8A). A total of 15 features were extracted (Table 4). Note that the second block of the milk data set is a multiway array; hence, the analysis presents multiblock multiway analysis. In the presence of no defined block order, the method selected two features from the first block, five features from the second block and then later eight features from the third block, that is, multiway array (Figure 8A). When the sequential block order is defined for SKCovSel (five features from each block sequentially), then the model performed a sequential variable selection as can be noted in the order of blocks selected in Figure 8B. Note that it is out of the scope of this work to find out which multiblock variable selection approach (SO-CovSel or ROCS) is better as that topic is already covered in an earlier article[25]; however, both the ROCS and SO-CovSel performed well in explaining the response by fusing information from different data blocks (Figure 8C).

The SKCovSel approach is a fast approach for selecting features for single block, multiblock and multiway data analysis. For example, time recording for executing the SKCovSel for selecting 30 features for single block, multiblock

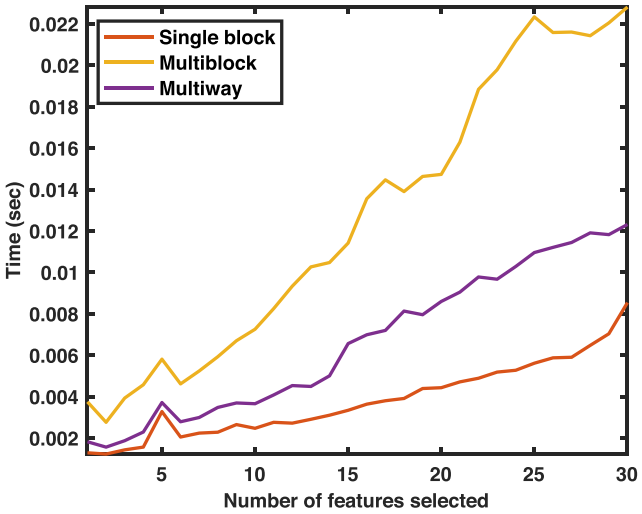


**FIGURE 9** Time requirements for single block, multiblock and multiway Swiss knife covariates selection (SKCovSel) analysis performed on the milk data set

and multiway milk data sets showed that it took less than 0.02 s for modelling all data scenarios. Note that the analysis presented in Figure 9 was using a computer with processor of 2.3-GHz 8-Core Intel Core i9 with 16-GB 2667-MHz DDR4 RAM. The time required for executing on a two way data block was the shortest due to the fact that there were fewer variables. The time execution for a multiway block was larger than the two way block due to having more variables. Finally, the time requirement for executing on a multiblock was the largest because the multiblock data set was both the two way and multiway in individual blocks.

## 6 | CONCLUSIONS

We developed and tested a new unified covariates selection algorithm called Swiss knife covariates selection (SKCovSel). With the test on wide data of types such as two-way, multiway and multiblock, the selection of the covariates was shown for both single and multiple responses. The SKCovSel technique is a single algorithm which covers all major types of covariates selection algorithms such as CovSel, fCovSel, ROCS, SO-CovSel and N-CovSel. Furthermore, the method supplies regression coefficient for the selected features such that selection and predictive modelling can be performed simultaneously. Just like the non-deflating fCovSel approach, the SKCovSel also does not require any deflation of the predictor matrices, hence, can be considered as a faster approach to perform all types of covariates selection analyses.

### PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/cem.3441.

### ORCID

*Puneet Mishra* https://orcid.org/0000-0001-8895-798X
*Kristian Hovde Liland* https://orcid.org/0000-0001-6468-9423
*Ulf Geir Indahl* https://orcid.org/0000-0002-3236-463X

### REFERENCES

1. Simon LL, Pataki H, Marosi G, et al. Assessment of recent process analytical technology (pat) trends: a multiauthor review. *Org Process Res Dev*. 2015;19(1):3-62.
2. Wang H-P, Chen P, Dai J-W, et al. Recent advances of chemometric calibration methods in modern spectroscopy: algorithms, strategy, and related issues. *TrAC Trends Anal Chem*. 2022;2022:116648.
3. Wold S, Albano CWJD, Dunn WJ, et al. Multivariate data analysis in chemistry. *Chemometrics*. Springer; 1984:17-95.

4. Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM J Sci Stat Comput*. 1984;5(3):735-743.
5. Prieto N, Pawluczyk O, Dugan MER, Aalhus JL. A review of the principles and applications of near-infrared spectroscopy to characterize meat, fat, and meat products. *Appl Spectrosc*. 2017;71(7):1403-1426.
6. Pasquini C. Near infrared spectroscopy: a mature analytical technique with new perspectives—a review. *Anal Chim Acta*. 2018;1026:8-36.
7. Bro R, Smilde AK. Principal component analysis. *Anal Methods*. 2014;6(9):2812-2831.
8. Wold S. *PLS Modeling With Latent Variables in Two or More Dimensions*: Verlag nicht ermittelbar; 1987.
9. Wold S, Sjöström M, Eriksson L. Pls-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst*. 2001;58(2):109-130.
10. Andersson M. A comparison of nine pls1 algorithms. *J Chemom: A J Chemom Soc*. 2009;23(10):518-529.
11. Indahl UG. The geometry of pls1 explained properly: 10 key notes on mathematical properties of and some alternative algorithmic approaches to pls1 modelling. *J Chemom*. 2014;28(3):168-180.
12. Mehmood T, Sæbø S, Liland KH. Comparison of variable selection methods in partial least squares regression. *J Chemom*. 2020;34(6):e3226.
13. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in partial least squares regression. *Chemom Intell Lab Syst*. 2012;118:62-69.
14. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (i pls): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl Spectrosc*. 2000;54(3):413-419.
15. Höskuldsson A. Variable and subset selection in pls regression. *Chemom Intell Lab Syst*. 2001;55(1-2):23-38.
16. Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. *Anal Chim Acta*. 2010;667(1-2):14-32.
17. Roger JM, Palagos B, Bertrand D, Fernandez-Ahumada E. Covsel: variable selection for highly multivariate and multi-response calibration: Application to ir spectroscopy. *Chemom Intell Lab Syst*. 2011;106(2):216-223.
18. Biancolillo A, Liland KH, Måge I, Næs T, Bro R. Variable selection in multi-block regression. *Chemom Intell Lab Syst*. 2016;156:89-101.
19. Biancolillo A, Marini F, Roger J-M. So-covsel: a novel method for variable selection in a multiblock framework. *J Chemom*. 2020;34(2):e3120.
20. Galindo-Prieto B, Trygg J, Geladi P. A new approach for variable influence on projection (vip) in o2pls models. *Chemom Intell Lab Syst*. 2017;160:110-124.
21. Wang ZX, He QP, Wang J. Comparison of variable selection methods for pls-based soft sensor modeling. *J Process Control*. 2015;26:56-72.
22. Van Loan CF, Golub G. Matrix computations (Johns Hopkins studies in mathematical sciences). Matrix Computations; 1996.
23. Mishra P, Roger J-M, Jouan-Rimbaud-Bouveresse D, Biancolillo A, Marini F, Nordon A, Rutledge DN. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends Anal Chem*. 2021;137:116206.
24. Andersson CA, Bro R. The n-way toolbox for matlab. *Chemom Intell Lab Syst*. 2000;52(1):1-4.
25. Mishra P, Metz M, Marini F, Biancolillo A, Rutledge DN. Response oriented covariates selection (rocs) for fast block order-and scale-independent variable selection in multi-block scenarios. *Chemom Intell Lab Syst*. 2022;224:104551.
26. Biancolillo A, Marini F, Roger J-M. N-covsel, a new strategy for feature selection in n-way data. In: 17th scandinavian symposium on chemometrics (ssc17); 2021.
27. Mishra P. A brief note on a new faster covariate's selection (fcovsel) algorithm. *J Chemom*. 2022;2022:e3397.
28. Liland KH, Næs T, Indahl UG. Rosa—a fast extension of partial least squares regression for multiblock data analysis. *J Chemom*. 2016;30(11):651-662.
29. Mishra P, Liland KH. 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. *Anal Chim Acta*. 2022;1206:339786.
30. Liland KH, Indahl UG, Skogholt J, Mishra P. The canonical partial least squares approach to analysing multiway datasets-cpls. *J Chemom*. 2022;36(7):e3432. doi:10.1002/cem.3432
31. Uusitalo S, Diaz-Olivares J, Sumen J, et al. Evaluation of mems nir spectrometers for on-farm analysis of raw milk composition. *Foods*. 2021;10(11):2686.
32. MATLAB version 9.10.0.1613233 (R2021a). The Mathworks, Inc., Natick, Massachusetts; 2021.
33. Osborne BG. Near-infrared spectroscopy in food analysis. Encyclopedia of analytical chemistry: applications, theory and instrumentation; 2006.

**How to cite this article:** Mishra P, Liland KH, Indahl UG. Swiss knife covariates selection: A unified algorithm for covariates selection in single block, multiblock, multiway, multiway multiblock cases including multiple responses. *Journal of Chemometrics*. 2022;e3441. doi:10.1002/cem.3441