

SCREAM: A novel method for multi-way regression problems with shifts and shape changes in one mode

Federico Marini ^{a,*}, Rasmus Bro ^b

^a Dept. of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy

^b Faculty of Science, University of Copenhagen, Frederiksberg C, Denmark

ARTICLE INFO

Article history:

Received 2 January 2013

Received in revised form 3 September 2013

Accepted 20 September 2013

Available online 18 October 2013

Keywords:

Multi-way calibration

Multi-way covariate regression

SCREAM (Shifted Covariate REgression Analysis for Multi-way data)

PARAFAC2

Principal covariates regression (PCovR)

ABSTRACT

Some fields where calibration of multi-way data is required, such as hyphenated chromatography, can suffer of high inaccuracy when traditional N-PLS is used, due to the presence of shifts or peak shape changes in one of the modes. To overcome this problem, a new regression method for multi-way data called SCREAM (Shifted Covariates REgression Analysis for Multi-way data), which is based on a combination of PARAFAC2 and principal covariates regression (PCovR), is proposed. In particular, the algorithm combines a PARAFAC2 decomposition of the X array and a PCovR-like way of computing the regression coefficients, analogously to what has been described by Smilde and Kiers (A.K. Smilde and H.A.L. Kiers, 1999) in the case of other multi-way PCovR models. The method is tested on real and simulated datasets providing good results and performing as well or better than other available regression approaches for multi-way data.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Regression problems, i.e. those problems that are formulated in terms of finding a relation between a set of predictors (independent variables) and one or more responses or properties (dependent variables) are ubiquitous in science. Examples include not only, e.g., the possibility of quantifying the concentration of a chemical substance by measuring a spectroscopic or chromatographic signal, but also cases where the focus is rather on understanding the relations between blocks of variables and not directly on prediction (as, for instance, in sensory science, where a model linking experimental profiles of products to the attributes measured by a panel is sought). As a consequence, several methods are available in the literature to calculate regression models when the blocks involved are either uni- or multivariate, the ones most often used relying on the concept of bilinear decomposition, such as PCR and PLS.

On the other hand, the development of analytical instrumentation is such that many techniques provide measurements that are most meaningfully held as a landscape or a higher order array for each sample. Examples of these are fluorescence excitation emission spectroscopy, multidimensional chromatographic systems or hyphenated instruments, such as GC-MS, HPLC-MS and HPLC-DAD. Common to all is that they produce multi-way data; also called tensor data.

Several methods have been proposed during the years to directly process multi-way data, the most popular being the so-called PARAFAC – PARAllel FACtor analysis – model [1,2]. Analogously, when calibration problems are involved, an extension of the traditional PLS regression to multi-way data has been developed [3].

Both multi-way PLS and PARAFAC are models which assume a common latent structure. Just like, e.g., principal component analysis assumes that any sample can be modeled as a linear combination of a few loadings, PARAFAC also assumes that each sample can be described by the same underlying set of loadings – now in two or more modes. Such models are perfectly suited, for example, when measured data follow Lambert-Beer's law, which specifically states that a measured spectrum is the weighted sum of the same underlying spectra regardless of concentration.

The models are less adequate, when the underlying profiles change shape from sample to sample. For instance, pure spectra of certain chemicals may change shape depending on temperature, or the elution profile of an analyte may have retention time shifts due to column degradation (as a consequence of contaminant buildup or general deterioration). In all cases, bi- or multilinear models are less adequate and need more components to describe the non-linear behavior. For decomposition models, it has been shown that the so-called PARAFAC2 model [4] is sometimes able to provide good models of such data especially in the case of chromatographic signals [5–9].

There have been no serious attempts, however, at making regression models that can handle similar problems as PARAFAC2 can do for decomposition models. Based on a combination of PARAFAC2 and

* Corresponding author at: Tel.: +39 06 4991 3680; fax: +39 06 445 7050.

E-mail addresses: fmmonet@hotmail.com, federico.marini@uniroma.it (F. Marini).

principal covariates regression [10], we developed a regression method called SCREAM (Shifted Covariates REgression Analysis for Multi-way data), which allows 'jittering' in one of the modes in a multi-way array and tested it on many different types of data, real and simulated.

2. Theory

The aim, in the following, is to develop a regression model that can provide latent variables which are useful for predictive models yet also allow a PARAFAC2 like deviation from multi-linearity. de Jong and Kiers [10] developed a framework for building multivariate bilinear regression models called principal covariates regression. This was later extended to multi-linear regression models by Smilde and Kiers [11]

and here we will show how we can enable a PARAFAC2 model structure and also develop the associated algorithm based on an Alternating Least Squares (ALS) approach.

The PARAFAC2 model is normally written as

$$\mathbf{X}_k = \mathbf{A} \mathbf{D}_k \mathbf{B}_k^T + \mathbf{E}_k \quad k = 1, \dots, K \quad (1)$$

where \mathbf{X}_k is one $I \times J$ slab of an $I \times J \times K$ three-way array, \mathbf{X} , to be modeled by the PARAFAC2 model [4,5,12,13]. The matrix \mathbf{A} ($I \times F$) is the first mode loadings of an F -component PARAFAC2 model. The matrix \mathbf{D}_k ($F \times F$) is a diagonal matrix that holds the k 'th row of the third mode loadings \mathbf{C} ($K \times F$) which is usually the sample mode in PARAFAC2. Finally \mathbf{B}_k ($J \times F$) is the loadings for second mode for sample k . These are subjected to the constraint that $\mathbf{B}_k^T \mathbf{B}_k = \mathbf{H}$ for all $k = 1, \dots, K$. This constraint is necessary

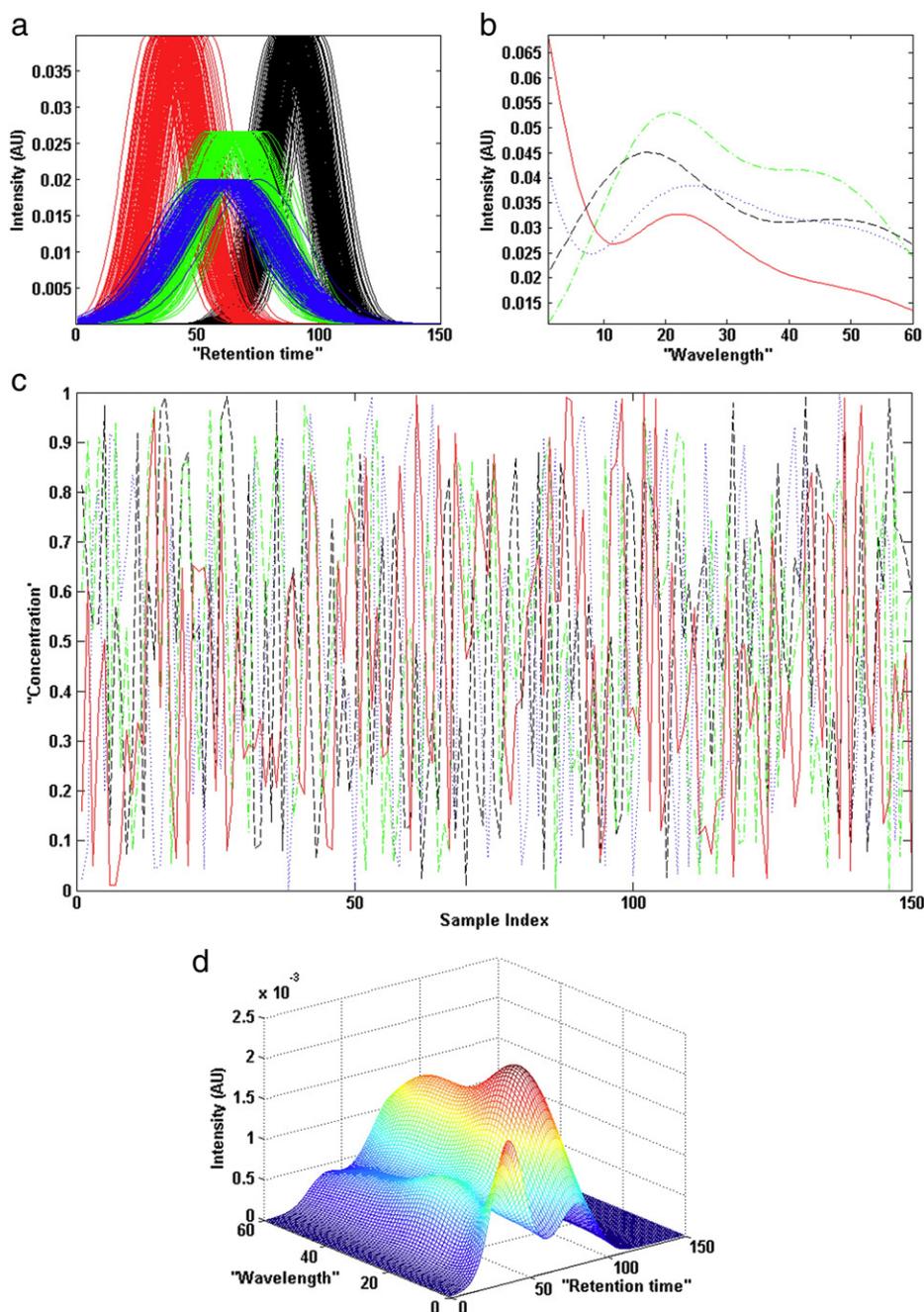


Fig. 1. Simulated dataset. Graphical representation of the profiles of the four constituents used for building the dataset: (a) elution-like mode; (b) spectral-like mode; (c) concentration mode [continuous red line: 1st constituent; dotted blue line: 2nd constituent; dashed dotted green line: 3rd constituent; dashed black line: fourth constituent]. An example of the 2D landscape for one of the simulated samples is reported in panel (d).

to provide an identified solution in PARAFAC2, which, under mild conditions, is unique. The matrix \mathbf{E}_k holds the residuals of the PARAFAC2 model.

Normally, the parameters of the PARAFAC2 model are determined in a least squares sense so as to minimize the loss function

$$\sum_{k=1}^K (\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}_k^T)^2 = \|\mathbf{X} - \mathbf{C}\mathbf{P}^T\|_F^2 \quad (2)$$

where \mathbf{X} is the three-way array unfolded into a $K \times IJ$ matrix and \mathbf{P} is a matrix holding the matrix \mathbf{A} and the matrices \mathbf{B}_k appropriately arranged.

In order to fit a PARAFAC2 model structure relevant for prediction, a model similar to this is fitted to \mathbf{X} with the third mode being the sample mode. Hence, the loadings, \mathbf{C} , will be the sample scores. At the same time, that \mathbf{C} is used for fitting the PARAFAC2-like model to \mathbf{X} , it is sought that the components in \mathbf{C} should be relevant for predicting a dependent variable, \mathbf{y} ($K \times 1$). This is done by minimizing the regression problem

$$\|\mathbf{y} - \mathbf{C}\mathbf{r}\|_F^2 \quad (3)$$

where \mathbf{r} ($F \times 1$) is a vector of regression coefficients. The two loss functions in Eqs. (2) and (3) can be combined into one, but there are two issues that must be handled. First of all, a metaparameter is needed

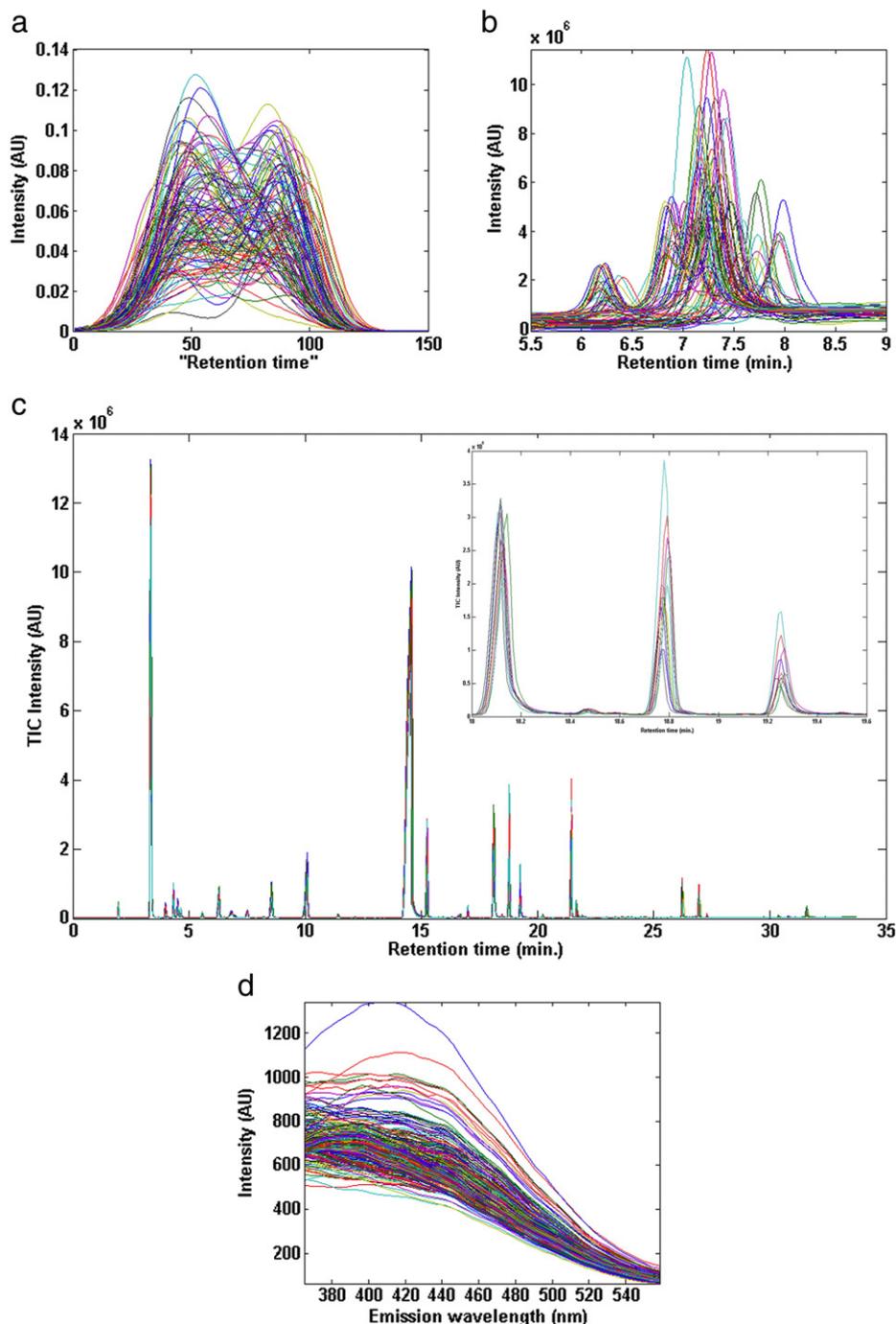


Fig. 2. (a) Simulated dataset: Overlapped "total absorbance chromatograms" for the 120 training samples, showing shifts in the elution-like mode. (b) Olive oil dataset: Overlapped total absorbance chromatograms for the 76 training samples. (c) Wine dataset: Overlapped TIC chromatograms for the 33 training samples (a magnification of a region of the profile better illustrating the kind of shifts involved is shown in the inset). (d) Sugar dataset: Overlapped total emission spectra for the 188 training samples.

for determining to which extent the model of \mathbf{X} or the model of \mathbf{y} is prioritized. This is done by introducing a scalar, α , $0 \leq \alpha \leq 1$, and minimizing

$$\alpha \|\mathbf{X} - \mathbf{C}\mathbf{P}^T\|_F^2 + (1 - \alpha) \|\mathbf{y} - \mathbf{C}\mathbf{r}\|_F^2 \quad (4)$$

over \mathbf{C} and \mathbf{P} . The regression vector will be implicitly defined by \mathbf{C} as in an ordinary multiple linear regression problem. Each block, \mathbf{X} and \mathbf{y} , is assumed scaled to a total sum of squares of one in order to make it easier to assess and decide α . Eqs. (3) and (4) can be easily generalized to the case when multiple \mathbf{Y} have to be predicted (See Appendices A and B for a formal treatment): in this case, it must be stressed that, differently to what occurs in the case of other component regression models (e.g., PLS), the \mathbf{Y} matrix is not decomposed further.

In practice, minimizing the above loss function could lead to a component matrix \mathbf{C} that is not in the space of \mathbf{X} and hence would be meaningless for predictions. In order to ensure that components are relevant for prediction, it is specified that $\mathbf{C} = \mathbf{X}\mathbf{W}$ where \mathbf{W} ($I \times F$) is a weight matrix that specifically ensures that \mathbf{C} is in the row-space of \mathbf{X} .

In order to choose the appropriate number of components as well as the optimal value of α , cross-validation is used. In particular, for each cross-validation split, models are built including an increasing number of components and, for each number of components investigated, varying α according to a grid search (in the present study, values from 0 to 1 in 0.1 steps were considered). Accordingly, the Root Mean Squared Error of Cross-Validation (RMSECV) is a function of both α and the number of factors:

$$\text{RMSECV}(\alpha_k, f) = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_{i,CV}(\alpha_k, f))^2}{N}} \quad (5)$$

where $\hat{y}_{i,CV}(\alpha_k, f)$ is the predicted value of the response for the i th sample in cross-validation (i.e., when the sample is left out), calculated using a model with f components and α set to the value α_k ; N is the total number of training samples, while y_i is the actual value of the response for sample i . Therefore, the combination of the two parameters will be chosen as the one leading to the minimum error.

In Appendices A and B, the algorithm is detailed and MATLAB implementation is also available at www.models.life.ku.dk (September 2013).

3. Experimental

To explore the characteristics of the proposed method in detail and to verify its performance, also compared to other approaches for building calibration models, one simulated and three real world datasets have been investigated.

3.1. Simulated dataset

To get insight into the method and study the effect of the adjustable parameter α on the results, also as a function of the amount of noise in the data, a three-way HPLC-DAD-like dataset (spectral signals \times elution profiles \times samples) was simulated. In particular, data have been generated considering a hypothetical problem involving four chemical constituents whose profiles are reported in Fig. 1. Shift was allowed in the second mode, i.e. the one corresponding to the elution profiles. Elution profiles for the four constituents (150 data points) were simulated as Gaussian peaks centered in 40, 60, 65 and 90, and having a standard deviation of 10, 15, 25 and 10, respectively. Then, to obtain the profiles for the individual samples, a random and Gaussian distributed shift (centered in zero and with a standard deviation of five data points) was applied to the peak positions. Simulated concentrations for the four constituents were randomly sampled from a uniform distribution in $[0,1]$. The spectral-like profiles were simulated

by combining two or three Gaussian bands opportunely spaced along the axis. In total, data for 120 training and 30 test samples were generated and collected in the two tensors \mathbf{X}_{Tr} and \mathbf{X}_{Ts} , having dimensions 60 (wavelengths) \times 150 (elution times) \times 120 (samples) and $60 \times 150 \times 30$, respectively. The chromatographic profiles for the training samples, obtained by summing, for each retention time, the spectral intensities at all wavelengths (i.e., the two-way matrix obtained summing the data tensor along the third mode), are shown in Fig. 2a, to illustrate which kind of shifts can be expected for these data. Noise (Gaussian iid) in different proportions (0%, 5%, 10% and 20%) was then added to these arrays, which were then used to build a calibration model for the quantification of the amount of the four constituents.

3.2. Olive oils dataset

The second dataset used is a real world dataset coming from the HPLC-DAD analysis of olive oil samples to quantify different phenolic acids. It contains a portion of the HPLC-DAD profiles recorded on 76 training and 21 test samples, corresponding to the retention time window where four compounds elute, namely *p*-hydroxybenzoic, syringic, vanillic and caffeic acids. The resulting multi-way arrays have dimensions 52 (wavelengths) \times 1052 (retention times) \times 76 (samples) and $52 \times 1052 \times 21$, respectively. Also in this case, shifts are present along the elution mode, as can be seen in Fig. 2b, where the chromatographic profiles for the training samples are shown. These are obtained by summing, for each retention time, the spectral intensities over all wavelengths (i.e., analogously to what was already described for the simulated dataset, they correspond to the two-way matrix obtained summing the data tensor along the third-spectral-mode). Calibration models were built to quantify the amount of the four chemical compounds in the mixtures. Further details on the dataset can be found in Ref. [14].

3.3. Wine dataset

The third dataset is made of the GC-MS landscapes measured on wine samples from different origins for the quantification of different analytes; it was originally published in Ref. [15], where models have been built on the data after alignment. In the present work, to illustrate the ability of the proposed method to deal with multi-way data

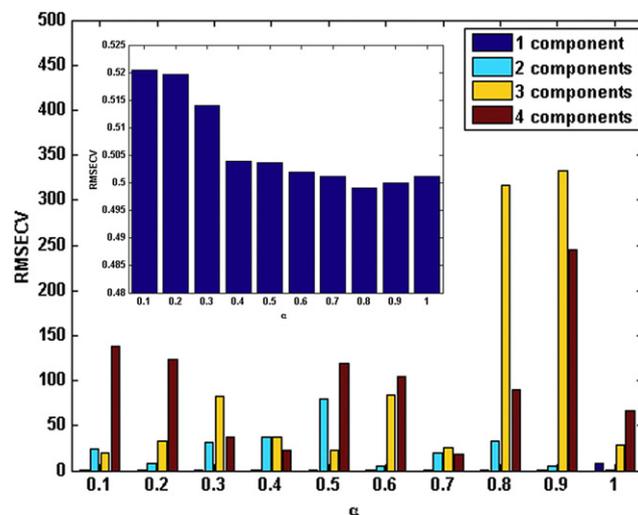


Fig. 3. Simulated dataset: selection of the optimal value of α for the prediction of the first constituent. Illustration of the dependence of the RMSECV on the number of SCREAM components and the value of α (inset) Dependence of the RMSECV on the value of α for one component SCREAM models.

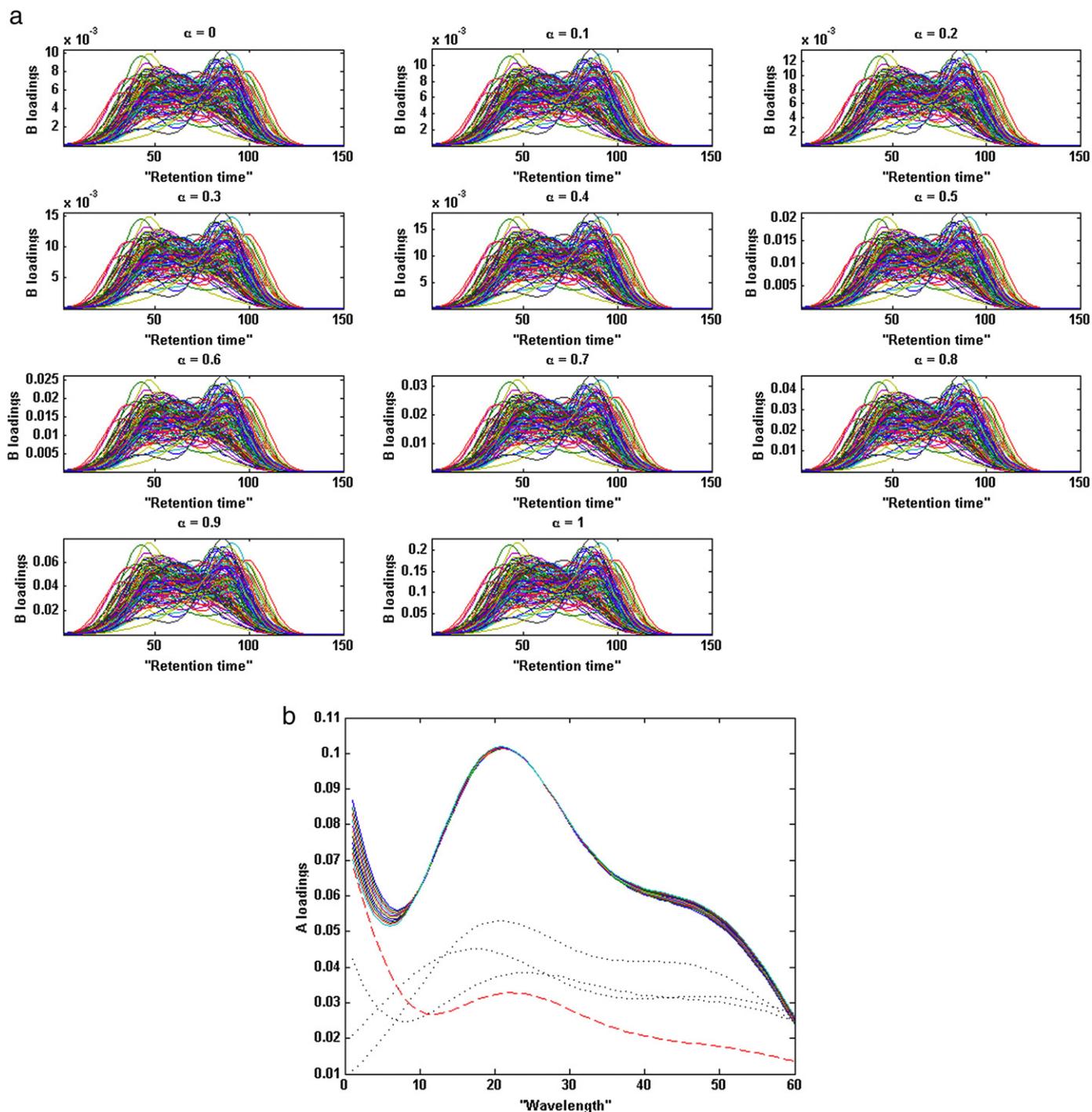


Fig. 4. Simulated dataset; prediction of the concentration of the first constituent. (a) Overlapped **B** loadings of the single SCREAM component for the 120 training samples as a function of the parameter α . (b) Overlapped **A** loadings of the single SCREAM component as a function of the parameter α , superimposed to the spectral-like profiles of the four constituents (red dashed line: 1st constituent; black dotted lines 2nd–4th constituents).

presenting shifts or shape changes along one mode, calibration models were built on the unaligned array. To this purpose, the dataset, which was originally made on the experimental profiles recorded on 44 samples was split into two data cubes of dimensions 200 (m/z ratios) $\times 2700$ (retention times) $\times 33$ (samples), and $200 \times 2700 \times 11$, respectively, for training and validation, using the duplex algorithm [16] on the TIC chromatograms (Fig. 2c). The duplex algorithm was chosen as it guarantees that the same diversity is preserved in both sets. Indeed,

it starts selecting the two objects in the data matrix that are farthest away from each other according to their Euclidean distance and putting them into the training set. Then, among the remaining candidates, the two objects farthest from each other are put into the test subset. At the next step, consecutive objects are selected and put alternatively in the training and test sets, the object added being the one farthest away from all the objects of the data matrix already selected in the considered set. To determine which object is the farthest one, a so-called maximin criterion

Table 1

Root Mean Square Error in Prediction (RMSEP) for the quantification of the 4 constituents of the simulated dataset as a function of the noise level.

Noise level	RMSEP			
	Y1	Y2	Y3	Y4
0%	0.11	0.15	0.16	0.12
5%	0.25	0.50	0.32	0.40
10%	0.48	0.56	0.65	0.45
20%	0.76	1.75	0.98	0.48

is used: the Euclidean distance between each candidate object and its closest neighbor already in the considered subset is computed and the object for which this distance is maximal is added. For the present study, calibration of only one of the quality parameters measured in the original study (total acidity) was considered.

3.4. Sugar dataset

The last dataset to be analyzed is the only one where no shifts are present and it was introduced to check whether the proposed method could induce severe artifacts in a situation where the trilinear structure of the X-block is not affected by shifts. The data contain the fluorescence excitation emission landscapes measured on 268 sugar process samples to predict the value of one quality parameter (color). The dataset was divided into training (180) and test (88) sets using the duplex algorithm [16] on the two-way matrix resulting after row-wise unfolding. Seven wavelengths are present in the excitation mode and 71 in the emission one, which was considered as the shifting mode. In Fig. 2d, the spectra resulting from summing, for each sample, the emission profiles at all excitation wavelengths are shown. Further details on the dataset can be found in the original paper [17].

4. Results

In order to explore the main characteristics of the proposed method and to evaluate its performances in different real world situations, four datasets have been analyzed. In the remainder of this paragraph, the results obtained in the different cases will be separately described and discussed.

4.1. Simulated dataset

The first dataset to be analyzed was the simulated one, containing HPLC-DAD-like profiles to be used for the quantification of four components in the presence of increasing noise levels. At first, the noiseless situation was studied. As there were four dependent variables (amount of constituents) to be predicted, a first choice to be made was whether to calibrate all the concentrations together or to build four separate models, one for each dependent variable. To get better insight in the method behavior, it was decided, in the case of the simulated dataset, to calibrate one Y variable at a time. Since the model contains two parameters to be optimized (the number of components and the value of the tuning parameter α), 5-fold cross-validation was used, so that the optimal complexity and the amount of X- and Y-blocks variance to be explained by the model were chosen as those leading to the minimum RMSECV. In particular, a grid search was performed by varying α from 0 to 1 in steps of 0.1 and examining from 1 to 8 components at each value of the metaparameter. The results for the calibration of the 1st simulated Y variable are graphically illustrated in Fig. 3, where the RMSECV values corresponding to the different model complexities are reported as a function of α . In all examined cases, increase in the value of α resulted in a corresponding increase in the explained X-variance which e.g., for the single component models, varied from 80–82% ($\alpha = 0$) to 89–92% ($\alpha = 1$). Here it must be stressed that the case corresponding to $\alpha = 0$ is degenerate, as is also clear from

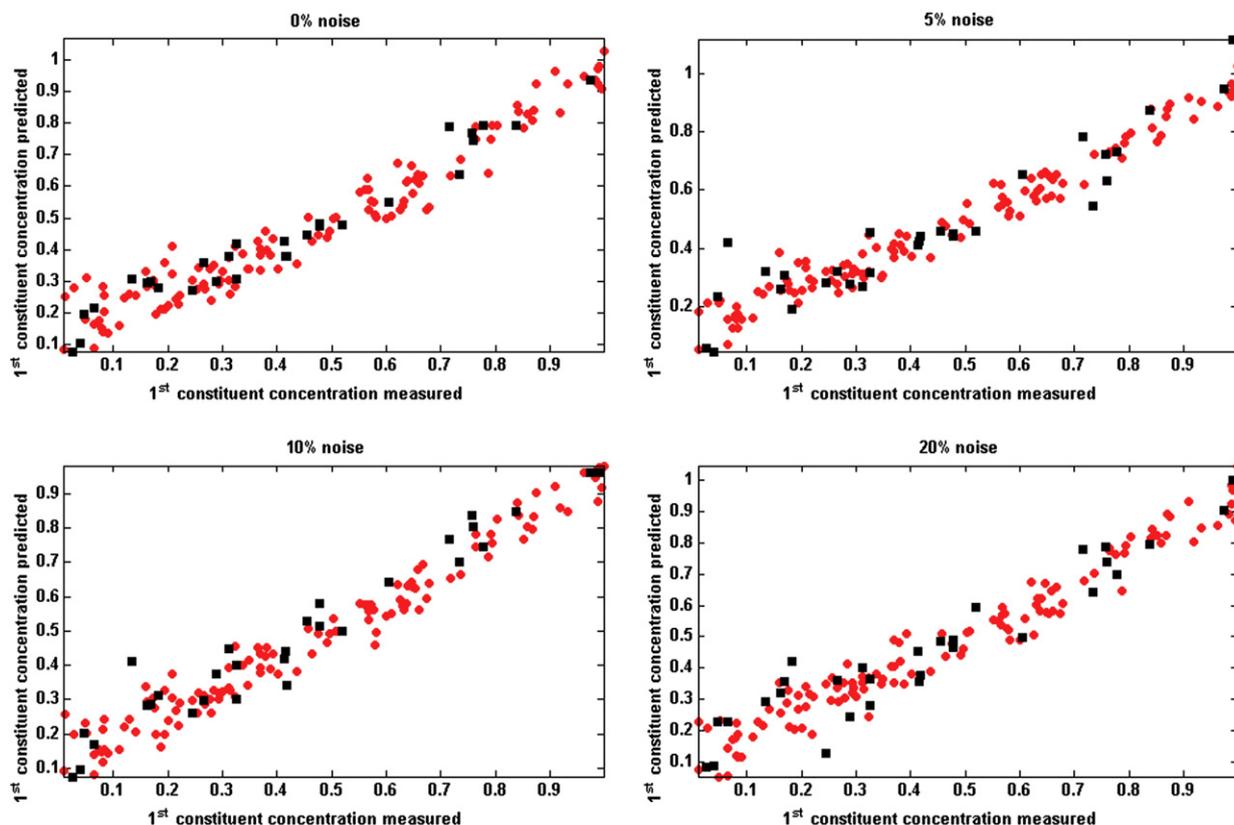


Fig. 5. Simulated dataset; SCREAM model for the prediction of the concentration of the first constituent. Observed vs. predicted plot for the training (●) and test (■) sets, as a function of the different noise levels investigated.

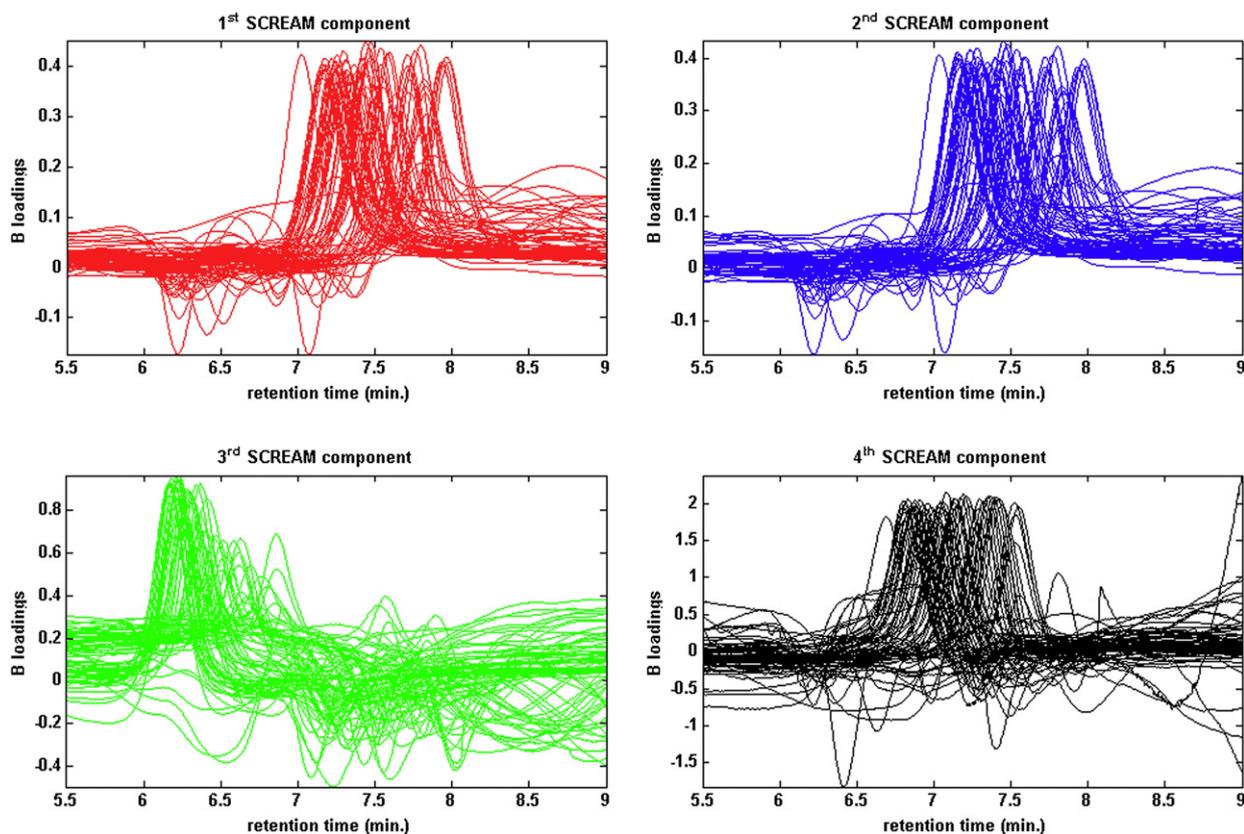


Fig. 6. Olive oil dataset. Graphical representation of the elution (B) loadings of the SCREAM model.

Eq. (4) and already pointed out by Smilde and Kiers [11]. This setting was therefore omitted. Similarly, since the values of the cross-validated error obtained with 5–8 LVs were significantly higher than the others, they were not shown in the graph for the sake of an easier visualization.

Fig. 3 suggests that the optimal model for the prediction of the first Y variable uses a single SCREAM component. The RMSECV is consistently lowest for one component. On the other hand, at least for the case where no noise is added, the value of α seems to have little impact on the results. Regardless, the plot in the inset of Fig. 3 shows that the value corresponding to the minimum cross-validation error is 0.8 and therefore this was the chosen one. Similar results were obtained also for the prediction of the other three Y variables, where the optimal models were those including a single factor, and the values of RMSECV changed only slightly as a function of the metaparameter α (the best results having been obtained setting $\alpha = 1$). The little impact of the value of α on the model performances when noise is absent is reflected also on the observation that the X-loadings are practically identical no matter the value of α . In particular, in Fig. 4a and b the “elution” and “spectral” loadings for the training samples in the case of the model for the prediction of the first Y variable are reported as a function of α , showing very little differences. Similar observations are made in the cases where the other three variables have to be calibrated.

The impact of noise on the results was considered, by adding 5%, 10% and 20% random noise to the X-block landscapes and rebuilding the calibration models using the proposed method. The first difference that was observed was that, while the optimal complexity remained one component for each of the models, the optimal value of α shifted to lower values (0.1 in all cases), indicating that a greater amount of the Y-block variation needed to be taken into account to obtain better predictions. When considering the predictive ability of the models, it

was observed that very good results could be obtained for all the Y variables also in the presence of 20% noise, as shown in Table 1, where the RMSEP on the 30 test set samples is reported as a function of the noise level for all the four constituents. The same information can also be represented in the form of an observed versus predicted scatterplot, which, in the case of the first Y variable, is shown in Fig. 5 as an example.

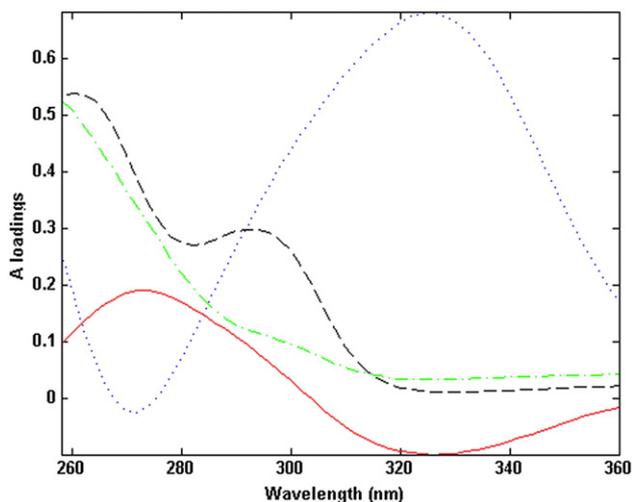


Fig. 7. Olive oil dataset. Representation of the spectral (A) loadings of the SCREAM model [continuous red line: 1st component; dotted blue line: 2nd component; dashed dotted green line: 3rd component; dashed black line: 4th component].

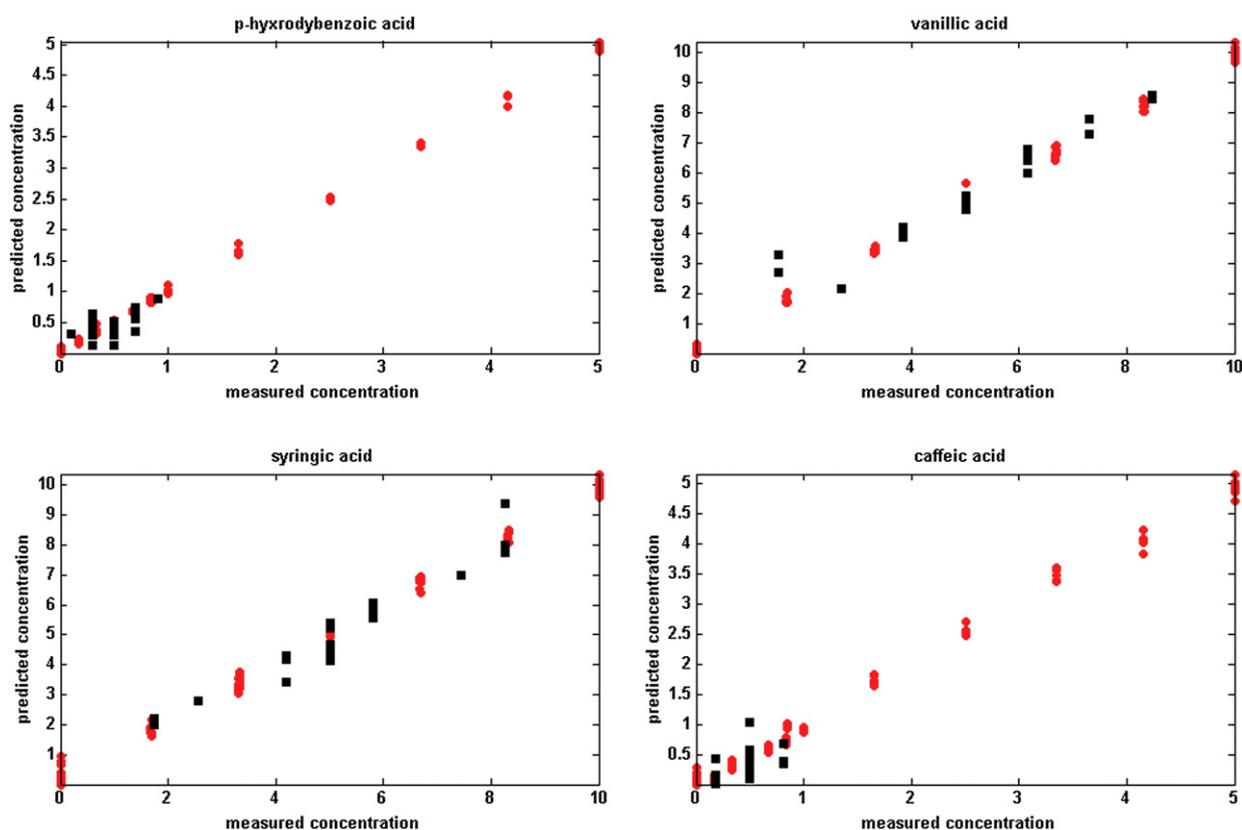


Fig. 8. Olive oil dataset; SCREAM model for the prediction of the concentration of the four analytes. Observed vs. predicted plot for the training (●) and test (■) sets.

It can be observed in the figure that even if the predictions are worsened by the addition of noise, still an acceptable error can be obtained.

For the sake of comparison, the same dataset was analyzed by building a single model for the simultaneous prediction of all four Y components instead of four individual ones. In this case, independently on the noise level considered, the optimal complexity was found to be four components, as expected, while the optimal α value was 0.7 for the noiseless data and 0.3 for the others. In terms of predictions, the results obtained are quite consistent with those reported in Table 1, as the difference in the observed RMSEP values was always less than 5%. On the other hand, the loadings obtained from the model computed on the full Y matrix more closely resemble the experimental fingerprints of the pure components and therefore are more easily interpretable. For this reason, and also to take advantage of the need to build a single model only, this approach was chosen for the analysis of the olive oil dataset, which is the other dataset considered in this study containing multiple responses.

4.2. Olive oil dataset

The second dataset to be analyzed was the HPLC-DAD dataset for the quantification of four phenolic acids in extra virgin olive oil samples. In

this case, building a model to quantify all the four dependent variables together in order to have loadings, which could more easily be interpreted in terms of chemical constituents was decided. Based on the results from a 10-fold cross-validation, it was observed that the optimal complexity of the model was with four components, while the best value of α was 0.3. Inspection of the model loadings for the elution and spectral modes, reported in Figs. 6 and 7, respectively, show that they very closely resemble the elution profiles and the spectra of the pure components, even with a relatively low value of α .

When the model was used to predict the concentrations of the four phenolic acids in the training and test mixtures, good results were obtained, as shown in Fig. 8, where the predicted versus measured plot for the four constituents are reported. The errors are of the same order of magnitude for all the four constituents and they are all sufficiently low to guarantee accurate predictions.

To compare the results of the proposed method on this dataset with other approaches, the same dataset was analyzed by PLS after unfolding, by N-PLS on the original data and by N-PLS on the data after alignment using the iCoshift algorithm [18]. In all cases, the optimal number of components was selected as the one, which resulted in the lowest error in 10-fold cross-validation. The parameters of iCoshift were selected manually in order to have visually assessed well-aligned data.

Table 2

Olive oil dataset: comparison of the results obtained by the different regression approaches.

Method	p-Hydroxybenzoic		Vanillic		Syringic		Caffeic	
	RMSEP	Bias	RMSEP	Bias	RMSEP	Bias	RMSEP	Bias
Unfolding + PLS	0.30	−0.025	1.42	−0.30	1.15	0.53	0.26	0.19
Alignment + N-PLS	0.55	−0.04	1.80	−0.48	1.79	0.90	0.34	0.21
N-PLS	0.47	0.32	3.39	1.38	4.50	−3.74	0.42	−0.37
SCREAM	0.25	0.01	0.43	−0.03	1.06	0.35	0.18	0.07

The outcomes are reported in Table 2. It can be seen from the table that SCREAM practically has no bias and, in general, significantly outperforms the other approaches. The bias was estimated as the average of the prediction residuals:

$$\text{bias} = \frac{\sum_{j=1}^{n_{\text{test}}} (y_j - \hat{y}_j)}{n_{\text{test}}} \quad (6)$$

y_j and \hat{y}_j being the actual and predicted value of the response for the j th test object and n_{test} being the total number of validation samples.

4.3. Wine dataset

In the case of the wine dataset, where the GC–MS profiles of the samples were used to predict the total acidity, the optimal number of components was found to be five while the value of the α parameter which led to the best results in 6-fold cross-validation was 0.3, as in the case of the olive oil dataset. The loadings for the three modes of the X-block, obtained from the model corresponding to these optimal settings, are reported in Fig. 9.

In the previous case, the property to be quantified was the concentration of the different constituents which were part of a coeluting signal (real or simulated). In the present case there is no direct correspondence between a specific portion of the chromatographic

landscape and the y variable to be predicted. This is implicitly reflected in that the elution and spectral loadings are affected by different portions of the signals. Still the loadings reflect meaningful portions of the profiles (see Fig. 2c). When the model was applied for the prediction of the total acidity on the 11 test samples, a reasonable value of the RMSEP (0.29) was obtained, as also observable in Fig. 10, where the observed versus predicted plot for the training and test samples is shown.

To compare the results of SCREAM on the wine dataset with other approaches, the data was analyzed by PLS after unfolding, by N-PLS on the original data and by N-PLS on the data after alignment using the iCoshift algorithm [18]. The outcomes are reported in Table 3. As observed for the olive oil dataset, the table evidences how SCREAM results in the lowest values of RMSEP and bias, and outperforms the other approaches.

4.4. Sugar dataset

Lastly, a dataset where no shifts are present was analyzed to verify to what extent the proposed method could introduce artifacts or result in overfitting. To this end, the sugar dataset was analyzed using the proposed method and the outcomes were compared to the ones of N-PLS. Both models, resulted in two components as the optimal complexity (as evaluated by 10-fold cross-validation) while the best value of the parameter α was 0.5 (which in covariate regression

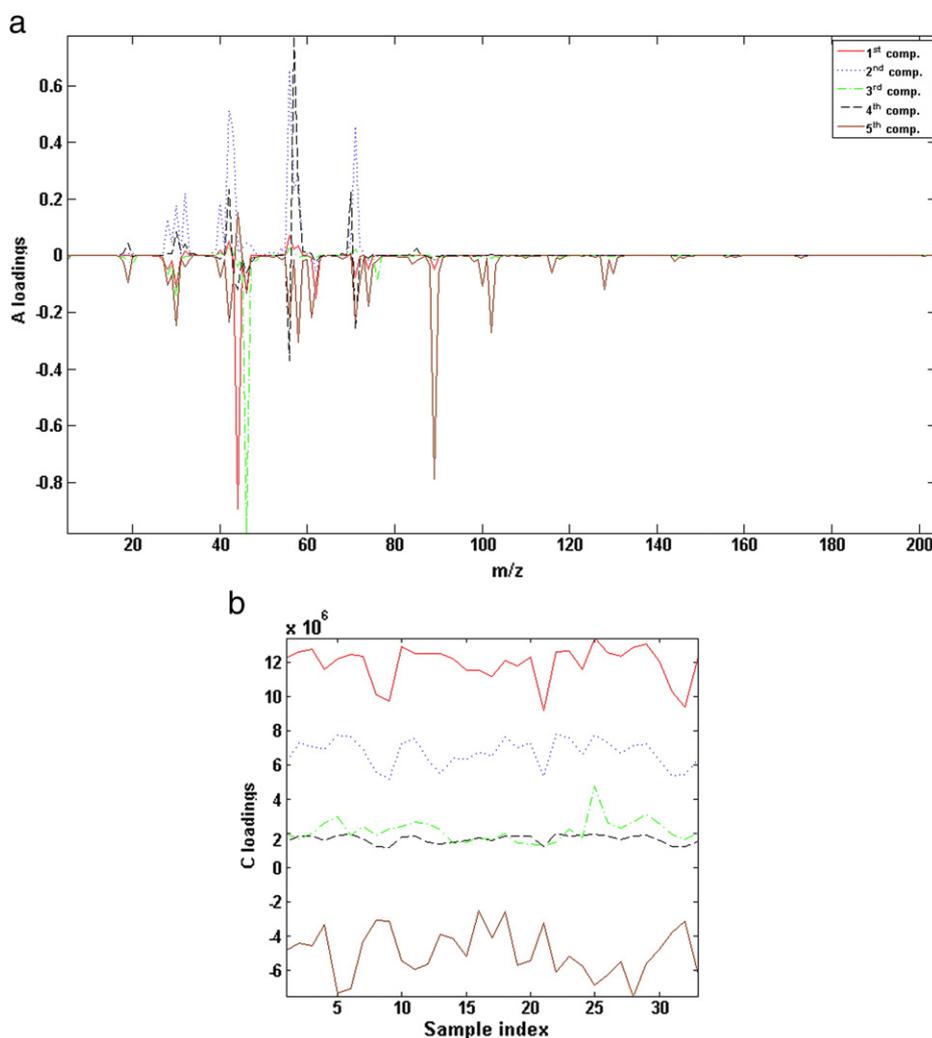


Fig. 9. Wine dataset. Characteristics of the optimal SCREAM model: (a) Spectral (A) loadings; (b) concentration (C) loadings; (c) elution (B) loadings. A detail of the elution loadings is reported in panel (d).

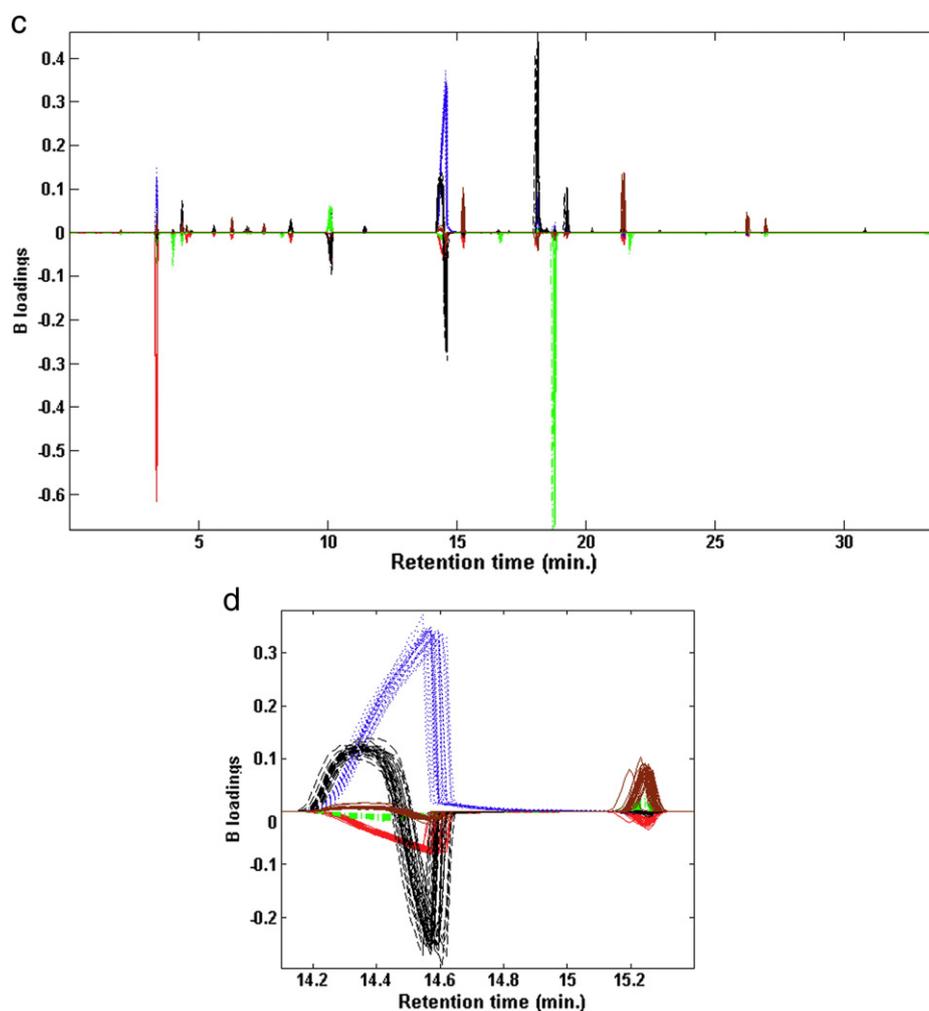


Fig. 9 (continued).

corresponds to a weighting scheme for the two blocks which is somewhat analogous to PLS). The two models showed comparable predictive ability both on the training and the test samples, as illustrated in Table 4, where the bias and SEP obtained using the two approaches are summarized. Especially when comparing the outcomes reported in Table 4, it can be seen that the differences in bias and standard error of prediction which result from the use of the two methods are small.

Moreover, when inspecting in greater detail the model structure, and in particular the loadings obtained by the two methods for the mode which was set as the “shifting” mode, i.e., the emission one (Fig. 11), it can be seen that very similar profiles are obtained (apart from a sign difference due to sign ambiguity). This confirms that even when no shifts are present, the proposed approach is apparently not introducing artifacts or severe overfitting.

5. Concluding remarks

A regression method for dealing with multi-way calibration problems in cases where shifts or shape changes are present along one of the modes was presented and discussed. The model is based on modifying the multi-way covariates regression algorithm proposed by Smilde and Kiers [11], to introduce a PARAFAC2 engine in the decomposition stage. The model has been tested on different simulated and real datasets with good results. It was shown that in the case of noiseless data, the value of the adjustable parameter α , which regulates the amount of X- and Y-blocks variation fitted by the model, does not have much influence on the results, while it becomes relevant when

noise is present. It was observed that lower values of α are favored in the investigated examples. These findings are in agreement with the conclusions of a recently published study by Vervloet et al. [21], who showed by an extensive simulation that, at least in the case of two-

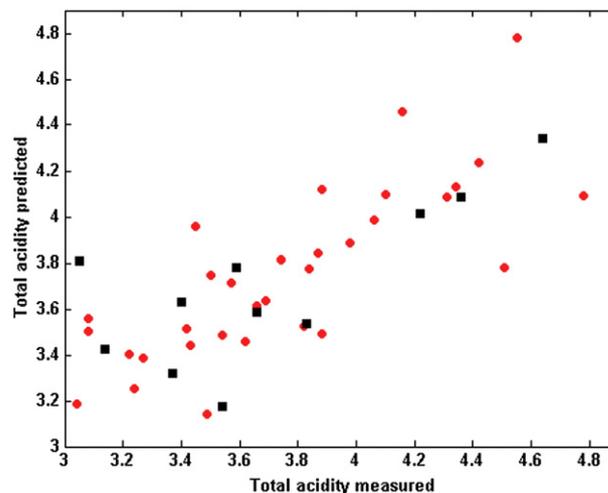


Fig. 10. Wine dataset; SCREAM model for the prediction of total acidity. Observed vs. predicted plot for the training (●) and test (■) sets.

Table 3
Wine dataset: comparison of the results obtained by the different regression approaches.

Method	RMSEP	Bias
Unfolding + PLS	0.33	0.01
Alignment + N-PLS	0.50	−0.04
N-PLS	0.51	0.09
SCREAM	0.29	0.01

way data, the value of the metaparameter, leading to the best recovery of underlying parameters or the best prediction of future data, depends at least on the number of predictors and on the ratio of the amount of error on the predictors and the amount of error on the criterion. In this respect, we plan to carry out a properly designed simulation study to verify how their conclusion can be extended to the proposed method. As the model corresponding to $\alpha = 1$ is statistically equivalent to MLR on PARAFAC2 scores, the observation that, in almost all the studied cases, lower values of the metaparameter are favored confirms that coupling the decomposition and the calibration stages in a single loss function leads to improved predictions, compared to performing the two steps separately.

When shifts are present along one mode, it was shown that the proposed method outperforms other calibration approaches, such as PLS after unfolding, N-PLS (without or after alignment of the profiles). On the other hand, even when no shifts are present in the dataset, the method seems to give comparable results when compared to N-PLS, thus suggesting that little overfitting is introduced by the additional flexibility of the model.

The theory of SCREAM is intrinsically related to PARAFAC2 and just like the PARAFAC2 model, SCREAM can be applied to a variety of data where normal models would fail. For instance, SCREAM can be applied to multiblock data. The PARAFAC2 model was originally developed for use in scenarios where individual observations were of varying dimensionality. That is, situations where each sample was measured at slightly different conditions. In the context of chromatographic data, it could, for example, be the case that individual experiments were of varying length. This would normally require some sort of alignment or truncation of the data, but PARAFAC2 was originally developed for modeling such data directly. As long as the underlying latent variables are the same, PARAFAC2 can handle that there are variations in how these latent variables are manifested in the actual measurements. Hence, similarly to e.g. Generalized Procrustes Analysis [19], PARAFAC2 and hence SCREAM may be used for more general multiblock data and with very few metaparameters that need to be decided.

We believe that SCREAM could be a general and valuable regression tool for dealing with multi-way calibration problems and believe that it will find use in diverse areas of applications yet to be discovered.

Appendix A

The complete ALS algorithm for estimating the parameter values of the proposed SCREAM method is given below. Here, it is assumed that the X-loadings for the second mode (the one which presents 'shifts') \mathbf{B}_k can be expressed as a product of a sample-dependent orthonormal

matrix \mathbf{Q}_k and the small matrix \mathbf{H} , as in the direct fitting PARAFAC2 approach:

$$\mathbf{B}_k = \mathbf{Q}_k \mathbf{H} \quad \text{for } k = 1, \dots, K. \quad (\text{A.1})$$

Accordingly, the steps of the SCREAM algorithm are the following:

Initialize the values of \mathbf{A} , \mathbf{C} , \mathbf{H} and \mathbf{R} . For instance, according to what was suggested by Kiers for PARAFAC2 [20], setting \mathbf{C} to ones, \mathbf{A} to the first singular values of $\sum_{k=1}^K \mathbf{X}_k \mathbf{X}_k^T$, and \mathbf{H} to the identity matrix. Then, the first estimate of \mathbf{R} can be obtained as:

$$\mathbf{R} = \mathbf{C}^+ \mathbf{Y}. \quad (\text{A.2})$$

Estimate the value of \mathbf{Q}_k , using the direct fitting PARAFAC2 approach. For every $k = 1, \dots, K$:

$$\mathbf{S}_k = \mathbf{X}_k^T \mathbf{A} \mathbf{D}_k \mathbf{H}^T \quad (\text{A.3})$$

$$\mathbf{Q}_k = \mathbf{S}_k \left(\mathbf{S}_k^T \mathbf{S}_k \right)^{-0.5}. \quad (\text{A.4})$$

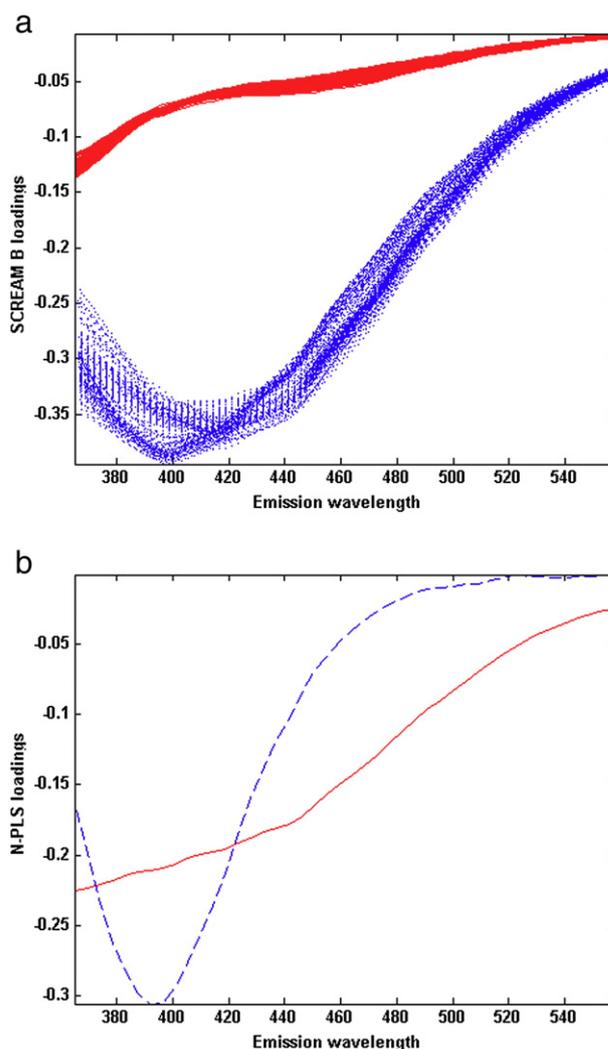


Fig. 11. Sugar dataset. (a) SCREAM loadings of the emission mode (\mathbf{B} loadings); (b) N-PLS loadings of the emission mode. In both cases, continuous red line indicates the first component and dashed blue line the second one.

Table 4
Sugar dataset: comparison of the results obtained by SCREAM and N-PLS.

Method	RMSEP	Bias
N-PLS	2.74	−1.05
SCREAM	2.85	−0.92

Use \mathbf{Q}_k to build the three-way array of cross-products:

$$\mathbf{Z}_k = \mathbf{X}_k \mathbf{Q}_k \quad \text{for } k = 1 : K. \quad (\text{A.5})$$

Update \mathbf{A} , \mathbf{C} , \mathbf{H} , \mathbf{R} and \mathbf{W} :

Update of \mathbf{H} :

$$\zeta_{CA} = (\mathbf{C} \circ \mathbf{A}) \quad (\text{A.6})$$

$$\mathbf{H} = \mathbf{Z}^{(F \times IK)} \zeta_{CA} \left(\zeta_{CA}^T \zeta_{CA} \right)^{-1}. \quad (\text{A.7})$$

Update of \mathbf{A} :

$$\zeta_{HC} = (\mathbf{H} \circ \mathbf{C}) \quad (\text{A.8})$$

$$\mathbf{A} = \mathbf{Z}^{(I \times FK)} \zeta_{HC} \left(\zeta_{HC}^T \zeta_{HC} \right)^{-1}. \quad (\text{A.9})$$

Update of \mathbf{W} using \mathbf{P} from Eq. (2) and using \mathbf{Z} unfolded to a $K \times IF$ matrix:

$$\mathbf{P} = (\mathbf{A} \circ \mathbf{H}) \quad (\text{A.10})$$

$$\mathbf{U} = \left[\sqrt{1 - \alpha} \mathbf{R} \mid \sqrt{\alpha} \mathbf{P} \right] \quad (\text{A.11})$$

$$\mathbf{V} = \left[\sqrt{1 - \alpha} \mathbf{Y} \mid \sqrt{\alpha} \mathbf{Z}^{(K \times IF)} \right] \quad (\text{A.12})$$

$$\mathbf{W} = \mathbf{Z}^+ \mathbf{V} \mathbf{U}^+. \quad (\text{A.13})$$

Update of \mathbf{C} :

$$\mathbf{C} = \mathbf{Z}^{(K \times IF)} \mathbf{W}. \quad (\text{A.14})$$

Update of \mathbf{R} :

$$\mathbf{R} = \mathbf{C}^+ \mathbf{Y}. \quad (\text{A.15})$$

Repeat steps 2–4 until convergence.

Appendix B

The steps for making predictions on new samples based on the trained model are described below.

1. Retrieve the values of \mathbf{A} , \mathbf{H} , \mathbf{W} and \mathbf{R} from the trained model.
2. Initialize \mathbf{C}_{new} .
3. Estimate $\mathbf{Q}_{k, \text{new}}$ based on \mathbf{X}_{new} , \mathbf{A} , \mathbf{H} , and \mathbf{C}_{new} ; for every $k = 1, \dots, K$:

$$\mathbf{S}_{k, \text{new}} = \mathbf{X}_{k, \text{new}}^T \mathbf{A} \mathbf{D}_{k, \text{new}} \mathbf{H}^T \quad (\text{A.16})$$

$$\mathbf{Q}_{k, \text{new}} = \mathbf{S}_{k, \text{new}} \left(\mathbf{S}_{k, \text{new}}^T \mathbf{S}_{k, \text{new}} \right)^{-0.5}. \quad (\text{A.17})$$

Build the three-way array \mathbf{Z}_{new} :

$$\mathbf{Z}_{k, \text{new}} = \mathbf{X}_{k, \text{new}} \mathbf{Q}_{k, \text{new}} \quad \text{for } k = 1 : K. \quad (\text{A.18})$$

Update \mathbf{C}_{new} using the weights \mathbf{W} :

$$\mathbf{C}_{\text{new}} = \mathbf{Z}_{\text{new}}^{(K \times IF)} \mathbf{W}. \quad (\text{A.19})$$

Repeat steps 3–5 until convergence.

Predict the value of the dependent variable(s) according to:

$$\hat{\mathbf{Y}}_{\text{new}} = \mathbf{C}_{\text{new}} \mathbf{R}. \quad (\text{A.20})$$

References

- [1] R.A. Harshman, Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-modal factor analysis, UCLA Work. Pap. Phon. 16 (1970) 1–84.
- [2] R. Bro, PARAFAC. Tutorial and applications, Chemometr. Intell. Lab. Syst. 38 (1997) 149–171.
- [3] R. Bro, Multiway calibration. Multilinear PLS, J. Chemometr. 10 (1996) 47–61.
- [4] R.A. Harshman, PARAFAC2: mathematical and technical notes, UCLA Work. Pap. Phon. 22 (1972) 30–47.
- [5] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2 – part II. Modeling chromatographic data with retention time shifts, J. Chemometr. 13 (1999) 295–309.
- [6] T. Skov, R. Bro, A new approach for modelling sensor based data, Sensors Actuators B 106 (2005) 719–729.
- [7] J.M. Amigo, T. Skov, J. Coello, S. Maspocho, R. Bro, Solving GC–MS problems with PARAFAC2, Trends Anal. Chem. 27 (2008) 714–725.
- [8] K. Wiberg, S.P. Jacobsson, Parallel factor analysis of HPLC-DAD data for binary mixtures of lidocaine and prilocaine with different levels of chromatographic separation, Anal. Chim. Acta. 514 (2004) 203–209.
- [9] B.M. Wise, N.B. Gallagher, E.B. Martin, Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch, J. Chemometr. 15 (2001) 285–298.
- [10] S. de Jong, H.A.L. Kiers, Principal covariates regression. Part 1. Theory, Chemometr. Intell. Lab. Syst. 14 (1992) 155–164.
- [11] A.K. Smilde, H.A.L. Kiers, Multiway covariates regression models, J. Chemometr. 13 (1999) 31–48.
- [12] H.A.L. Kiers, J.M.F. ten Berge, R. Bro, PARAFAC2 – part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemometr. 13 (1999) 275–294.
- [13] R. Bro, Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications, (Ph.D. thesis) University of Amsterdam (NL), 1998, [Available at <http://www.models.life.ku.dk/research/theses/> (Last accessed July 13th 2013)].
- [14] F. Marini, A. D'Aloise, R. Bucci, F. Buiarelli, A.D. Magri, A.L. Magri, Fast analysis of 4 phenolic acids in olive oil by HPLC-DAD and chemometrics, Chemometr. Intell. Lab. Syst. 106 (2011) 142–149.
- [15] T. Skov, D. Ballabio, R. Bro, Multiblock variance partitioning. A new approach for comparing variation in multiple data blocks, Anal. Chim. Acta. 615 (2008) 18–29.
- [16] R.D. Sneer, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–418.
- [17] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, Chemometr. Intell. Lab. Syst. 46 (1999) 133–147.
- [18] G. Tomasi, F. Savorani, S.B. Engelsen, iCoshift: an effective tool for the alignment of chromatographic data, J. Chromatogr. A 1218 (2011) 7832–7840.
- [19] J.C. Gower, Generalized procrustes analysis, Psychometrika 40 (1975) 33–51.
- [20] H.A.L. Kiers, An alternating least squares algorithm for PARAFAC2 and three-way DEDICOM, Comput. Stat. Data Anal. 16 (1993) 103–118.
- [21] M. Vervloet, K. Van Deun, W. Van den Noortgate, E. Ceulemans, On the selection of the weighting parameter value in Principal Covariates Regression, Chemometr. Intell. Lab. Syst. 123 (2013) 36–43.