Received: 18 August 2022 | Revised: 3 November 2022 | Accepted: 13 November 2022

DOI: 10.1002/cem.3458

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Iterative re-weighted covariates selection for robust feature selection modelling in the presence of outliers (irCovSel)

Puneet Mishra[1] | **Kristian Hovde Liland**[2]

[1]Food and Biobased Research, Wageningen University and Research, Wageningen, 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, Netherlands.
Email: puneet.mishra@wur.nl

Abstract

A new method for feature selective modelling in the presence of outliers is presented. The method is a combination of iterative re-weighted partial least squares and the covariates selection approach. The method relies on iterative down-weighting of the outlying samples prior to estimating the squared covariance for covariates selection. In this way, the outlying samples carrying low weights have minimal influence on the squared covariance estimation, while the inliers have the maximum influence on the squared covariance estimation. The method allows selecting robust features, and models based on such features in general perform better in terms of prediction accuracy (lower error) than selecting features using equal sample weights for all samples. The algorithm description and tests of the method in single and multiple response scenarios are presented. Method performance is also demonstrated on a real spectral data set.

KEYWORDS

feature selection, multivariate, robustness, spectroscopy

1 | INTRODUCTION

Multicollinear data are frequently encountered in different areas of science.[1] In chemistry, such data are widely generated by laboratory or handheld analytical instruments such as optical spectrometers. Often, the challenges with multicollinear data generated in analytical experiments are twofold; first, data are multicollinear, and predictive modelling based on ordinary least squares (OLS) can have redundant feature information, which is detrimental for model generalisation. Second, the usually lower number of samples available than the variables causes an ill-posed problem for matrix inversion during estimation of OLS models. The common approach to deal with both the challenges is to perform some form of dimensionality reduction using approaches such as variance or covariance maximisation. For example, popular chemometric techniques such as principal component analysis (PCA)[2] and partial least squares (PLS)[1,3] are used to reduce the dimensions of data and to acquire orthonormal score vectors carrying the most useful information. These score vectors are then used in an OLS to make predictive models. Note that for predictive cases, PLS is the most preferred approach as it is based on the maximisation of covariance between the predictor and the response.

Apart from multicollinearity and lower number of samples than variables, one other key challenge commonly encountered with data generated in the domain of analytical chemistry is outlying samples.[4,5] Outlying samples can occur due to instrumental errors or because of human errors in the sample handling and reference analysis. The

outliers can be both in the predictor (**X**) and response (**Y**) matrix. The outlying samples have a major influence in modelling approaches such as PLS. This is because the outlying samples by default contribute equally to the estimation of covariance ($\mathbf{X}^t\mathbf{DY}$), which is the first step of the PLS algorithm. The matrix **D** is a diagonal matrix carrying equal weight for each sample, i.e., $\mathbf{D} = \mathbf{I}$ ($d_{i=j} = 1, d_{i \neq j} = 0$). A solution to handling outliers is to either do a prior outlier detection and setting the sample weights (diagonal of **D**) to 0 and 1 for the outlying and in-lying samples, respectively. The other more practical approach is to integrate the outlying sample detection directly as an iterative step in the PLS modelling. In such an approach, the outlying samples are down-weighted while the in-lying samples retain higher weights to have the covariance estimation rely more on in-lying samples. This was also the foundation of the method called iterative re-weighted PLS (irPLS),[6] where the Y residuals were used in combination with a weighting function to update the sample weights during the estimation of the latent variables. The irPLS relied on the sample weight estimation solely on the residuals; hence, it was only capable of handling vertical outliers, that is, poorly predictive samples. To also be able to detect high leverage samples, that is, samples far from the model centre, the irPLS idea was extended through a method called partial robust-M (PRM) regression,[7] where the sample weights were defined as the product of the weights estimated using both the residuals and score vectors. Weight estimation using residuals allows down-weighting outliers in **Y**, while weight estimation using score vectors allows down-weighting high leverage samples, that is, outliers in **X**. Recently, the idea of irPLS and PRM was further extended to include handling of multiple responses under the method name RoBoost PLS[8] and PLS2.[9] The key idea behind the RoBoost PLS2[9] approach was to estimate sample weights for multiple responses as a product of weights estimated for individual responses similar to the idea of PRM[7] of combining weights from different sources as a product.

Although major attention has been paid to the development of robust PLS methods[4–9] and a wide range of feature selection methods are available,[10–13] there is currently limited literature on selection of features in the presence of outliers, that is, robust feature selection. Typically, such methods use some sort of subsampling to estimate feature stability from an ensemble of models,[14–17] that is, quite computationally intensive. In the chemometric domain, a key approach to selecting features is using the squared covariance between predictors and responses. For example, the method covariates selection (CovSel)[18] performs stepwise feature selection by repeated steps of squared covariance maximisation and Gram-Scmidt (GS) orthogonalisation,[19] similar to the NIPALS PLS modelling. The major difference between CovSel and NIPALS PLS is that in the case of CovSel the associated weight vector is chosen as a (sparse) standard basis vector in the direction of the feature of maximum squared covariance with the response(s). Subsequently, just like the NIPALS algorithm, the data matrices are deflated, and the process continues extracting the desired number of features. Note that the first step of CovSel, that is, estimation of covariance ($\mathbf{X}^t\mathbf{DY}$), is the same as the PLS. Hence, just like PLS, for CovSel the sample weights (**D**) are equal for all the samples ($d_{i=j} = 1$), indicating equal contribution of samples. Considering equal sample weights can be sub-optimal in the same way as it can be for PLS approaches as outlying samples can influence the estimation of covariance and the selection of features using the estimated squared covariance. Hence, a natural solution to allow a robust feature selection with CovSel is to allow a robust estimation of the squared covariance. The robust estimation of the squared covariance can be performed in the same way as proposed in the irPLS, PRM, and RoBoost PLS by using an iterative re-weighting approach. By doing such a re-weighting, it is expected that the outlying samples will have minimal influence on the squared covariance estimation, hence also minimal influence on the feature selection. The underlying assumption is that the outlying samples are deviations from the expected distribution of samples, that is, that they contain some kind of error that will be harmful to predictive modelling if included fully.

The aim of this study was to develop a new chemometric tool for feature selection in the presence of outliers. The method is a combination of irPLS and the CovSel approach. The method is termed iterative re-weighted covariates selection (irCovSel). The method relies on iterative down-weighting of outlying samples prior to estimating the squared covariance for covariates selection. The algorithm description and test of the method on simulated and real data are presented.

2 | MATERIAL AND METHOD

At first, the algorithm description is presented. Later, the analysis performed with the irCovSel is explained. In the following description of the algorithm, all matrices are denoted with bold uppercase typeface such as **X**. All vectors are denoted with bold lowercase typeface such as **w**. All scalars are denoted with italic typeface such as a. Note that irCovSel is a novel combination of irPLS and CovSel feature selection. The key algorithm behind both irPLS and CovSel

is the NIPALS algorithm for PLS. In NIPALS, the first step is the estimation of the covariance (or weight vector), which is used to estimate the scores, that is, the projection of data in the direction of maximum covariance. The scores are used to fit OLS on the response and later the already explained information is removed from both the **X** and **Y** matrices, and the steps continues in a loop until the desired number of latent variables are extracted. In the case of irPLS, the residuals from the OLS step in PLS are used for sample weighting using the bisquare function and later a new NIPLAS step is performed but using sample weights. irPLS also continues until the desired number of latent variables are extracted. CovSel is slightly different than irPLS and NIPALS, due to the fact that as a first step it estimates the squared covariance. The squared covariance allows estimating the associated weight vector as a (sparse) standard basis vector in the direction of the variable of maximum covariance with the response(s). Later, just like in the NIPALS PLS, the data matrices are deflated, and the process continues for extracting the desired number of variables according to minimisation of the (residual) covariance with the response(s). The irCovSel takes advantage of the irPLS reweighing strategy to estimate sample weights and the squared covariance for selection of the next feature. The algorithm of the irCovSel is further detailed in Section 2.1.

2.1 | Algorithm

Define $\mathbf{Y}$ $(n \times k)$ as the response matrix, $\mathbf{X}$ as the predictor data matrix, and let A be the desired number of features to be extracted. Let $\mathbf{D}$ be the initial sample weight matrix. Note that $\mathbf{D}$ is the diagonal matrix having $1/n$ as the weight for all samples. Both the predictors and the responses are assumed to be median centred (less influence of outliers than mean centred). Let α be the tuning parameter defining the aggressiveness in weighting down outliers, and let C be the number of responses. The iterative reweighing is performed in a continuous loop until the sum of absolute differences of weights of two consecutive iterations is smaller than a user defined limit. In case of a hypothetical case of non-convergence, a maximum number of iterations (J) can be set; however, our experience is that convergence is typically achieved in less than 10 iterations for practical cases.

2.2 | Comments on the algorithm

The proposed irCovSel approach is a direct combination of the two chemometric methods irPLS and CovSel. In particular, the key idea of sample weight estimation was used from irPLS and the step-wise selection of features was taken from the CovSel approach. In the presented algorithm, the sample weights are updated using the bisquare function $(1-u^2)^2$; however, the user is free to chose one of many weighting functions as described in the earlier work.[6] An illustration of the effect of the bisquare function given various values of α applied to residuals ranging from -3 to $+3$ is seen in Figure 1. In the presented algorithm, the **Y** residuals (single and multiresponse) were only used for sample reweighting. However, as discussed in the PRM and the RoBoost PLS methods, the scores and **X** residuals can also be used to update the sample weights. Furthermore, multiple criteria can also be used to compute global weights; for example, **Y** residuals can be used to down-weight vertical outliers, while the scores can be used to down-weight high leverage samples, that is, samples far from the model centre. One of the ideas behind using multiple criteria to update sample weights is to simply compute their product. However, in some recent works,[8,9] it is mentioned that the user can also try summation or averaging.

In the irPLS method, the **Y** residuals were directly used for weight estimation; hence, it was only capable of dealing with vertical outliers. This is also the reason behind the proposal of the PRM method where the product of the weights obtained with **Y** residuals and scores were used to deal with both the vertical outliers as well as high leverage samples. In the present study, we used adjusted residuals as estimated by Equation (1).

$$r_{adj} = \frac{r_i}{\sqrt{1-h_i}}, \tag{1}$$

where r_i is the OLS residuals and h_i is the least-squares fit leverage values (estimated by squared score values t_i^2). Leverages adjust the residuals by reducing the weight of high-leverage data points, which have a large effect on the least-squares fit. Adjusted residuals were later standardised as Equation (2).

Algorithm for Iterative Re-weighted Covariates Selection (irCovSel)

for $a = 1 : A$ - loop over A features to be selected

  $while\ crit > 10^{-5}$ & $iter \leq J$ - loop for sample re-weighting

    $\mathbf{V} = \sum_c (\mathbf{X}^t \mathbf{D} \mathbf{Y_c})^2$ - weighted sum of squared covariances

    $(m, s) = argmax(\mathbf{V})$ - maximum squared covariance value and its position

    $\mathbf{t} = \frac{\mathbf{X}_s}{\|\mathbf{X}_s\|}$ - estimate normalised score vector

    $\mathbf{Q}_t = \mathbf{Y}^t \mathbf{D} \mathbf{t}$ - temporary regression coefficients

    $\mathbf{R} = \mathbf{Y} - \mathbf{t}\mathbf{Q}_t^t$ - estimate residuals

      for $c = 1 : C$ - loop for multiple responses

        $\mathbf{R}_c = \mathbf{R}_c / \sqrt{1\text{-diag}(\mathbf{t}\mathbf{t}^t)}$ - compute adjusted residuals

        $\mathbf{R}_c = \frac{\mathbf{R}_c \times 0.6745}{\alpha \times \text{ MAD}(\mathbf{R}_c)}$ - standardise the adjusted residuals

        for $i = 1 : n$ - loop over samples

          if $|\mathbf{R}_c| > 1$ - limit large residuals

            $\mathbf{R}_c \Leftarrow 0$ (implemented as element-wise replacement)

          else

            $\mathbf{R}_c \Leftarrow (1 - \mathbf{R}_c^2)^2$ - bisquare function based weight estimation

          end

        end

      end

      $\mathbf{r} = \prod_{c=1}^{C} \mathbf{R}_c$ - product of weights for multiple responses

      $crit = \sum(|\mathbf{r}| - |\text{diag}(\mathbf{D})|)$ - updating criterion for loop

      $\mathbf{D} \Leftarrow \mathbf{I}_n \cdot \mathbf{r}$ - updating weights

  end

  $\mathbf{V} = \sum_c (\mathbf{X}^t \mathbf{D} \mathbf{Y_c})^2$ - weighted sum of squared covariances

  $(m, s) = argmax(\mathbf{V})$ - maximum squared covariance value and its position

  $\mathbf{T}_a = \frac{\mathbf{X}_s}{\|\mathbf{X}_s\|}$ - scores

  $\mathbf{W}_a = \{0\}, \mathbf{W}_{s,a} = 1$ - loading weights (zero vector with 1 at selected index)

  $\mathbf{P}_a = \mathbf{X}^t \mathbf{D} \mathbf{T}_a$ - X loadings

  $\mathbf{Q}_a = \mathbf{Y}^t \mathbf{D} \mathbf{T}_a$ - Y loadings

  $\mathbf{Y} \Leftarrow \mathbf{Y} - \mathbf{T}_a \mathbf{Q}_a^t$ - Y deflation

  $\mathbf{X} \Leftarrow \mathbf{X} - \mathbf{T}_a \mathbf{P}_a^t$ - X deflation

end

$\mathbf{R} = \mathbf{W}(\mathbf{P}^t \mathbf{W})^{-1}$ - projections for score prediction

$\mathbf{B} = \text{cumsum}(\mathbf{R}\mathbf{Q}^t)$ - regression coefficients

$\mathbf{B}_0 = \bar{\mathbf{Y}} - \bar{\mathbf{X}}\mathbf{B}$ - median compensation

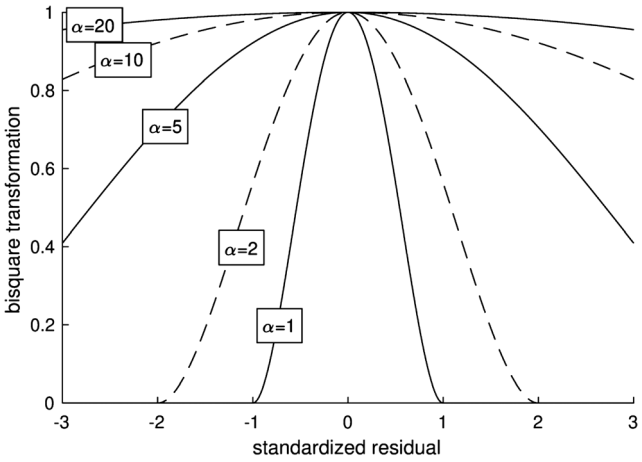

FIGURE 1 Effect of bisquare transformation on residuals ranging from −3 to +3, given chosen values of α from 1 (highly truncated weights) to 20 (almost flat weights)

$$u = \frac{r_{adj}}{\alpha s} = \frac{r_i}{\alpha s\sqrt{1-h_i}}, \tag{2}$$

where α is a tuning constant and s is an estimate of the standard deviation of the error term given by $s = \text{MAD}/0.6745$. MAD is the median absolute deviation of the adjusted residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution. The tuning constant α is a user defined input which defines the aggressiveness toward down-weighting outliers. Also note that when $\alpha \rightarrow$, then irCovSel becomes the standard CovSel as all samples will be given equal weights. Furthermore, as $\alpha \rightarrow 0$, the method will become highly aggressive and will end up down-weighting inliers as well. Hence, it is important to tune the α parameter, which can be performed using approaches such as cross-validation or using a separate validation set as performed in earlier studies.[9] The sample weights are updated as $u_i \Leftarrow (1-u_i^2)^2$ for $|u_i| = < 1$ and $|u_i| \Leftarrow 0$ for $|u_i| > 1$.

The algorithm allows robust feature selection for multiresponse scenarios as well. This was missing in the irPLS, PRM and RoBoost PLS method but was recently proposed in the RoBoost PLS2 method. The irCovSel uses a similar strategy as RoBoost PLS2, where the key idea behind estimating the sample weights in the multi-response scenario is to first estimate the sample weights for each response individually and later multiplying the weights for each response to have a single weight per sample. Estimating the weights individually for each response was noted as more robust when the responses have different variances.[9]

The irCovSel like the irPLS, PRM and RoBoost PLS approaches is a step-wise approach. This means that irCovSel selects one feature at a time, which gives it full freedom to have a wide range of extensions such as for selecting features in multiblock and multiway data. The extension will be similar to the already existing extensions of CovSel to multiblock[20,21] and multiway[22] scenarios, but the main difference will be that the CovSel step will be replaced by the irCovSel to deal with outlying samples. Furthermore, the irCovSel can also be adapted to local irCovSel by defining the weights using the dissimilarity of the samples just as usually performed in locally weighted PLS approaches.[23,24] The extension of methods are not discussed here as they involve different types of weight estimation, but they directly fit in the algorithm of irCovSel. irCovSel computes regression coefficients, **B**, and projections, **R**, for score predictions as a final part of the algorithm. This means that the user has direct access to predictions using anywhere from 1 to A selected variables without any remodelling of the selected features using OLS or PLS. The user can also easily predict scores of new samples, for example, for finding out which training samples these are most similar to or for assessing the degree of outlyingness based on T^2 (T-squared).

2.3 | Analysis

The capability of the irCovSel to select features in the presence of outliers will be demonstrated. For demonstration purposes, we used a potential outlier free data set and simulated the desired type of outlier in the data. The data contain

NIR spectra and reference measurements of total solids and fats performed on 296 milk samples and were already used for developing calibrations in an earlier study.[25] The spectral sensor used for the measurement was the NIRONE 1.4 (1,100 to 1,400 nm) from Spectral Engines (Helsinki, Finland). All measurements were performed in transmission mode. More information on the data set and reference total solids and fats analysis protocol can be obtained in the earlier study.[25] The data set can be considered as outlier free as potential outliers were removed manually before the data set was made public.[25] To confirm it, a PCA on the spectral data was performed, and histograms for responses are presented in Figure 2. In Figure 2A, some sample points on the right side appear strange; however, double checking the reference values for those samples, it was noted that those samples having lower values of reference measurements as well, hence, cannot be considered as outliers.

Using the milk data set, different outlier scenarios were simulated. In the first case of single response modelling, 11 fat content values for some samples were set to zero. Second, for the multiresponse case, a different set of 11 fat and total solids content values were set to zero. However, before any simulation, one out of three samples was selected as test set for model evaluation. From the remaining 2/3 part of the data, 48 samples were selected as validation set to tune the model hyperparameters (α) and total number of features to select. Finally, there were 149, 48 and 99 samples available in the calibration, validation and test set. Note that in the presented analysis, outliers were only simulated in the response matrix $\mathbf{Y}$; hence, the analysis only included using the residuals for weighting of the samples. However, if needed, one can also simulate outliers in the predictor matrix $\mathbf{X}$. Since the current algorithm uses adjusted residuals, it should be able to capture high leverage samples, but it will require adjustment if separate handling of the score vector and residuals of the predictor matrix is wanted for sample re-weighting. The only challenges then will be the optimisation of extra hyperparameters as any new criterion for weight estimation will carry a hyperparameter on its own, which may limit practicality of the method. Also, these days measurements with analytical sensors have become highly sophisticated, and there are minimal chances of having outliers in the spectral data matrix. All data analyses were carried out in MATLAB®.[26]

3 | RESULTS AND DISCUSSION

3.1 | irCovSel vs CovSel

The irCovSel method requires two main parameters to be optimised: the total number of optimal features to be selected and the α parameter for aggressiveness of down-weighting outliers. The validation analysis to select the optimal parameters for single- and multi-response scenarios involved exploring their combination in the interval of [1–20]. The root mean squared error of validation (RMSEV) for model tuning are shown in Figure 3. For the single response scenario (Figure 3A), the minimum RMSEV was found at 18 features and the α parameter as 9. For the multiresponse scenario (Figure 3B), the minimum RMSEV was again found at 17 features but with an α parameter at 11.

For the single response modelling case (Figure 4), the irCovSel model based on optimal parameters identified in Figure 3A, achieved an RMSEP = 0.30 %. The (unweighted) CovSel analysis with the same number of variables achieved a higher RMSEP = 0.72 %. Similarly, for the multiresponse modelling case (Figure 5), the RMSEP for the

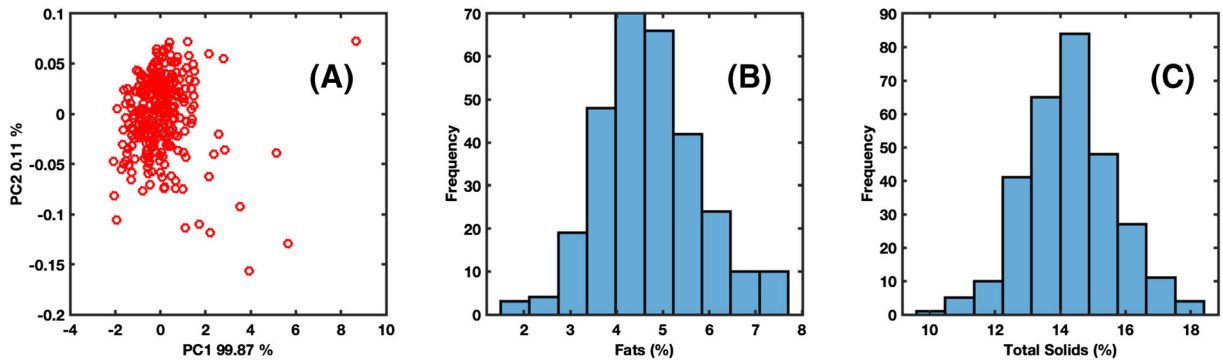

FIGURE 2 Plots related to data set. (A) PC2 vs PC1 for spectral data, (B) distribution of fat content, and (C) distribution of soluble solids content

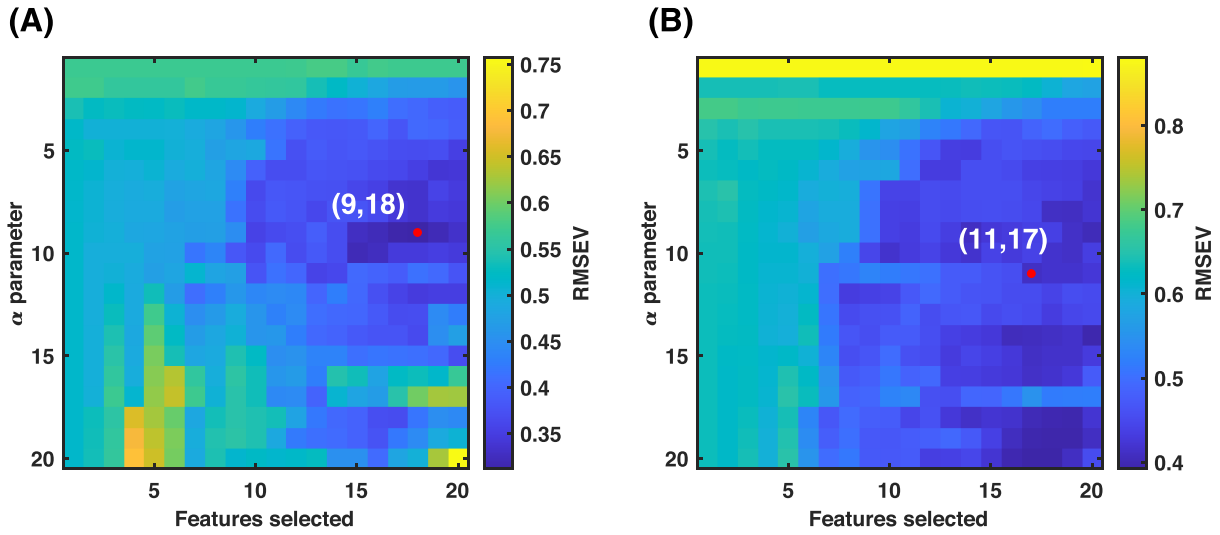

FIGURE 3 Alpha parameter versus number of features selected to select the optimal value attaining minimum root mean squared error of validation (RMSEV) for single response (A) and multiresponse (B) modelling. The parameters leading to minimum RMSEV are highlighted with white text and marked in red

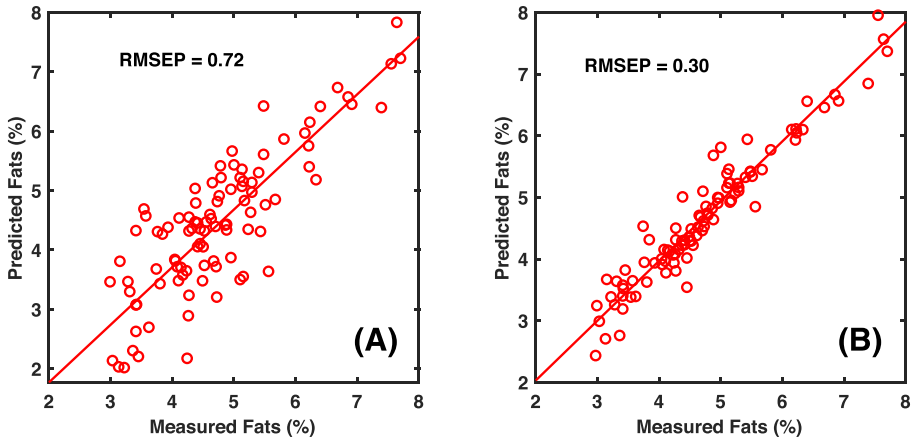

FIGURE 4 The performance of CovSel (A) and irCovSel (B) variable selection models for predicting fat content in milk

irCovSel based model was lower than the CovSel based model for both the responses. The improvements with the irCovSel model were larger for the total solids content as the RMSEP was almost 70 % lower than the RMSEP achieved with the CovSel selected variables. The improved predictive performance indicates that CovSel was heavily influenced by the presence of outliers in the data while irCovSel was less influenced by the presence of outliers and was able to select features carrying better predictive power than the CovSel selected features.

Using the α parameters at 9 and 11 for the single and multiresponse cases (Figure 3), the irCovSel models were trained and tested for the feature range of 1–30 (Figure 6). The analysis was carried out as a posterior analysis to understand the general trend of the predictive power of irCovSel and CovSel selected variables. As can be noted in Figure 6, the irCovSel selected variables in general led to lower RMSEP compared to the CovSel selected variables. This trend continued even after selecting more than the optimal number of features.

The main difference between CovSel and irCovSel is the change in the weights given to the samples during the estimation of the covariance according to $\mathbf{X}^t\mathbf{DY}$. In CovSel, all samples carried equal weights; hence, the outliers contributed equally to the estimation of covariance, which possibly hindered the selection of informative features. In the case of irCovSel, the weights given to the outlying samples were lowered during the iterative process (Figure 4), hence, allowed selecting informative features which in general led to lower RMSEP as shown in Figure 6. Note that features selected by CovSel and irCovSel were different. In Figure 7A-C, we see how the weights of the outlying samples were

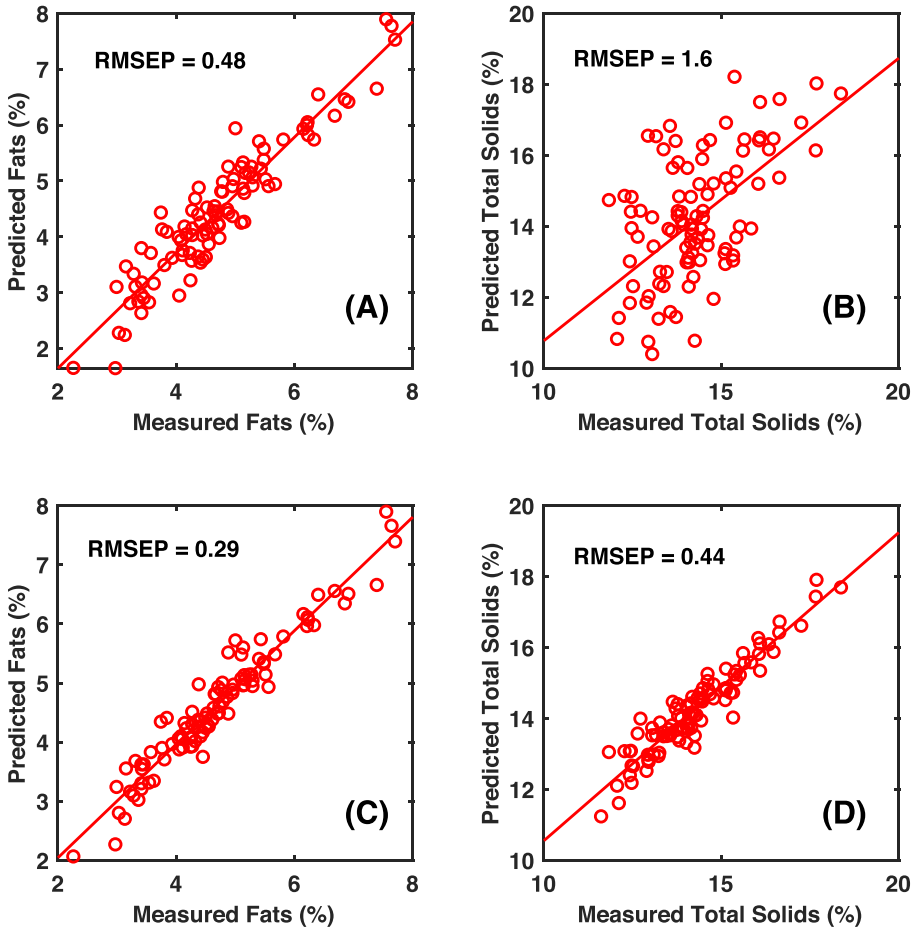

FIGURE 5 The performance of CovSel (top row) and irCovSel (bottom row) variable selection models for predicting fat (A, C) and total solids content in milk (B, D)

lowered compared to the inliers using irCovSel. In Figure 7, only sample weights corresponding to the first three extracted features are shown as the subsequent features followed the same trend of having lower weights for outlying samples and higher weights for inliers. The irCovSel method is a fast method for robust feature selection. A run on the milk data set to select up to 100 features only required less than 1 s on a computer with the following processor: 2.3 GHz 8-Core Intel Core i9 and memory: 16 GB 2667 MHz DDR4 RAM, as the iterative part of the procedure is light-weight.

3.2 | Effect of α parameter on outliers

The α parameter of the irCovSel method, or in general iterative re-weighted methods such as irPLS, PRM and RoBoost PLS, is the most important parameter that has a direct impact on sample weights. In the current version of irCovSel, the bisquare function was used, which is tuned by the α parameter. As highlighted also in earlier sections, for a bisquare weighting function with $\alpha \rightarrow$, the irCovSel becomes the standard CovSel as all samples will be given equal weight. As the $\alpha \rightarrow 0$, the method will become highly aggressive and will end up down-weighting inliers as well. This trend was also noted in the analysis of the milk data set. For example, the sample weights for the decreasing α parameter are shown in Figure 8. As can be noted in Figure 8A, with a relatively high $\alpha = 16$, the outlying samples were given weights in the range of 0.4–0.8. As the α parameter decreased to 4, all the outlying samples were given 0 weights, while at the same time, more inliers were given lower weights as well. Furthermore, when the $\alpha = 1$, then all the outlying samples were given 0 weights, while several of the inliers were also given 0 weights, indicating a highly aggressive down-weighting that is also detrimental for modelling. Hence, in the irCovSel method unlike CovSel, it is essential to tune the α parameter to achieve optimal performance.

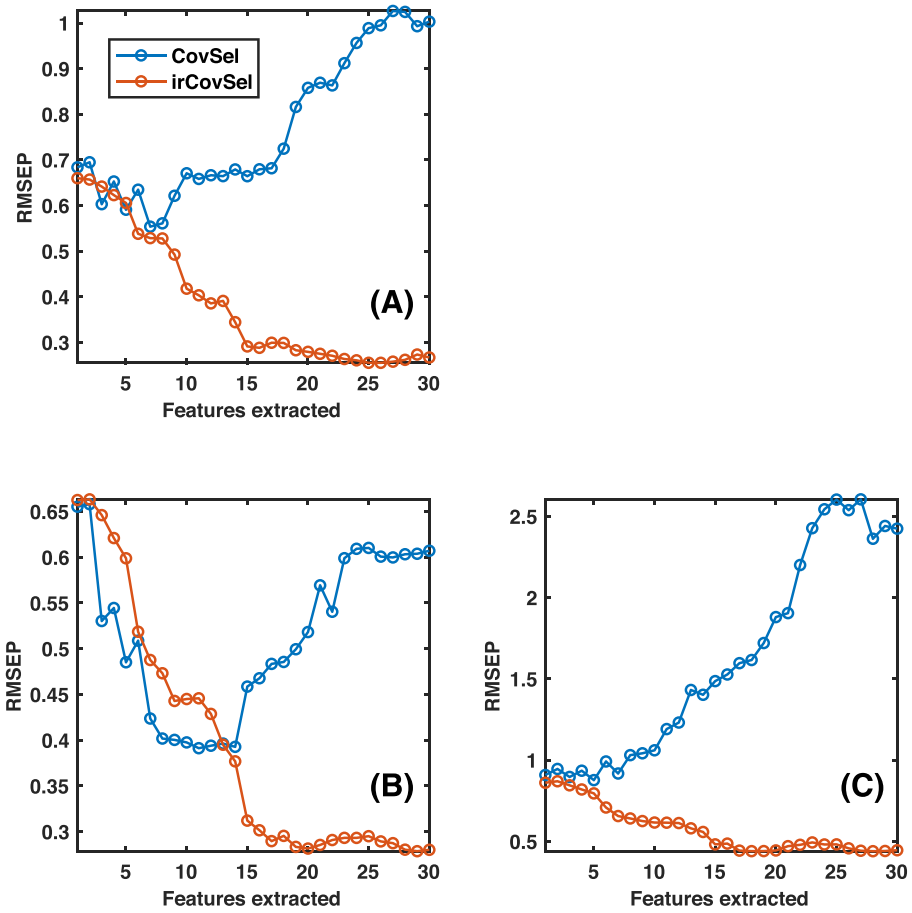

FIGURE 6 Exploring the behaviour of number of selected features on the RMSEP. (A) Single response case to predict fat in milk. Multiresponse case to predict fat (B) and total solids (C) in milk

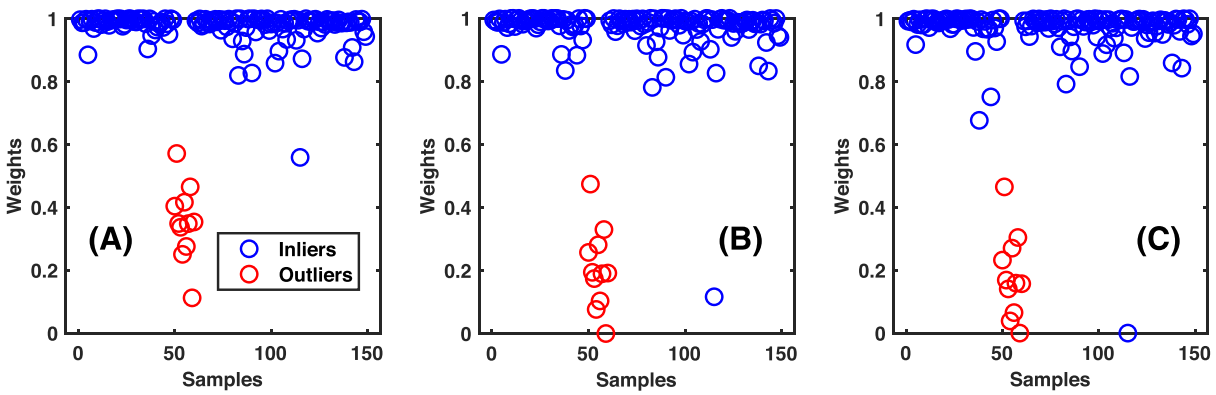

FIGURE 7 Sample weights obtained during the run of irCovSel for selecting the first three features in the milk data

3.3 | irPLS vs irCovSel

irCovSel can be considered as a special case of irPLS where the associated weight vector is chosen as a (sparse) standard basis vector in the direction of the variable of maximum weighted covariance (less affected by outliers) with the response(s). The key advantage of irCovSel over irPLS is to identify the key variables of interest without compromising too much in the predictive performance of the model. As a demonstration, the irPLS and irCovSel analysis was performed on the milk data set to predict fat content (Figure 9). The analysis was accompanied with the PLS and CovSel

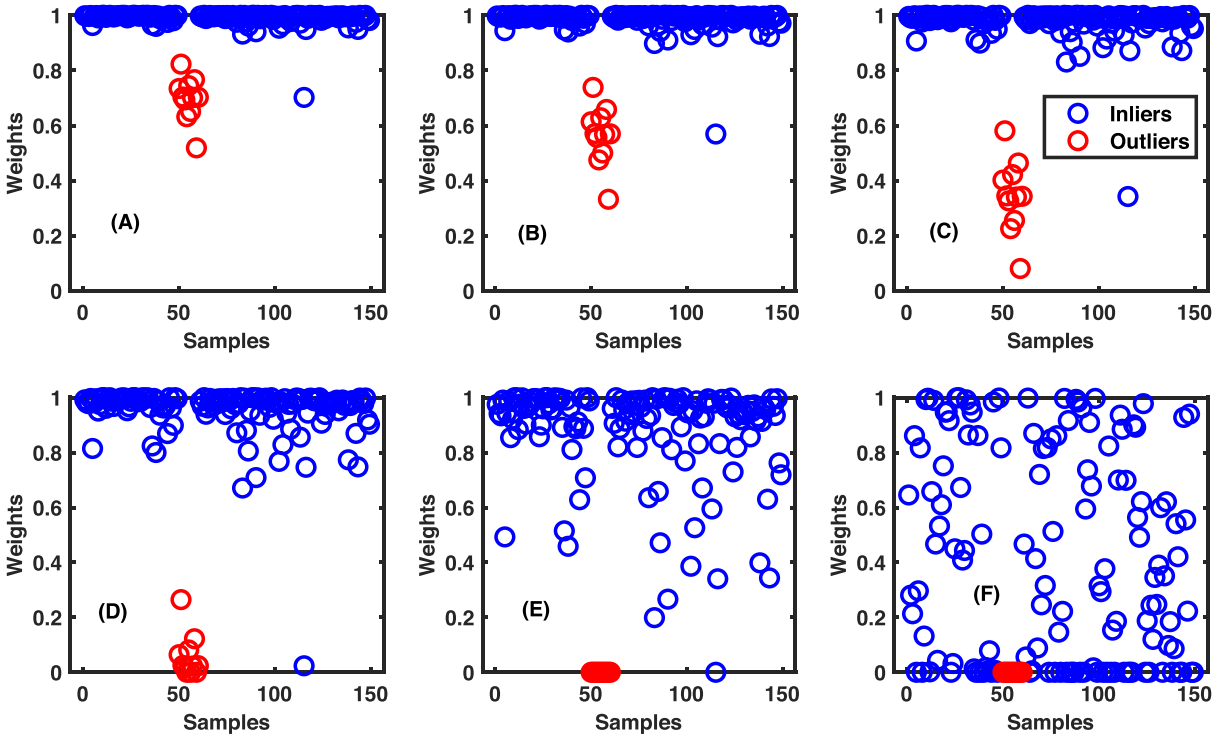

FIGURE 8 Effect of α parameter on the sample weights. (A) $\alpha = 16$, (B) $\alpha = 13$, (C) $\alpha = 10$, (D) $\alpha = 7$, (E) $\alpha = 4$ and (F) $\alpha = 1$

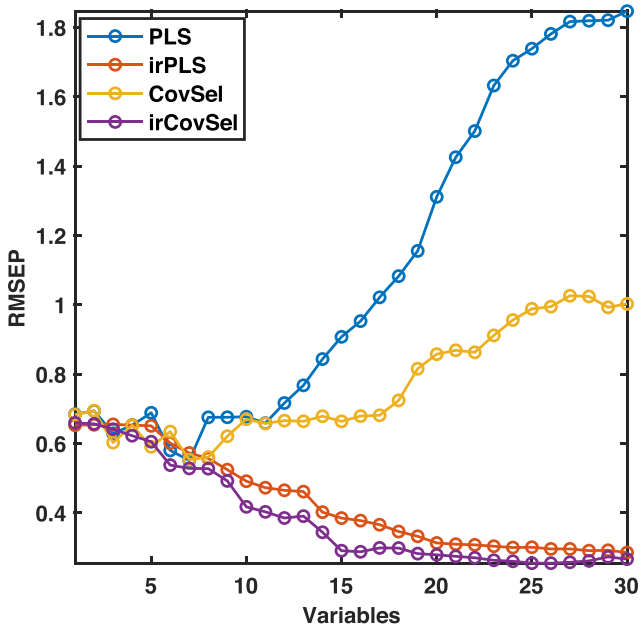

FIGURE 9 Performance of CovSel, PLS, irCOvSel and irPLS to predict fat content in milk including artificial outliers

analysis to also understand the predictive potential of irPLS compared to PLS and CovSel in the presence of outliers. In general, the predictive performance of irPLS was better than PLS and CovSel. The performance of irCovSel was better than irPLS. Note that the irCovSel model (10 features) was only based on 10 variables, while the irPLS model was based on 126 variables (10 latent variables). irCovSel reduced the total number of used variables by a factor of 13. The performance of PLS and CovSel was similar, but poorer than irCovSel.

3.4 | Test on a real data set

The method was also tested on a well know NIR spectroscopy data set related to the compositional analysis of biscuits.[27] The data set is particularly interesting as it has been used widely to test robust chemometric methods. The data set includes spectral (1,100–2,498 nm) and chemical compositional information on biscuit dough just before baking of the biscuits. The data set is multiresponse, and the four chemical constituents measured were fat, sucrose, dry flour and water. There are 40 samples in the calibration and 32 samples in the test set. Furthermore, sample number 23 is already know as the key outlying sample, and the authors have suggested to remove that sample prior to data modelling.[27] To have a comparison with already existing robust analysis of the biscuit data set, the data were preprocessed in the same way as in earlier studies.[5,28] Note that in earlier studies, the authors only modelled three responses and removed the fat content from the analysis as fat content was not correlated to the other response variables. In this study, we modelled all four responses. The results of the irCovSel analysis ($\alpha = 11$) for joint modelling of all four responses are shown in Figure 10. For comparison, the CovSel analysis is also presented in the same plot. The models were calibrated on the

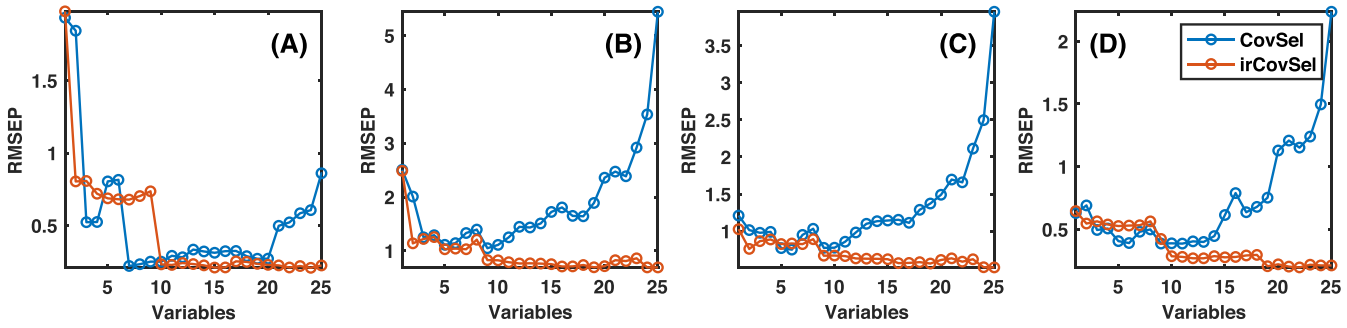

FIGURE 10 irCovSel analysis for joint prediction of four chemical constituents in biscuits. (A) Fat, (B) sucrose, (C) dry flour and (D) water

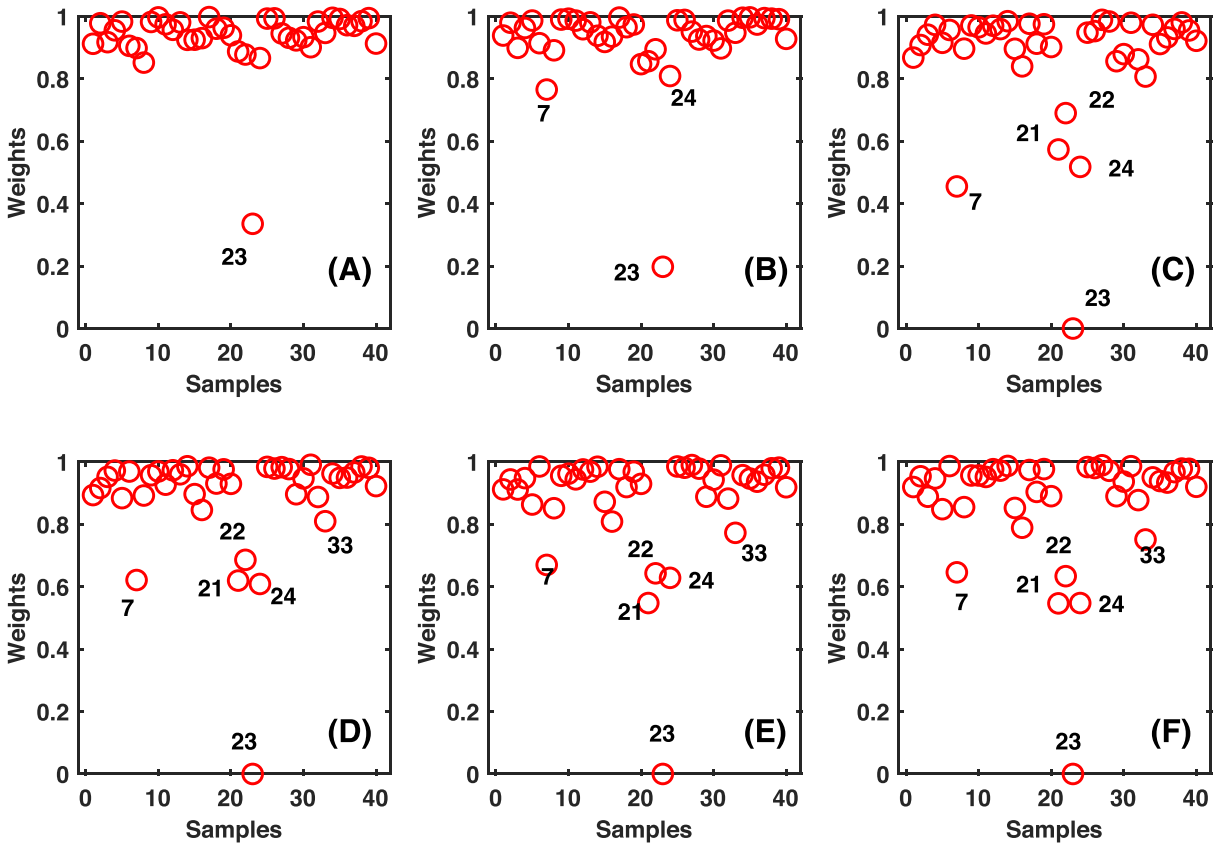

FIGURE 11 Sample weights for the first six (A–F) selected features. The outlying samples are highlighted with sample number

calibration set, and the results in Figure 10 are the predictions for the test set. In three out of four responses (sucrose, dry flour and water), the minimum RMSEPs were achieved with the irCovSel method, while for the fat content, the lowest RMSEP was the same for both CovSel and irCovSel, but irCovSel achieved it with a lower number of features. The key interesting aspect to note is that outlying samples 7, 21, 22, 23, 24 and 33 were given lower weight during the execution of the irCovSel (Figure 11). These are the same outlying samples, which have been identified in earlier studies using robust analysis.[5,28]

4 | CONCLUSIONS

An iterative re-weighted feature selection approach to select features in the presence of outlying observations was presented. The key strategy behind the proposal was the iterative re-weighting of samples, where the outlying samples were given lower weights, while the inliers were given higher weights. Since outlying samples have lower weights, they also have lower influence on the estimation of the squared weighted covariance. The squared weighted covariance was used to select the feature carrying maximum squared weighted covariance, and the data matrices (predictor and response) were orthogonalised with respect to that feature. The process was repeated until the desired number of features were selected, similar to the CovSel approach to variable selection. On comparison of the irCovSel and CovSel for selecting features in data containing outlying observations, it was noted that the features selected by irCovSel achieved in general better prediction (lower RMSEP) compared to the CovSel selected features. On comparison of the irCovSel with the irPLS approach, it was noted that irCovSel selected features that maintained predictive performance similar to irPLS, even though the total number of variables were reduced by up to a factor of 13. The irCovSel method is capable of handling multiple responses as this was achieved by estimating the weights for individual responses and then taking their product as global sample weights. The weighting strategy demonstrated in the current algorithm was based on the residuals, but the weighting strategy can also be modified and explored based on the need of the user. Note the irCovSel can also be easily extended to multiblock and multiway scenarios, just as currently available for CovSel. The multiblock and multiway extensions will require replacing the CovSel step with the irCovSel step and tuning of extra weighting parameters such as α either globally or individually for each data block. Extensions of irCovSel will be covered in our future works. Note that although in this study the algorithm was demonstrated also for multiple responses cases, user should take precautions when modelling multiple responses as sometimes the optimal number of features may not be the same for all the responses and selecting features individually for each response could be a better option. Similarly, if one of the responses has bad reference values, then the sample weights are down-weighted for all responses, as the current approach to combine weights from different responses is to take products of weights obtained through individual responses. Iterative weighting algorithms like irCovSel do not estimate sample weights for the test set. However, standard diagnostics such as T^2 (T-squared) can be calculated to estimate degree of outlyingess and thereby how much one can trust the predictions. Note that the final irCovSel models has the same number of features and model components, hence, the spectral residuals cannot be calculated to estimate Q statistics. Extending the iterative weighting to also encompass test set samples is a possible direction for future work.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/cem.3458.

DATA AVAILABILITY STATEMENT

Data are available on request from the authors.

ORCID

Puneet Mishra https://orcid.org/0000-0001-8895-798X
Kristian Hovde Liland https://orcid.org/0000-0001-6468-9423

REFERENCES

1. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst*. 2001;58(2):109-130.
2. Bro R, Smilde AK. Principal component analysis. *Anal Methods*. 2014;6(9):2812-2831.
3. 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.

4. Verboven S, Hubert M. Libra: a matlab library for robust analysis. *Chemom Intell Lab Syst*. 2005;75(2):127-136.
5. Hubert M, Branden KV. Robust methods for partial least squares regression. *J Chemom: A J Chemom Soc*. 2003;17(10):537-549.
6. Cummins DJ, Andrews CW. Iteratively reweighted partial least squares: a performance analysis by monte carlo simulation. *J Chemom*. 1995;9(6):489-507.
7. Serneels S, Croux C, Filzmoser P, Van Espen PJ. Partial robust m-regression. *Chemom Intell Lab Syst*. 2005;79(1-2):55-64.
8. Metz M, Abdelghafour F, Roger J-M, Lesnoff M. A novel robust PLS regression method inspired from boosting principles: Roboost-plsr. *Anal Chim Acta*. 2021;1179:338823.
9. Metz M, Ryckewaert M, Mas-Garcia S, et al. Roboost-PLS2-R: an extension of roboost-plsr method for multi-response. *Chemom Intell Lab Syst*. 2022;222:104498.
10. 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.
11. Mehmood T, Sæbø S, Liland KH. Comparison of variable selection methods in partial least squares regression. *J Chemom*. 2020;34(6):e3226.
12. Höskuldsson A. Variable and subset selection in PLS regression. *Chemom Intell Lab Syst*. 2001;55(1-2):23-38.
13. Xiaobo Z, Jiewen Z, Povey Malcolm JW, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. *Anal Chim Acta*. 2010;667(1-2):14-32.
14. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Daelemans W, Goethals B, Morik K, eds. *Machine learning and knowledge discovery in databases*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008:313-325.
15. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc: Ser B (Stat Methodol)*. 2010;72(4):417-473.
16. Wiegand P, Pell R, Comas E. Simultaneous variable selection and outlier detection using a robust genetic algorithm. *Chemom Intell Lab Syst*. 2009;98(2):108-114.
17. Jenul A, Schrunner S, Liland KH, Indahl UG, Futsæther CM, Tomic O. Rent-repeated elastic net technique for feature selection. *IEEE Access*. 2021;9:152333-152346.
18. 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.
19. Van Loan CF, Golub G. Matrix computations (Johns Hopkins studies in mathematical sciences). *Matrix Comput*. 1996.
20. 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.
21. 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.
22. 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.
23. Lesnoff M, Metz M, Roger J-M. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic nir data. *J Chemom*. 2020;34(5):e3209.
24. Shen G, Lesnoff M, Baeten V, et al. Local partial least squares based on global PLS scores. *J Chemom*. 2019;33(5):e3117.
25. 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.
26. The Mathworks, Inc. Natick, Massachusetts. MATLAB version 9.10.0.1613233 (R2021a); 2021.
27. Osborne BG, Fearn T, Miller AR, Douglas S. Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *J Sci Food Agric*. 1984;35(1):99-105.
28. Hubert M, Rousseeuw PJ, Van Aelst S. High-breakdown robust multivariate methods. *Stat Sci*. 2008;23(1):92-119.

How to cite this article: Mishra P, Liland KH. Iterative re-weighted covariates selection for robust feature selection modelling in the presence of outliers (irCovSel). *Journal of Chemometrics*. 2022;e3458. doi:10.1002/cem.3458