Received: 1 March 2023 | Revised: 30 March 2023 | Accepted: 31 March 2023

DOI: 10.1002/cem.3480

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

An algorithm for robust multiblock partial least squares predictive modelling

Puneet Mishra[1] | **Kristian Hovde Liland**[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 new algorithm for robust multiblock (data fusion) modelling in the presence of outlying observations is presented. The method is a combination of a robust modelling technique called *iterative reweighted partial least squares* and the block order and scale-independent component-wise multiblock partial least squares modelling. The method is based on automatic down-weighting of outlying observations such that their contribution is minimal during the estimation of block-wise partial least squares models, thus leading to robust modelling minimally affected by outliers. The algorithm and test of the methods for modelling multiblock data sets (simulated and real) in the presence of outlying observation are demonstrated.

KEYWORDS

data fusion, multiblock, multivariate, robustness, spectroscopy

1 | INTRODUCTION

Combining data from multiple sources is becoming increasingly popular in analytical sciences.[1] Often the aim is to combine different complementary sources of data to either have better insights into the data patterns or to improve the predictive performance of models.[2] In the domain of chemometrics, the most common methods to combine data from multiple sources are known as multiblock techniques.[3] Multiblock techniques are available for both exploratory analysis and predictive modelling.[4]

The key feature that makes multiblock techniques very useful and unique compared with traditional data fusion approaches such as neural networks based, is that multiblock techniques are highly parsimonious.[3,4] Model parsimony is highly important in several domains of sciences, such as analytical chemistry, where the interest is usually to understand the background chemical process and causes. The multiblock techniques allow such parsimony as they usually model latent spaces which capture the intrinsic background knowledge about the data being modelled. Furthermore, a key point to note is that most of the multiblock techniques are extensions of traditional latent space approaches such as principal component analysis, partial least squares and canonical correlation analysis.

Multiblock predictive modelling is a key topic of research in the chemometric domain where the aim is to combine data from different sources to improve the prediction of responses. For example, combining data from two different complementary analytical modalities,[5,6] or combining multiple forms of the same data.[7] There are several multiblock

methods available to perform predictive modelling. For example, *multiblock partial least squares* (MBPLS), which is based on the hierarchical PLS modelling concept of extracting global scores corresponding to shared latent spaces in multiblock data,[8] *sequential and orthogonalised PLS* (SO-PLS) to extract the complementary latent spaces in a sequential fashion,[9] *parallel and orthogonalised PLS* (PO-PLS) to extract the common and distinct latent spaces from multiblock data,[10] and *response oriented sequential alternation* (ROSA), relying on the extraction of the latent spaces from multiple blocks in a competition to minimise residuals.[11] Different methods carry their own advantages and disadvantages, for example, the MBPLS method is a fast block order independent method but is highly influenced by the scale of data, the SO-PLS is scale independent as it models each data block individually, SO-PLS is also block order dependent. The PO-PLS method allows modelling distinct and common information, whereas the ROSA method is both block order and scale independent, but prone to get stuck in local minima during the competition for selection of subspaces. More methods can also be found in a recent review.[4]

Although several methods are now available to perform multiblock predictive modelling, there is one area of research which is still unexplored for multiblock predictive modelling methods. The area is multiblock predictive modelling in the presence of outlying samples. Outlying samples are widely encountered in the area of spectral measurements, for example, due to improper measurements with sensors or sample handling, and due to instrumental or human errors during reference analysis. This means that samples can be outlying both in the predictor space and response space. Unlike multiblock modelling approaches, for single-block modelling, there is a wide range of methods available to handle outlying samples during the calibration process. The family of methods is commonly known as robust methods. Some methods are *robust SIMPLS*,[12] *iterative reweighted PLS*,[13] *partial robust M (PRM) regression*[14,15] and *RoBoost PLS/PLS2*.[16,17] The methods handle outliers with varying criteria ranging from an estimation of simple distances from the model centre to repeated weighting of the samples to adjust the covariance estimation process. However, most of the multiblock predictive methods, such as MBPLS, SO-PLS, PO-PLS and ROSA, are simple extensions of the PLS approach to latent space modelling. Hence, a robust multiblock method can be achieved as the extension of single-block robust PLS methods. As an example, we demonstrate the development of an iterative reweighted multiblock modelling method. However, the extension of any robust PLS method to its multiblock form should be feasible.

The aim of this study was to develop a new robust multiblock predictive method. The method is a combination of a robust modelling technique called iterative reweighted partial least-squares (irPLS) and the block order and scale-independent component-wise multiblock PLS modelling. The method is the first robust multiblock method to fuse data from multiple sources in the presence of outlying observations. The algorithm and test of the method on simulated and real data set are demonstrated.

2 | MATERIALS AND METHOD

At first, the algorithm is presented as is, followed by a discussion of details of the algorithm. Later the method is tested on several outlier scenarios on simulated data. Finally, the method is tested on real data. In the following description of the algorithm, all matrices are denoted with bold uppercase typeface such as $\mathbf{X}$. All vectors are denoted with bold lowercase typeface such as $\mathbf{w}$. All scalars are denoted with italic typeface such as a.

2.1 | Algorithm

Define $\mathbf{Y}$ $(n \times c)$ as the multi-response matrix, $\mathbf{X_1}, \mathbf{X_2}, .., \mathbf{X_B}$, as B predictor data matrices and let A be the desired number of components to be extracted. Let $\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_B$, be the initial sample weight matrices for each data block. 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. Let $\mathbf{o}$ $(1 \times A)$ be the order of the components to be extracted from different data blocks to explain $\mathbf{Y}$. The use of an order vector, $\mathbf{o}$, enables component extraction similar to both SO-PLS and ROSA.

Algorithm for robust multiblock partial least squares modelling

$for\ a = 1 : A$	- loop over A components to be extracted
$\mathbf{V} = \mathbf{X}_{\mathbf{o}_a}{}^t\mathbf{Y}$	- candidate loading weights
$\mathbf{Z} = \mathbf{X}_{\mathbf{o}_a}\mathbf{V}$	- candidate scores
$while\ crit > 10^{-5}$	- loop for sample re-weighting
$\mathbf{c} \Leftarrow canoncorr(\mathbf{Z}, \mathbf{Y}, \mathbf{D}_{\mathbf{o}_a})$	- weighted canonical correlation analysis (the vector of dominant left canonical weights)
$\mathbf{w} = \mathbf{V}\mathbf{c}$	- transform $\mathbf{V}$ by $\mathbf{c}$ to obtain the loading weights $\mathbf{w}$
$\mathbf{t} = \frac{\mathbf{X}_{\mathbf{o}_a}\mathbf{w}}{\lVert\mathbf{X}_{\mathbf{o}_a}\mathbf{w}\rVert}$	- estimate normalised score vector
$\mathbf{Q}_t = \mathbf{Y}^t\mathbf{D}_{\mathbf{o}_a}\mathbf{t}$	- temporary regression coefficients
$\mathbf{R} = \mathbf{Y} - \mathbf{t}\mathbf{Q}_t^t$	- estimate residuals
$for\ j = 1 : c$	- loop over multiple responses
$\mathbf{R}_j = \mathbf{R}_j / \sqrt{1\text{-diag}(\mathbf{t}\mathbf{t}^t)}$	- compute adjusted residuals
$\mathbf{R}_j = \frac{\mathbf{R}_j \times 0.6745}{\alpha \times \text{MAD}(\mathbf{R}_j)}$	- standardise the adjusted residuals with mean absolute deviation (MAD)
$for\ i = 1 : n$	- loop over samples
$if\ \lvert\mathbf{R}_{j,i}\rvert > 1$	- limit large residuals
$\mathbf{R}_{j,i} \Leftarrow 0$	(implemented as element-wise replacement)
$else$	
$\mathbf{R}_{j,i} \Leftarrow (1 - \mathbf{R}_{j,i}^2)^2$	- bisquare function based weight estimation
end	
end	
end	- end of response loop
$\mathbf{r} = \prod_{j=1}^{c} \mathbf{R}_j$	- product of weights for multiple responses
$crit = \sum(\lvert\mathbf{r}\rvert - \lvert\text{diag}(\mathbf{D}_{\mathbf{o}_a})\rvert)$	- update criterion for loop
$\mathbf{D}_{\mathbf{o}_a} \Leftarrow \mathbf{I}_n \odot \mathbf{r}$	- update weights matrix with $\mathbf{r}$ on the diagonal, otherwise zeros
end	- end of re-weighting loop
$\mathbf{c} \Leftarrow canoncorr(\mathbf{Z}, \mathbf{Y}, \mathbf{D}_{\mathbf{o}_a})$	- weighted canonical correlation analysis (the vector of dominant left canonical weights)
$\mathbf{w}_{\mathbf{o}_a} = \mathbf{V}\mathbf{c}$	- transform $\mathbf{V}$ by $\mathbf{c}$ to obtain the final loading weights $\mathbf{w}_{\mathbf{o}_a}$
$\mathbf{t}_a = \frac{\mathbf{X}_{\mathbf{o}_a}\mathbf{w}_{\mathbf{o}_a}}{\lVert\mathbf{X}_{\mathbf{o}_a}\mathbf{w}_{\mathbf{o}_a}\rVert}$	- estimate final normalised score vector
$\mathbf{q}_{\mathbf{o}_a} = \mathbf{Y}^t\mathbf{D}_{\mathbf{o}_a}\mathbf{t}_a$	- Y loadings
$\mathbf{Y} \Leftarrow \mathbf{Y} - \mathbf{t}_a\mathbf{q}_a^t$	- robust Y deflation
$for\ i = 1 : B$	- loop over B blocks to extracted block specific loadings
$\mathbf{p}_{i_a} = \mathbf{X}_{i_a}^t\mathbf{D}_{\mathbf{o}_a}\mathbf{t}_a$	- Block specific X loadings
$\mathbf{X}_{i_a} \Leftarrow \mathbf{X}_{i_a} - \mathbf{t}_a\mathbf{p}_{i_a}^t$	- Block specific robust deflation using extracted score
end	- accumulate loading weights, scores and loadings in matrices (not shown)
end	- end of component loop
$\mathbf{R} = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1}$	- 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

*Calculation of projection for score predictions **(R)** assumes block-specific loadings and loading weights stacked with matrices of zeros for blocks not used for components extraction.

2.2 | Comments on the algorithm

The proposed algorithm is a combination of the irPLS and a block-wise components extraction strategy for multiblock data modelling popular in the chemometric domain. In particular, the sample weights estimation was based on irPLS[13] and the component-wise multiblock modelling was similar to the SO-PLS[9] and the ROSA[11] multiblock modelling. In the current implementation, the sample weights are updated using a bisquare function, but other weighting functions can also be explored.[13] Currently, for sample weighting, the **Y** residuals were used; however, the scores and **X** residuals can also be combined[17] to update sample weights by taking products of sample weights obtained with different criteria.

In ordinary PLS modelling, data (both **X** and **Y**) are usually mean-centred. However, the mean can be a non-robust measure to estimate the model centre as the outlying observations will have a major effect on mean estimation. Hence, median-centring was used as it is less affected in the presence of outlying observations. A recently proposed algorithm for robust feature extraction has also used the median-centring[18] and has already reported its benefit for handling data containing outlying observations.

In the first proposed irPLS method, the **Y** residuals were directly used for weight estimation and it was only capable of dealing with outliers in the vertical direction (vertical outliers). Later, the PRM method, estimated sample weights jointly with **Y** residuals and scores 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) and earlier used in Mishra and Liland.[18]

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

where r_i are the OLS residuals, and h_i are the least-squares fit leverage values. Leverages adjust the residuals by reducing the weight of high-leverage data points that have a large effect on the least-squares fit. Adjusted residuals were later standardised as in Equation (2).

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

where α is a tuning constant, and s is an estimate of the standard deviation of the error term given by $s = \mathrm{MAD}/0.6745$. 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 defines the aggressiveness towards down-weighting outliers. For example, the $\alpha \to \infty$, then all samples will be given equal weights and the algorithm will converge to SO-PLS or ROSA depending on the component extraction order. As the $\alpha \to 0$, the method will become highly aggressive and down-weighting inliers. It is important to tune α using validation approaches like earlier studies.[17,18]

The algorithm handles multi-response scenarios as well. It uses a similar strategy as RoBoost PLS2[17] and irCovSel,[18] where estimating the samples weights in the multi-response scenario involves first estimating 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.[17]

The algorithm can deal with responses of different scales and variances because it uses the canonical PLS (CPLS)[19] approach to loading weight estimation. In usual PLS2 modelling, the scale and variance of different responses can have a major influence on the estimation of the loading weights; however, in CPLS, the loading weights are estimated using a canonical correlation analysis step. The canonical analysis step operates independently of the scales and variances of different responses, hence making the CPLS also scale and variance independent of the multiple responses to be modelled. Due to the use of CPLS, the method is also suitable for other multi-response tasks such as multi-class classification or using additional information about responses to enhance the component extraction.[19]

The block component extraction order in the algorithm is provided as a user input. Hence, it gives the freedom to the algorithm to not only do sequential multiblock modelling but allows to extract components in any possible order. If the user defines a sequential order then the algorithm will lead to a robust sequential PLS solution. Such flexibility in component extraction has several benefits, for example, when the user is interested in finding the best possible order of model components then the user can freely explore all possible block orders (including sequential) typically by using a validation set to find the best order. The other benefit could be when the user has some insight into the importance of blocks, then the user can restrict the total number of components to extract from each block and in a predefined order.

In general, leaving the user to define the block order adds freedom to the algorithm which is not possible with the traditional SO-PLS[9] type heuristic to model components.

In recent chemometric literature, some iterative robust modelling approaches such as RoBoost -PLS and -PLS2,[16,17] have proposed estimating sample weights using wide information criteria such as Y-residuals, X-residual and scores. The main benefit of using different information criteria is that a wide type of outliers can be covered; however, at the same time, the number of parameters to be optimised increases proportionally. For example, using the three different criteria such as Y-residuals, X-residual and scores indicates an optimisation of four parameters, three for the criteria and one for the total number of latent variables. To avoid the complexity of optimising several parameters, this algorithm only implements the Y-residuals (adjusted by least-squares leverages) to update sample weights. However, the user is free to change the criterion and even add multiple criteria for estimating sample weights.

2.3 | Data sets for method demonstration

2.3.1 | Milk data set

The milk data set contains spectral data and reference fats and total solids content for 296 milk samples.[20] The spectral data were of multiblock form as the measurements on the same samples were performed with three different complementary near-infrared (NIR) spectral sensors: NIRONE 1.4, NIRONE 2.0 and NIRONE 2.5 from Spectral Engines (Helsinki, Finland). The NIRONE 1.4 was in 1100–1400 nm, NIRONE 2.0 was in 1550–1950 nm and NIRONE 2.5 was in 2000–2450 nm spectral ranges. Measurements were performed in transmission mode.

2.3.2 | Biscuit data set

Biscuit data set[21] has NIR spectral and reference biscuit constituents information related to fat, sucrose, dry flour and water. The measurements were performed on biscuit dough pieces. There were a total of 72 spectral measurement and reference constituents. The spectral data consist of 700 points measured from 1100 to 2498 nm in steps of 2 nm. Note that the biscuit data set has been widely used in literature to test robust chemometric methods. In this study, it has been used to directly compare the results with the results of earlier developed methods.

2.3.3 | Soil data set

Soil data set[22] consisted of 102 X-ray fluorescence (XRF) and laser-induced breakdown spectroscopy (LIBS) measurements on loose soil samples (dry with grain size <2 mm). For XRF data acquisition, the X-ray tube was set to voltage and current of 35 kV and 7 μA, respectively. No vacuum condition or filters were used for the XRF spectra acquisition. For LIBS data acquisition, the instrument used laser pulses with 65 mJ, 19.5 cm of lens-to-sample distance (given that 255 J cm^{-2} laser fluence), 15 accumulated laser pulses, 2 μs of delay time, and 7 μs of integration time gate. XRF had 2048 variables in the energy range of 0.01 to 40.74 keV. LIBS data had 53,717 variables in wavelengths ranging from 200.01 to 779.99 nm. The reference properties were exchangeable calcium (ex-Ca) and magnesium (ex-Mg).

3 | RESULTS

3.1 | Single-response analysis—Simulated outliers

The milk spectra from three different NIR sensors are shown in Figure 1. In all three different spectral ranges, the spectra have different peaks for overtones of OH and CH bonds which are in abundance in macromolecules such as fat and water in milk. To demonstrate the first case of robust multiblock single-response modelling, some simulations were performed based on the milk data set. At first, the data were divided into calibration and test sets by selecting every 3rd sample as the test set and the remaining as the calibration set. This led to a total of 197 samples in the calibration set and 99 samples in the test set. Later, in the calibration set, the fat content values for samples 20–50 were set to zero as

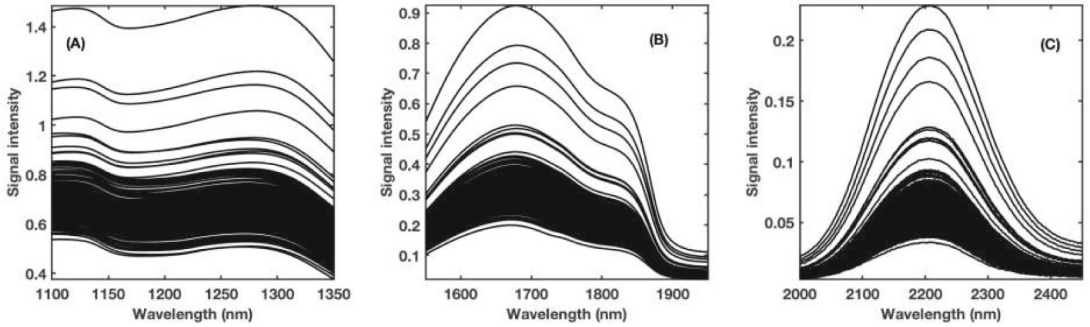

FIGURE 1 Three-block spectral data from milk data set. (A). NIRONE 1.4 was in 1100–1400 nm, (B) NIRONE 2.0, and (C) NIRONE 2.5.

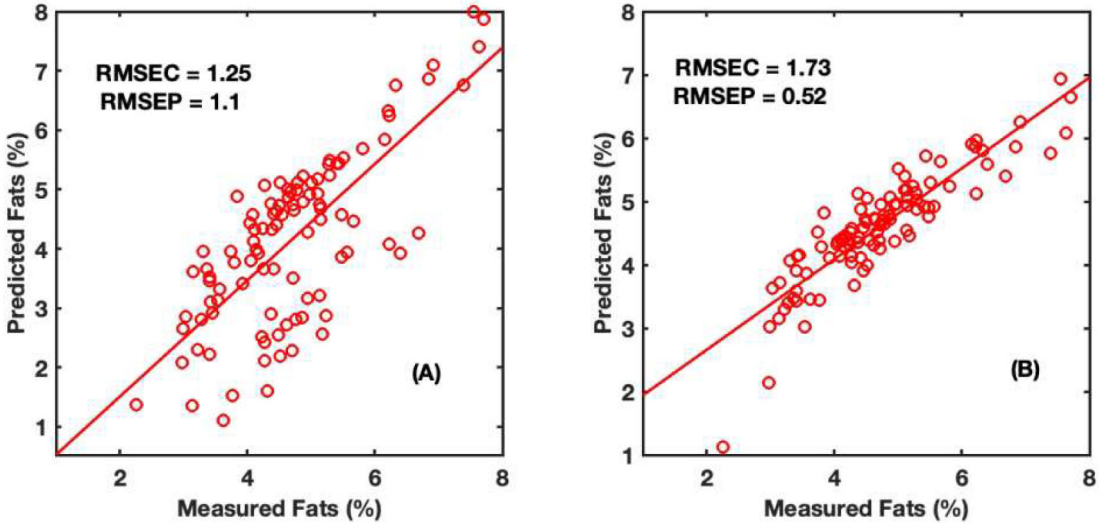

FIGURE 2 Unweighted (A) and iterative weighted (B) multiblock modelling performed on three-block milk data to predict fat content.

potential outliers. Unweighted and iterative weighted multiblock models were developed using three-block spectral data to predict the fat content. In total 12 components were extracted following the block order for component extraction as [1 2 3 1 2 3 1 2 3 1 2 3]. Note that the same component order was used for both the unweighted and iterative weighted modelling.

Models were later tested on the remaining test set and the results are shown in Figure 2. The root mean squared error of prediction (RMSEP) for the iterative weighted model ($\alpha = 7$) was 50% lower compared with unweighted modelling. The root mean squared error of calibration (RMSEC) was higher for the iterative weighted model compared with the unweighted model. This was due to the fact that the iterative weighted model fits poorly to the outlying samples in the calibration set, whereas the unweighted model gives equal weights to outlying samples and tries to fit equally well even to the outlying samples.

Figure 3 shows the sample weights for the first three model components extracted from blocks 1, 2 and 3, for iterative weighted modelling. As can be noted, the iterative modelling assigned nearly zero weights to the simulated outlying samples 20–50 in the calibration set. The capability of the iterative weighted approach to down-weight outlying observation gives it the power to achieve robust models.

A key parameter in the iterative weighted modelling is the α parameter. The parameter defines the aggressiveness for sample down-weighting. For the case of the milk data set, it can be noted that $\alpha < 15$ always led to lower RMSEP compared with unweighted modelling (Figure 4). After $\alpha > 15$, the performance of the weighted and unweighted gets similar. This is mainly due to the fact that the effect of sample weighting reduces with increasing α, with equal sample weighting when $\alpha \to \infty$.

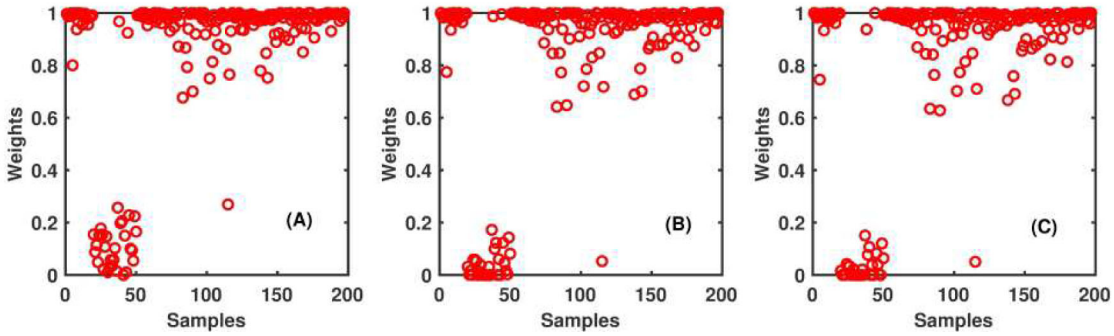

FIGURE 3 Sample weights for first three model components after iterative weighted multiblock modelling to predict fat content in milk.

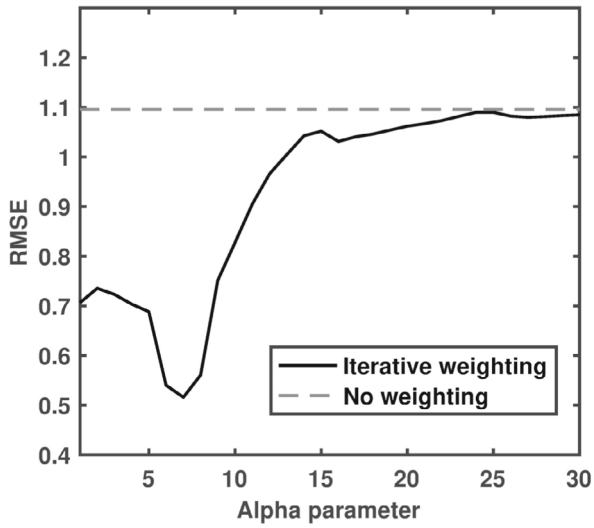

FIGURE 4 A posterior prediction analysis on test set to demonstrate the effect of the α (alpha) parameter on the model performance.

3.2 | Multi-response analysis—Simulated outliers

One of the unique features of the proposed robust multiblock method is that it is capable of handling multiple responses as well, thanks to the CPLS strategy integrated into the algorithm. To demonstrate this, outliers were simulated in two responses (fats and solids) of the milk data set. In the calibration set, samples 20–35 were set to zero for fat content, whereas samples 36–50 were set to zero for solids. Note that no common samples were set to zero, hence, the outliers were unique to each response. The results of unweighted and iterative weighted analysis are shown in Figure 5. The performance of the iterative weighted model ($\alpha = 9$) was better for predicting both the responses compared with unweighted models. Note that an equal number of components were extracted in both cases and following the block order [1 2 3 1 2 3 1 2 3 1 2 3]. The RMSEP for predicting fats was 65% lower with iterative weighted compared with unweighted modelling. The RMSEP for predicting solids was 80% lower with iterative weighted compared with unweighted modelling. The RMSEC for the iterative weighted models were higher compared with the unweighted model. This is mainly due to the fact that iterative weighted models fit poorly for the outlying samples, whereas the unweighted models fit equally to all samples.

Figure 6 shows the sample weights for the first three model components extracted from blocks 1, 2 and 3, for iterative weighted modelling. As can be noted, the iterative modelling assigned lower and nearly zero weights to the simulated outlying samples 20–50 in the calibration set.

The effect of α on the predictive performance of the multi-response model is shown in (Figure 7). It can be noted that $\alpha < 15$ always led to lower RMSEP compared with unweighted modelling (Figure 4). After $\alpha > 15$, the performance

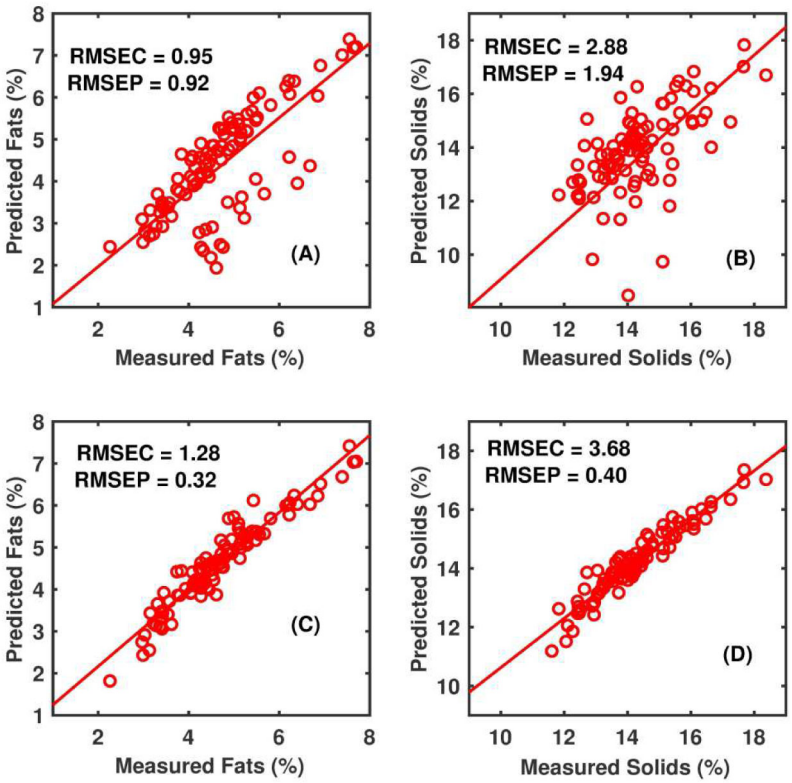

FIGURE 5 Unweighted and iterative weighted multiblock modelling performed on three-block milk data to predict fat and total solids content. Unweighted modelling to predict (A) fat, and (B) solids content. Iterative weighted modelling to predict (C) fat, and (D) solids content.

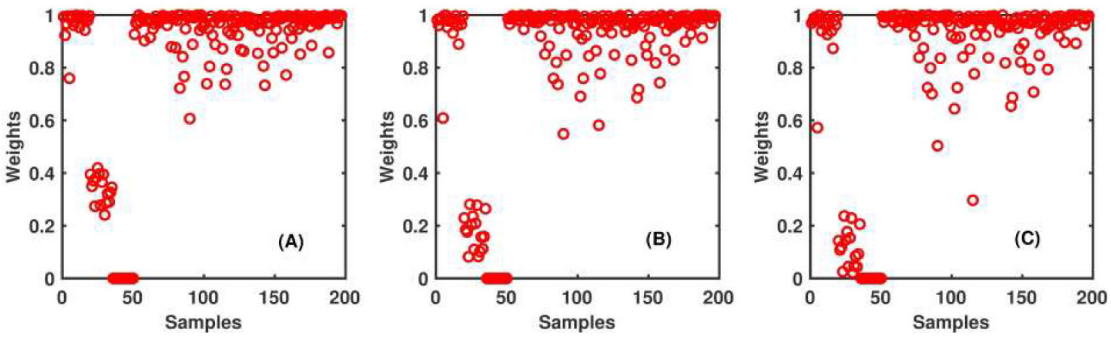

FIGURE 6 Sample weights for the first three model components after iterative weighted multiblock modelling to jointly predict fat and solids content in milk. Samples weights (A) first latent variable, (B) second latent variable, and (C) third latent variable.

of the weighted and unweighted gets more similar. This is mainly due to the fact that the effect of sample weighting reduces with increasing α, with equal sample weighting when $\alpha = 1$.

3.3 | Analysis with real outliers—Biscuit data

The iterative reweighted modelling was also demonstrated on real data sets. The first real data set was the biscuit data set carrying four different responses to be predicted with NIR data. The biscuit data set is the most commonly used spectral data to demonstrate the performance of robust modelling. However, note that the biscuit data set is not originally multiblock data. In this study, we divided the spectral range of the NIR data into two parts to create multiblock

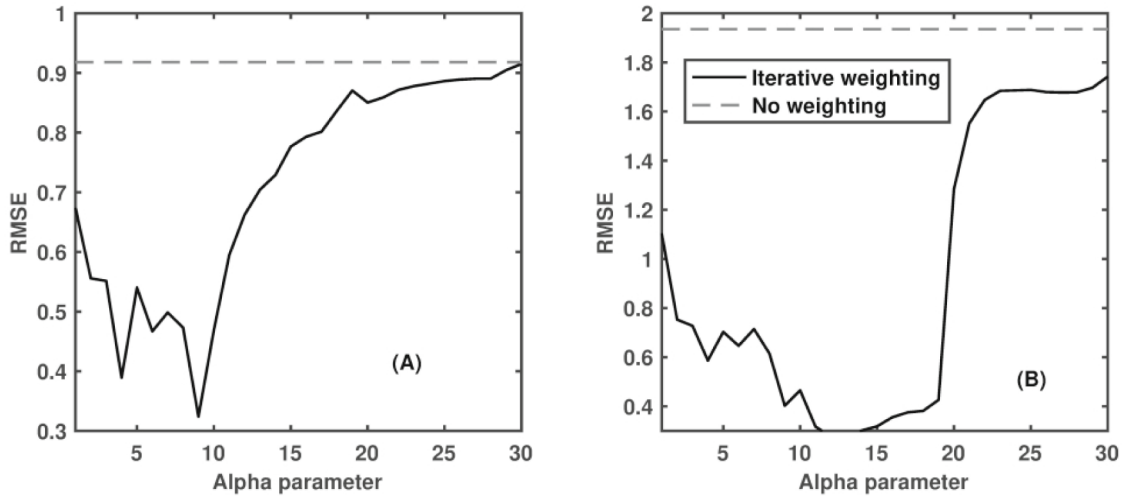

FIGURE 7 A posterior prediction analysis on test set to demonstrate the effect of the α (alpha) parameter on the model performance. (A) Fats and (B) solids content.

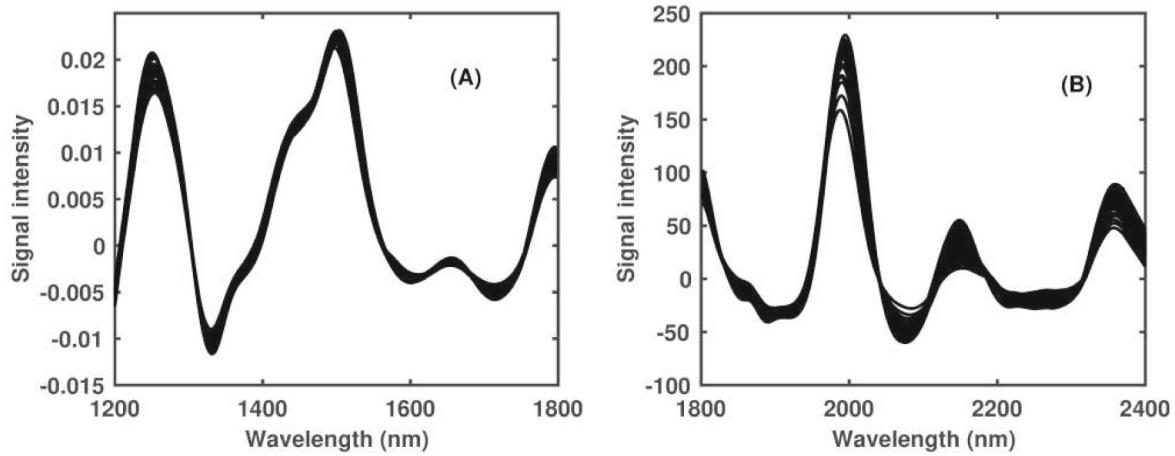

FIGURE 8 Simulated multiblock biscuit spectra data carrying different scales to demonstrate scale-independent robust multiblock modelling. (A) Spectra data in 1st and 2nd overtones range and (B) combination bonds vibrations range.

data. The spectra range of 1200–2400 nm was divided into 1200–1800 nm, covering the 1st and 2nd overtones and the range of 1801–2400 nm covering the combination bonds vibrations. Further, to demonstrate the scale independence of the proposed multiblock method, the scale of one of the blocks (1801–2400 nm) was made 1000 times higher than the other block (Figure 8). The biscuit data now have two blocks and four responses to be predicted. Note that the biscuit data set was already partitioned from the source into calibration (40 samples) and test set (32 samples). In the literature, earlier works have also used similar partitions.

The results of the unweighted and iterative weighted models ($\alpha = 15$) predicting fat, sucrose, dry flour, and water in biscuits are shown in Figure 9. It can be noted that for all the responses, the RMSEP was lower for the iterative weighted models (bottom row in Figure 9) compared with the unweighted models (top row in Figure 9). Note also that the component extraction order was the same for both models [1 2 1 2 1 2 1 2 1 2]. The RMSEC in all the cases was higher for the iterative weighted models but that was due to the fact that iterative models fit poorly to the outlying training samples, whereas the unweighted models fit equally for outlying samples and inlying samples.

The sample weights for the first five components extracted are shown in Figure 10. As can be noted, the key outlying samples (7, 21, 22, 23, 24, 33) were down-weighted automatically by the iterative weighting approach. Note also that outliers detected were the same outliers as has been detected in earlier studies using the same data.[12,18,23] The capability

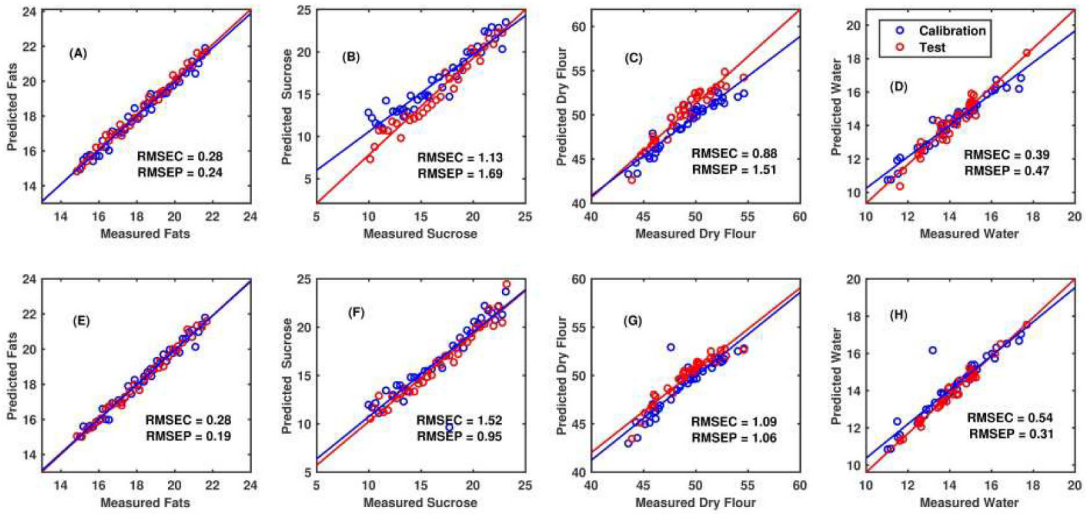

FIGURE 9 Prediction plots for unweighted (top row) and iterative weighted multiblock modelling (bottom row) performed on the biscuit data set. Unweighted multiblock modelling to predict: (A) fat, (B) sucrose, (C) dry flour and (D) water. Iterative weighted multiblock modelling to predict: (E) fat, (F) sucrose, (G) dry flour and (H) water.

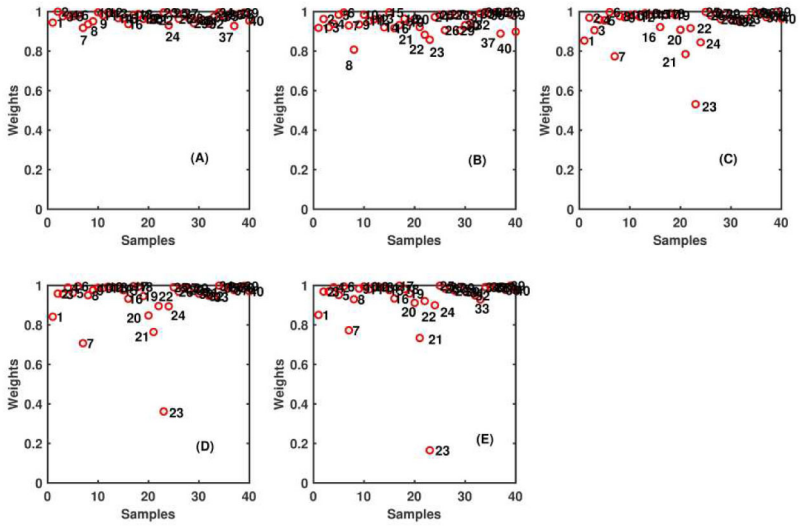

FIGURE 10 Sample weights for the first five (A–E) model components after iterative weighted multiblock modelling of the biscuit data set.

of the presented algorithm to detect the same outliers as detected in earlier studies confirms the usefulness of the presented algorithm for robust multiblock data modelling.

3.4 | Analysis with real outliers—Soil data

Finally, the robust algorithm was also tested on a real multiblock data set related to soil. In the case of soil, the data set has XRF and LIBS spectra to predict exchangeable Ca and Mg content in the soil. The spectral set was divided into calibration and test set by selecting every 3rd sample as the test set and the remaining as the calibration set. Models with 12 components following the extraction order [1 2 1 2 1 2 1 2 1 2] were built on the calibration set and tested on the test set. The models were multi-response models for joint prediction of Ca and Mg content. The prediction plots are shown in Figure 11. The iterative weighted model ($\alpha = 7$) achieved lower RMSEP for prediction of both Ca and Mg

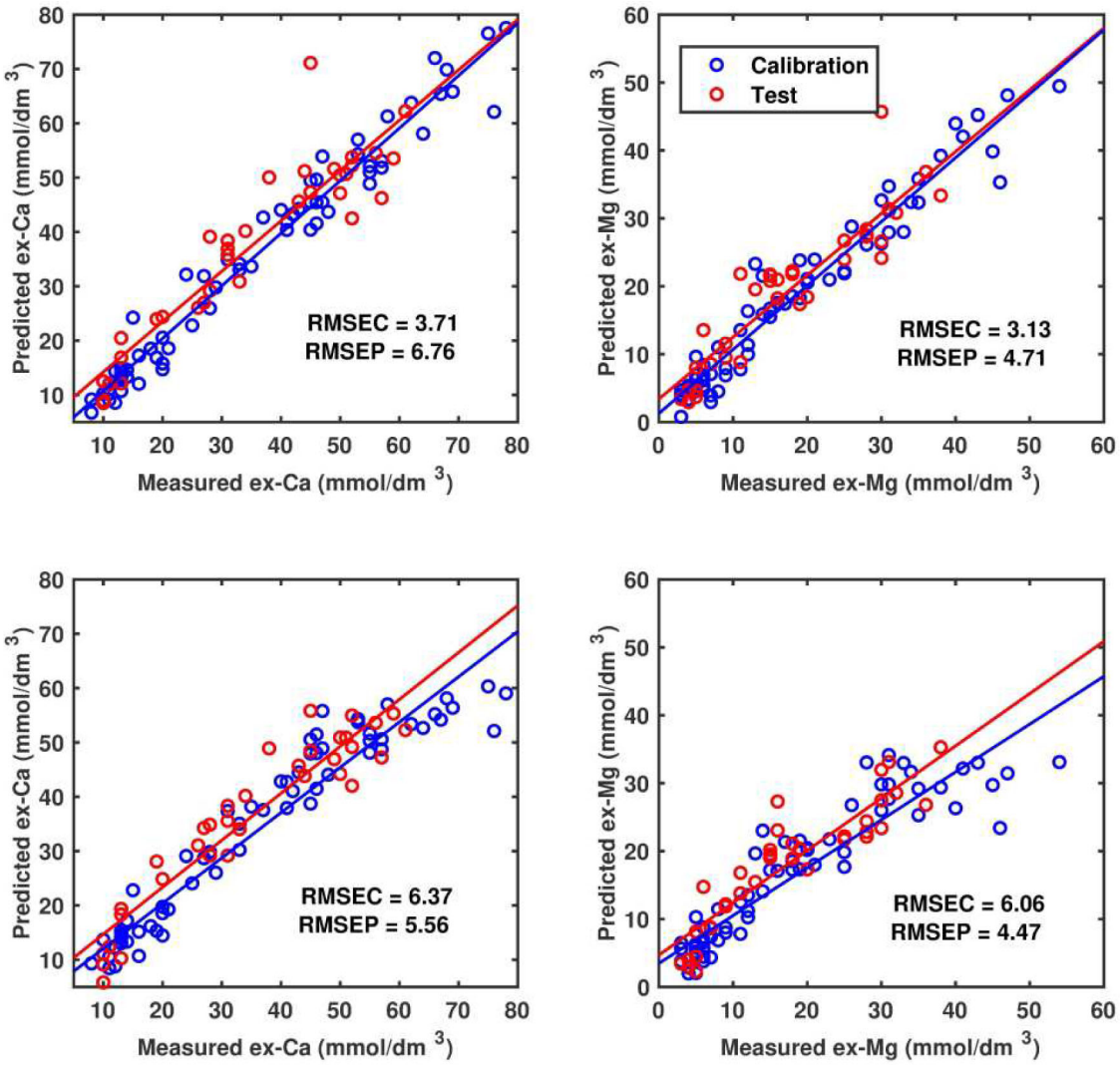

FIGURE 11 Unweighted and iterative weighted multiblock modelling performed on two block soil data. Unweighted modelling to predict (A) calcium and (B) magnesium. Iterative weighted modelling to predict (C) calcium and (D) magnesium.

concentration compared with the unweighted models. The RMSEC for the iterative weighted models was higher than for the unweighted model because the model was poorly fitted on the outlying samples.

4 | DISCUSSION

A new robust method and algorithm for scale-independent multiblock modelling was proposed. The method was tested on a wide range of data modelling scenarios using both simulated and real outliers in the data sets. The test of methods in all cases demonstrated that the method was successfully able to down-weight the outliers and reduce their effect on the model. The models where outliers were down-weighted in general performed better on the test set compared with unweighted models. In general, the robust model has a higher error on the calibration set compared with models that do not deal with outliers; however, that was due to the fact that the robust method fits poorly to the outlying samples in the calibration set. As the obtained sample weights are different from component to component, a weighted calibrated RMSE measure becomes complicated, and may not add value to the assessment of fit.

The component order strategy used in the analyses was one of cycling through the different blocks. There is room for improvement of the predictions and interpretations by exploring other choices for component order. One strategy could be to optimise the order based on unweighted SO-PLS or ROSA and then apply it to the iterative weighted

multiblock method. This would not guarantee optimality for the iterative weighted models as the progression of finding subspaces for each component will be different when samples are weighted differently.

The presented method is the first of its kind, as there currently exist no robust predictive multiblock methods in the scientific literature. There are several open issues to explore regarding sample weighting strategies and extensions, for example, a robust way of estimating common, local and distinct components,[24] as in PO-PLS.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

ORCID

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

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3480.

REFERENCES

1. Zhou L, Zhang C, Qiu Z, He Y. Information fusion of emerging non-destructive analytical techniques for food quality authentication: A survey. *TrAC Trends Anal Chem*. 2020;127:115901.
2. Azcarate SM, Ríos-Reina R, Amigo JM, Goicoechea HC. Data handling in data fusion: methodologies and applications. *TrAC Trends Anal Chem*. 2021;143:116355.
3. Smilde AK, Måge I, Naes T, et al. Common and distinct components in data fusion. *J Chemometr*. 2017;31(7):e2900.
4. Mishra P, Roger J-M, Jouan-Rimbaud-Bouveresse D, et al. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends Anal Chem*. 2021;137:116206.
5. Biancolillo A, Bucci R, Magrì AL, Magrì AD, Marini F. Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication. *Anal Chimica Acta*. 2014;820:23-31.
6. Mishra P, Marini F, Brouwer B, et al. Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit. *Talanta*. 2021;223:121733.
7. Mishra P, Biancolillo A, Roger JM, Marini F, Rutledge DN. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends Anal Chem*. 2020;132:116045.
8. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical pca and pls models. *J Chemometr: A J Chemometr Soc*. 1998;12(5):301-321.
9. Biancolillo A, Næs T. The sequential and orthogonalized pls regression for multiblock regression: theory, examples, and extensions. *Data handling in science and technology*, Vol. 31: Elsevier; 2019:157-177.
10. Naes T, Tomic O, Afseth NK, Segtnan V, Måge I. Multi-block regression based on combinations of orthogonalisation, pls-regression and canonical correlation analysis. *Chemometr Intell Lab Syst*. 2013;124:32-42.
11. Liland KH, Næs T, Indahl UG. Rosa a fast extension of partial least squares regression for multiblock data analysis. *J Chemometr*. 2016;30(11):651-662.
12. Hubert M, Branden KV. Robust methods for partial least squares regression. *J Chemometr: A J Chemometr Soc*. 2003;17(10):537-549.
13. Cummins DJ, Andrews CW. Iteratively reweighted partial least squares: A performance analysis by monte carlo simulation. *J Chemometr*. 1995;9(6):489-507.
14. Serneels S, Croux C, Filzmoser P, Van Espen PJ. Partial robust m-regression. *Chemometr Intell Lab Syst*. 2005;79(1-2):55-64.
15. Hoffmann I, Serneels S, Filzmoser P, Croux C. Sparse partial robust m regression. *Chemometr Intell Lab Syst*. 2015;149:50-59.
16. Metz M, Abdelghafour F, Roger J-M, Lesnoff M. A novel robust pls regression method inspired from boosting principles: Roboost-plsr. *Anal Chimica Acta*. 2021;1179:338823.
17. Metz M, Ryckewaert M, Mas-Garcia S, et al. Roboost-pls2-r: An extension of roboost-plsr method for multi-response. *Chemometr Intell Lab Syst*. 2022;222:104498.
18. Mishra P, Liland KH. Iterative re-weighted covariates selection for robust feature selection modelling in the presence of outliers (ircovsel). *J Chemometr*; 37:e3458.
19. Indahl UG, Liland KH, Næs T. Canonical partial least squares a unified pls approach to classification and regression problems. *J Chemometr: A J Chemometr Soc*. 2009;23(9):495-504.
20. 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.
21. 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.

22. Tavares TR, Molin JP, Nunes LC, Alves EEN, Krug FJ, de Carvalho HWP. Spectral data of tropical soils using dry-chemistry techniques (vnir, xrf, and libs): A dataset for soil fertility prediction. *Data Brief*. 2022;41:108004.
23. Hubert M, Rousseeuw PJ, Van Aelst S. High-breakdown robust multivariate methods. *Stat Sci*. 2008;23(1):92-119.
24. Smilde AK, Næs T, Liland KH. *Multiblock data fusion in statistics and machine learning: Applications in the natural and life sciences*: John Wiley & Sons; 2022.

How to cite this article: Mishra P, Liland KH. An algorithm for robust multiblock partial least squares predictive modelling. *Journal of Chemometrics*. 2023;37(6):e3480. doi:10.1002/cem.3480