

Automated resolution of overlapping peaks in chromatographic data

Lea G. Johnsen^{a,b*}, José Manuel Amigo^a, Thomas Skov^a and Rasmus Bro^a



Parallel factor analysis 2 (PARAFAC2) has been shown to be a powerful tool for resolution of complex overlapping peaks in chromatographic analyses. It is particularly useful because of its ability to handle shifts in the elution time mode and peak shape changes. Like all curve resolution techniques, PARAFAC2 will only find chemically meaningful parameters (elution time profiles and mass spectra) if the correct number of factors are determined. So far, the primary way to determine an appropriate number of factors, when using PARAFAC2, is to calculate models with different number of factors and then inspect the models manually. This approach is time consuming, and the result may be biased because of the manual assessment of the model quality, making PARAFAC2 inaccessible for analytical chemists in general. Here, we develop a method that can determine an appropriate number of factors in an automated way. The automation is based on a number of model diagnostics (quality criteria) collected from models with different numbers of factors. Combining these diagnostics, it is possible to assess what the appropriate number of components is. In this work, only gas chromatography–mass spectrometry data are considered. However, it will most likely be fairly straightforward to expand the work to also cover liquid chromatography data (with a multivariate detector). Automating the model quality evaluation of the PARAFAC2 model enables both the inexperienced and trained user to perform comprehensive and advanced analysis of chromatographic data with a minimum of manual work. Copyright © 2013 John Wiley & Sons, Ltd.

Additional supporting information may be found in the online version of this article at the publisher's web site.

Keywords: chromatography; PARAFAC2; automated; GC–MS

1. INTRODUCTION

Nowadays, the most widespread approach for chromatographic data analysis is using the software provided by the manufacturer of the instrument. However, it has been shown that the algorithms implemented in many commercial packages can result in suboptimal utilization of the information in the data, compared to what can be provided by curve resolution techniques such as multivariate curve resolution or parallel factor analysis (PARAFAC or PARAFAC2) [1,2]. A quite typical example is shown in Figure 1, where the deconvolution procedure from commonly used manufacturer's software has been applied to the data from 45 samples. The peaks in the total ion chromatogram (TIC) (left side of the figure) seem to be divided into two groups, indicating that the data represent two different chemical compounds. If the individual mass traces (not shown) are inspected, it is clearly seen that there are indeed two chemical compounds. The result from the manufacturer's software is shown in the middle plot in Figure 1. Several problems can be observed with this result. First of all, the software only finds a total of six peaks in all samples; this means that many of the peaks from the raw data are not described by this model. Furthermore, several of the peaks contain a shoulder (circle in Figure 1), indicating that the resolution of the two co-eluting peaks has not been successful. A third problem is that none of the peaks in the model have been separated from the baseline and that the level of the included baseline is fluctuating. On the other hand, by using PARAFAC2 on the dataset, these problems are seemingly solved.

The PARAFAC2 model is able to split the peak into three main contributions arising from the two chemical compounds and signal from the baseline (rightmost plot in Figure 1).

PARAFAC2 is a method able to handle many chromatographic artifacts (e.g., baseline drift, overlapping, and elution time shifts) [3]. It allows separating each source of variability in the data by using the spectral information gathered for each elution time. The resulting model provides three important sets of parameters: an estimated elution time profile for each compound in each sample, an estimated pure spectrum for each compound, and relative concentrations for each of these chemical compounds [4].

Previous publications [5–7] have shown that the use of PARAFAC2 enables a comprehensive analysis of chromatographic data. However, a common practical issue in all curve resolution techniques is that the correct number of factors must be determined in order to obtain chemically meaningful profiles [8]. Unfortunately, it is not straightforward to determine the proper number of factors to include. This may be one reason why curve resolution methods are not usually seen in routine chemical analysis. Therefore, an automated selection of the appropriate PARAFAC2 model would provide a significant improvement in terms of allowing non-chemometrically skilled chemists to take advantage of modern data analysis solutions.

* Correspondence to: Lea G. Johnsen, Quality & Technology, Department of Food Science, Faculty of Science, University of Copenhagen, Copenhagen, Denmark. E-mail: dklgj@chr-hansen.com

^a L. G. Johnsen, T. Skov, R. Bro, Quality & Technology, Department of Food Science, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

^b L. G. Johnsen
Chr. Hansen A/S, Bøge Alle 10-12, 2970 Hørsholm, Denmark

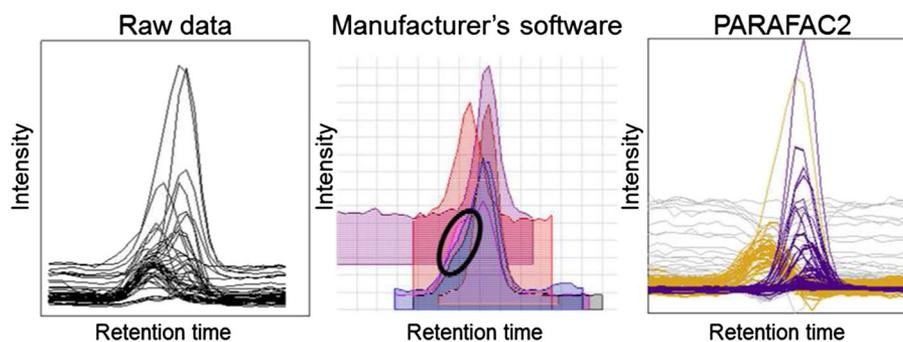


Figure 1. Left: TIC of 45 elution profiles. The example shows the performance of, respectively, manufacturer's software deconvolution algorithm (middle) and PARAFAC2 (right) on the raw data (illustrated with the TIC, leftmost of the figure). The manufacturer's software finds a total of six compounds in all the 45 samples (illustrated with six elution profiles in the figure). The PARAFAC2 model finds three components in each sample (illustrated with estimated elution profiles to the right): one of these describes baseline (gray) and two describe two different chemical compounds (orange and purple).

When a PARAFAC2 model is calculated, several statistical and empirical diagnostics can be used to evaluate the reliability of the model. Traditionally, only explained variance, residuals, and mere observation of the obtained elution time and spectral profiles have been used when determining the appropriate number of factors in PARAFAC2. It has recently been suggested to use core consistency, which is a diagnostic tool indicating whether the model is over-fitted [9], in the evaluation of PARAFAC2 models [10]. Nevertheless, there exist a number of additional diagnostics that can be used to evaluate the obtained PARAFAC2 model. In this manuscript, we suggest 102 different diagnostics that all aim to describe some aspect of the quality of a PARAFAC2 model.

From these initial statistical and empirical diagnostics, we determine a classification model, which, in an automated way, can find an appropriate number of factors to include in the PARAFAC2 model. In order to obtain a good classification model, we determine which of the initial descriptive quality criteria are most important for the classification, and only those will be used in the final classification model. The proposed method is tested on four different gas chromatography–mass spectrometry (GC–MS) datasets originating from different chromatographic instruments and from different sample matrices (apples, wine, aroma standards, and cheese).

The manuscript will be initiated by a section describing the theory behind PARAFAC2 and a brief description of the diagnostics developed for assessing PARAFAC2 model complexity. The theory behind the classification will not be described in this paper but can be found elsewhere in the literature [11,12]. The description of the diagnostics will be followed by a section that describes how the classification model is optimized and a validation of the final classification model.

Throughout this paper, the word compound refers to the chemical compounds contained in a sample, while the words factor and component refer to the outcomes of the model.

2. THEORY

PARAFAC was introduced simultaneously by Harshman [8] and Carroll and Chang [13] (who named it canonical decomposition). Harshman based his PARAFAC model on the idea of parallel proportional profiles by Cattell [14]. The idea behind parallel proportional profiles is that if the relative amounts of overlapping

phenomena are changing across samples, then it is possible to resolve the unique patterns for each of these phenomena.

The PARAFAC solution is uniquely identified (up to scaling and permutations) under mild conditions [15]. Essentially, PARAFAC is unique as long as the chemical compounds do not have identical spectra or identical elution profiles. This indirectly implies that a correctly specified PARAFAC model, applied to, for example, GC–MS data, can provide estimates of the pure mass spectra, concentration profiles, and pure elution time profiles when there are no elution time shifts (as illustrated in Figure 2).

Chromatographic data often contain shift in the elution time dimension, and PARAFAC is unable to handle shifted data efficiently without the data being aligned prior to modeling with, for example, correlation optimized warping as described by Skov *et al.* [1], as only one common elution time profile is estimated for each compound. Hence, PARAFAC assumes that the *same* elution profile can model the compound in all samples. However, it has been shown that PARAFAC2 can be used to solve the problem with shifts in retention time [4,16]. In PARAFAC2, an elution time profile is found for each compound in each sample as illustrated in Figure 2. As for PARAFAC, the PARAFAC2 solutions are also unique. A thorough description of PARAFAC2 has been made by Bro *et al.* [4,15].

For both PARAFAC and PARAFAC2, a model with too many factors will describe a variation that is not chemically meaningful. Sometimes, one or a few extra components do not disturb the model. For example, for PARAFAC, it has been shown that in some cases, more than one model can be chemically meaningful and provide estimates of the underlying patterns even though the models have different numbers of factors [8]. This implies that in some cases, there is no *one* optimal model but rather a range of appropriate models. In our experience, the same goes for PARAFAC2.

Figure 3 shows two examples of under-fitted models obtained from the data presented in Figure 1. Both residuals and scaled elution profiles indicate that the models do not have enough factors included. The residuals are, especially in the one-factor model, behaving in a very systematic manner. This indicates that there is still chemical information in the data, which is not explained by the model. In the elution profiles, the under-fitting is indicated by the modeled peak shapes, which suggest that two (or more) co-eluting compounds are described by the same elution profile.

In Figure 4, a good model is shown to the left and an over-fitted model to the right. Both models show how the residuals

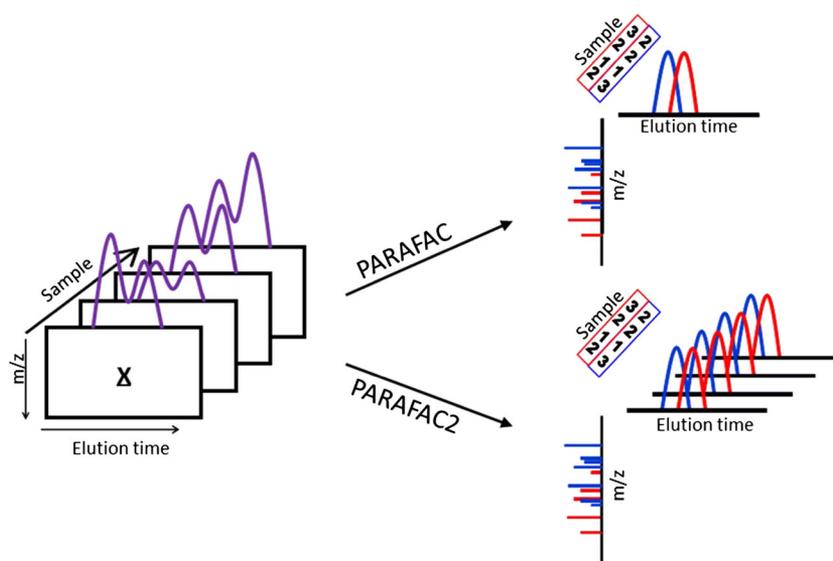


Figure 2. Illustration of the differences between PARAFAC and PARAFAC2. In PARAFAC, one common elution profile is used to describe each compound in all samples, whereas in PARAFAC2, each compound in each sample is modeled with a distinct elution profile. Figure adapted with permission from Amigo *et al.* [3]. Copyright © 2010 American Chemical Society.

become smaller and more unsystematic when a sufficient number of factors are included. The over-fitted model, in this example, is characterized by having low core consistency (below zero) as well as an increase in the proportion of negative values in the obtained mass spectra, compared to the model with one less factor. Also, the number of iterations, which has increased considerably, compared to the model with one less factor, indirectly indicates that this model is over-fitted.

One disadvantage in the aforementioned evaluation of obtained models is that all of the diagnostics are subjective, and such evaluations will inevitably be biased by different personal opinions on which of the diagnostics one thinks is more important in the evaluation. Another disadvantage is that the evaluation is a time-consuming task. To counter these problems, an automated quality control is proposed by investigating 34 diagnostics (see supplementary material), as descriptors for quality of the PARAFAC2 model. All the diagnostics are calculated on models where the problem with sign indeterminacy has been solved with the method proposed by Bro *et al.* [17].

For each of the 34 diagnostics, also the difference (“diff”) between the diagnostic value of the present model and the model with one factor less was included (diagnostics #35–68). This approach was inspired by the DIFFIT approach described by Timmerman and Kiers [18]. The diff value for a model with k factors is defined as the value at k factors minus the value at $k - 1$ factors. Also, the relative change is determined for all 34 diagnostics (diagnostics #69–102). The relative change is defined as the diff divided by the original value. This gives a total of 102 diagnostics.

3. MATERIALS AND METHODS

3.1. Datasets

A total of four different GC–MS datasets have been included in this study:

- analysis of apples (apple juice analyzed with headspace–GC–MS (HS–GC–MS) using a Hewlett Packard G1800A GC/MS

system. More details can be found in the publication by Amigo *et al.* [6],

- analysis of wine (HS–GC–MS analysis carried out on an HP 6890 GC coupled with an Agilent 5973 Mass Selective Detector. More details can be found in the publication by Ballabio *et al.* [19]),
- analysis of standards with different additions of aroma compounds [1] (HS–GC–MS analysis carried out on a Hewlett-Packard G1800A GC–MS system),
- analysis of cheese derivatized with methoxamine (20 mg/mL in pyridine) followed by derivatization with *N*-methyl-*N*-(trimethylsilyl) trifluoroacetamide as described by Kanani *et al.* [20]. The derivatized cheeses were analyzed using an Agilent technologies 7890A GC system coupled to an Agilent technologies 5975C inert XL MSD with Triple-Axis detector.

As curve resolution techniques work in small and regional intervals, the four datasets were divided into intervals with an estimated maximum of five compounds in each interval. The TICs from the four datasets are shown in Figure 5, with blowups F5 showing examples of these intervals. In total, 155 intervals were created. PARAFAC2 models with one to seven factors were calculated for each of these intervals, and the diagnostics described in the supplementary material were determined for each of these 1085 models. By working on baseline-separated intervals, the problems are typically much easier to analyze, and more unambiguous results are obtained. PARAFAC2 models were calculated without any constraints (such as non-negativity) as these constraints might disguise indications of over-fit.

The ‘correct’ number of factors for each model was determined by having a skilled GC–MS user evaluating the models by assessing not only the core consistencies, number of iteration used, residuals, and obtained elution and spectral profile (as described in the preceding section) but also the chemical information in the raw data.

The classification model was constructed with partial least squares discriminant analysis (PLS-DA) as follows: The 155 intervals were randomly divided into a calibration (75 intervals) and a validation (80 intervals) set. The PLS-DA model was optimized,

Under-fitted models

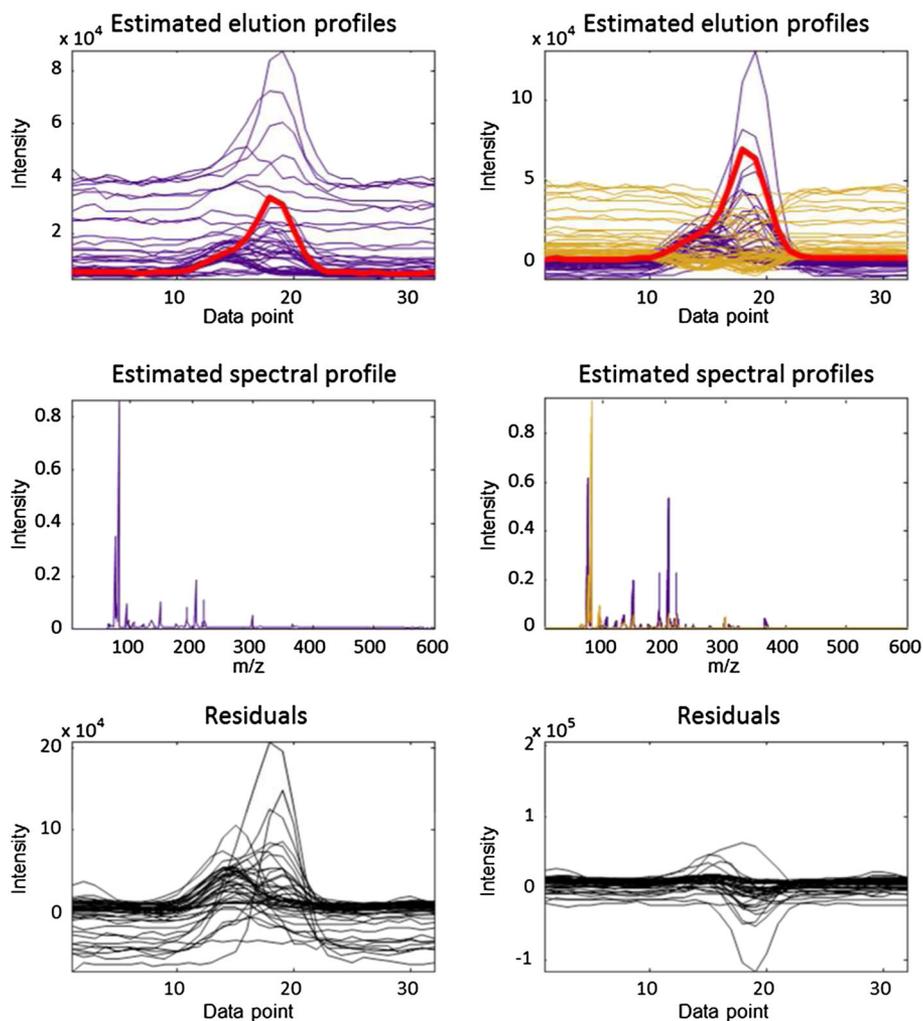


Figure 3. PARAFAC2 models with too few factors applied to the data presented in Figure 1. Left, one factor; core consistency, 100; iterations, 2. Right, two factors; core consistency, 99; iterations, 2. Core consistency and iterations are used in the evaluation of the models. The colors in the figure refer to the different components; for example, purple in elution and spectral profile refers to the same component. The thick red lines illustrate elution profiles referred to in the text.

and the number of variables (diagnostics) included reduced by using the calibration set with 75 intervals and the 102 diagnostics, and subsequently tested using the test set containing the rest of the intervals and the remaining diagnostics. Eleven of the intervals were not modeled well by PARAFAC2. The reason for that could be that two compounds were totally overlapping (embedded peaks), the peak shapes were severely changed, or that the spectra of different compounds were too similar. These intervals were removed from the calibration set before any variable selection was performed in order for them not to influence the selection.

3.2. Software

All algorithms and models have been developed using MATLAB R2012a (Mathworks, Inc., Natick, MA, USA). PLS_Toolbox (Eigenvector Research, Inc., Wenatchee, WA, USA) has been used for principal component analysis and PLS-DA models. PARAFAC2 models have been calculated using the algorithm available from www.models.life.ku.dk (April 2013).

4. RESULTS AND DISCUSSION

By visual inspection of the raw diagnostics, it became obvious that most of them were following an overall pattern. In Figure 6, the values of core consistency (diagnostic 20), the number of iterations (diagnostic 2), and the relation between the absolute area and the area of the spectral profile (diagnostic 10) are shown. The values are obtained from the PARAFAC2 models illustrated in Figures 3 and 4.

Small changes in the diagnostic value are observed as more and more factors are included until the model becomes over-fitted. At this point, there is a significant change after which the diagnostic value only changes slowly again. The relative difference obtained from the first over-fitted model describes this jump in the raw diagnostic values, as illustrated in Figure 6. Therefore, the first over-fitted model will be distinguishable from the ones with fewer components. This approach is similar to what was described by Hoggard and Synovec [21] in their paper concerning automated determination of the number of factors to include in PARAFAC models.

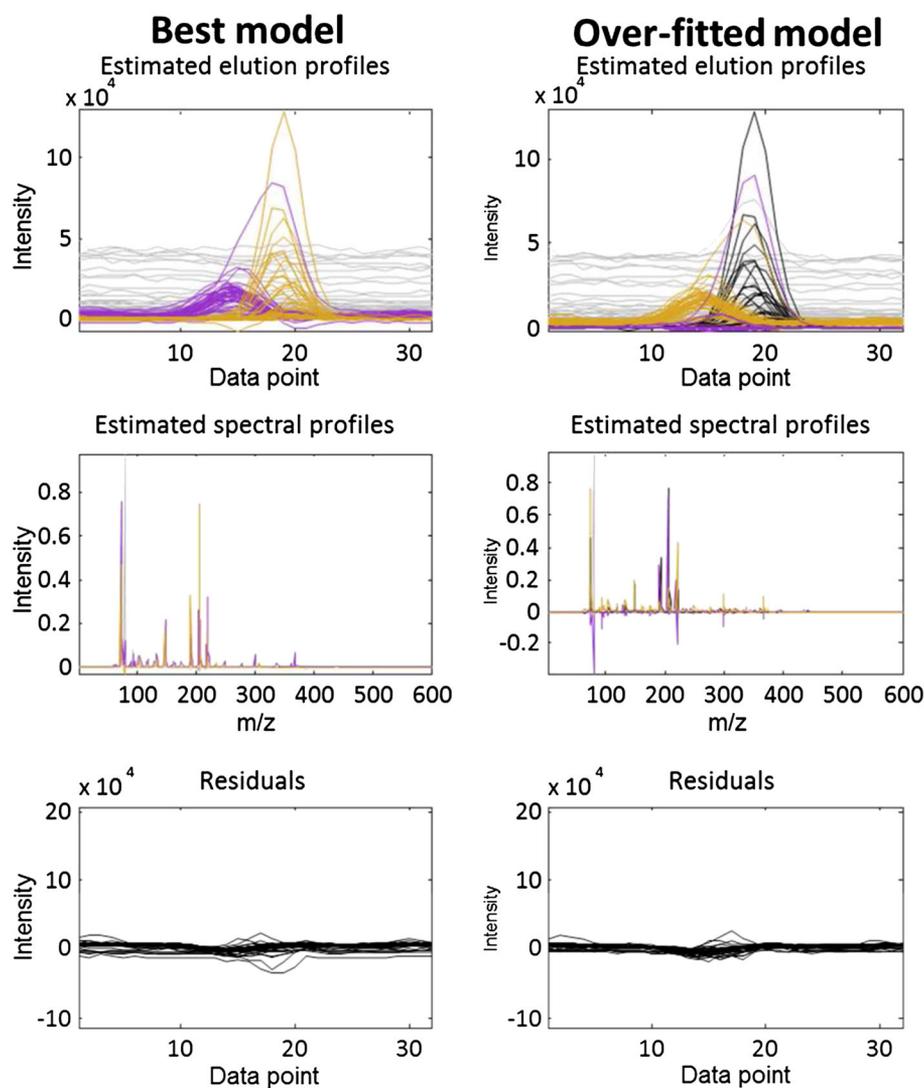


Figure 4. Best model: core consistency, 97; iterations, 30. Over-fitted model: core consistency, <0; iterations, 383. Core consistency and iterations are used in the evaluation of the models. The colors in the figure refer to the different components; for example, purple in elution and spectral profile refers to the same component.

In order to determine which of the 102 selected diagnostics were useful in classification of the first over-fitted model, variable selection was conducted. By using a combination of three different variable selection techniques, the most important diagnostics were selected. The variable selection techniques used were VIP (which indicates the importance of a variable for the model), iPLS (which identifies a few number of variables that give better predictions than the other), and genetic algorithms (which selects the variable combination that gives the best model). All of these techniques are thoroughly described by Andersen and Bro [22].

The following seven diagnostics represent a good compromise between having a good classification power without too many variables (the numbers in parentheses refer to the table in the supplementary material, which includes all 102 diagnostics):

- core consistency (#20),
- change in the negative area in the elution profile (#43),
- change in negative correlation between spectra (#53),
- logarithm of the number of iterations (#3),
- the positive correlation between spectra (#18),

- the relative change in how systematic the residuals are (indicated with Durbin Watson) (#97),
- the relative change in the correlation between the TIC from raw data and the TIC from the obtained model (#94).

The specificity (0.97 for over-fit and 0.95 for not over-fit) obtained with a model including these seven diagnostics is very similar to the model including five additional diagnostics.

None of the seven final diagnostics could be removed without a significant loss of classification power (decrease in specificity close to 10%).

The regression vector obtained in the PLS-DA model (Figure 7) F7 indicated that the core consistency, changes in the negative area in the elution profile, and changes in negative correlation between spectra are negatively correlated to over-fit. On the other hand, the following diagnostics were positively correlated with the first over-fitted model: the logarithm of the number of iterations, the positive correlation between spectra, the relative change in how systematic the residuals are (indicated with Durbin Watson), and the relative change in the correlation between the TIC from raw data and the reconstructed TIC from the obtained model.

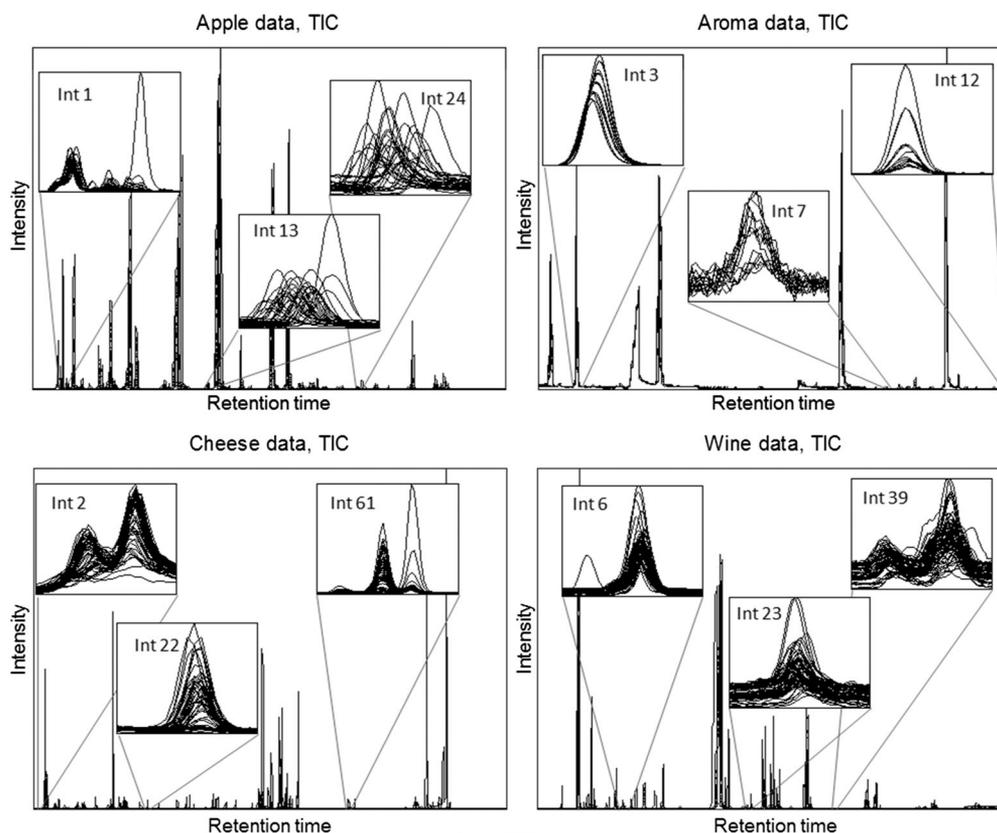


Figure 5. Overview of the raw data; the blowups show examples of different intervals.

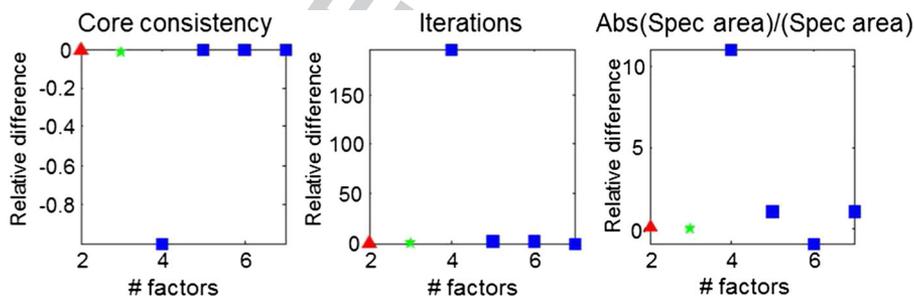


Figure 6. Development of a typical diagnostic with increasing number of factors included. Illustrated here are three of the 34 diagnostics listed in the supplementary material. The diagnostics are obtained from the PARAFAC2 models illustrated in Figures 3 and 4. Red triangles: under-fitted models. Green stars: best model. Blue squares: over-fitted model.

The PLS-DA model based on these seven diagnostics was used to classify the models in the test set (a total of 80 intervals). Most of these intervals were correctly classified according to the manual evaluation. Furthermore, it was found that the majority of the cases where there was incongruence between manual and automated evaluation were either models that were misclassified by the manual evaluation or cases where PARAFAC2 was simply unable to describe the data in an appropriate way.

As an example of the wide application range of the presented method, the elution profiles from three over-fitted models, with very different characteristics, are shown at the right of Figure 8. These were all correctly classified as the first over-fitted models, leading to the models with one factor less, shown at the left of the figure, to be determined as being appropriate models. The diversity in these models indicates that the classification is performing well for many different types of GC-MS data.

In Table I, the cases of incongruence between manual and automated evaluation are listed. These cases can be divided into three categories; misclassified (eight models), intervals not well described by any PARAFAC2 model (10 models), and models where both the results of the manual and automated classifications can be assessed as correct (nine models). In the table, it can be seen which of these categories the individual models falls into. In the following, a thorough inspection of the three types will be given.

4.1. Intervals not well described by PARAFAC2

These are cases where it is not possible to obtain a model that looks meaningful from a chemical point of view. An example of this is interval 2 from the cheese data (Figure 9). In the model with two factors (left side), the two peaks have

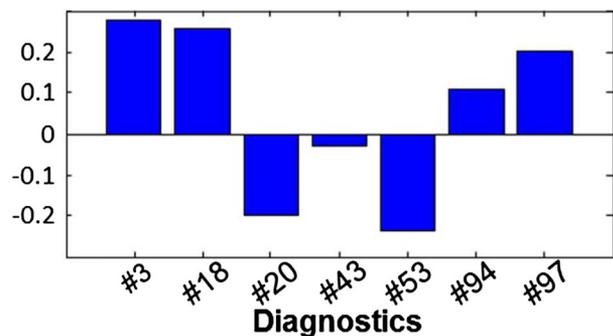


Figure 7. Regression vector for the PLS-DA model.

not been separated. Furthermore, the factor that describes the baseline contains a number of small peaks. This indicates that this component is actually describing baseline plus at least one chemical compound. These observations indicate

that additional factors should be included. In the model with three factors (right side), there is still one elution profile that describes both of the peaks. Additionally, the spectra from component 2 (green profiles) are exclusively negative, which indicate that the model is over-fitted. In models with more than three factors included, the signs of over-fit become even more obvious. When inspecting the mass spectra of the two peaks (not shown), it can be seen that the two compounds have many fragments in common but that they also have some differences. In this case, we must therefore conclude that this interval cannot be well described with PARAFAC2. Imposing non-negativity in the estimated spectral profile and concentrations did not improve the models. Note that traditional chromatographic software would also not be able to provide meaningful results of this interval.

It might be possible to identify infeasible intervals by including the correct diagnostics. The development of such a method is beyond the scope of this paper but should be investigated further in future work.

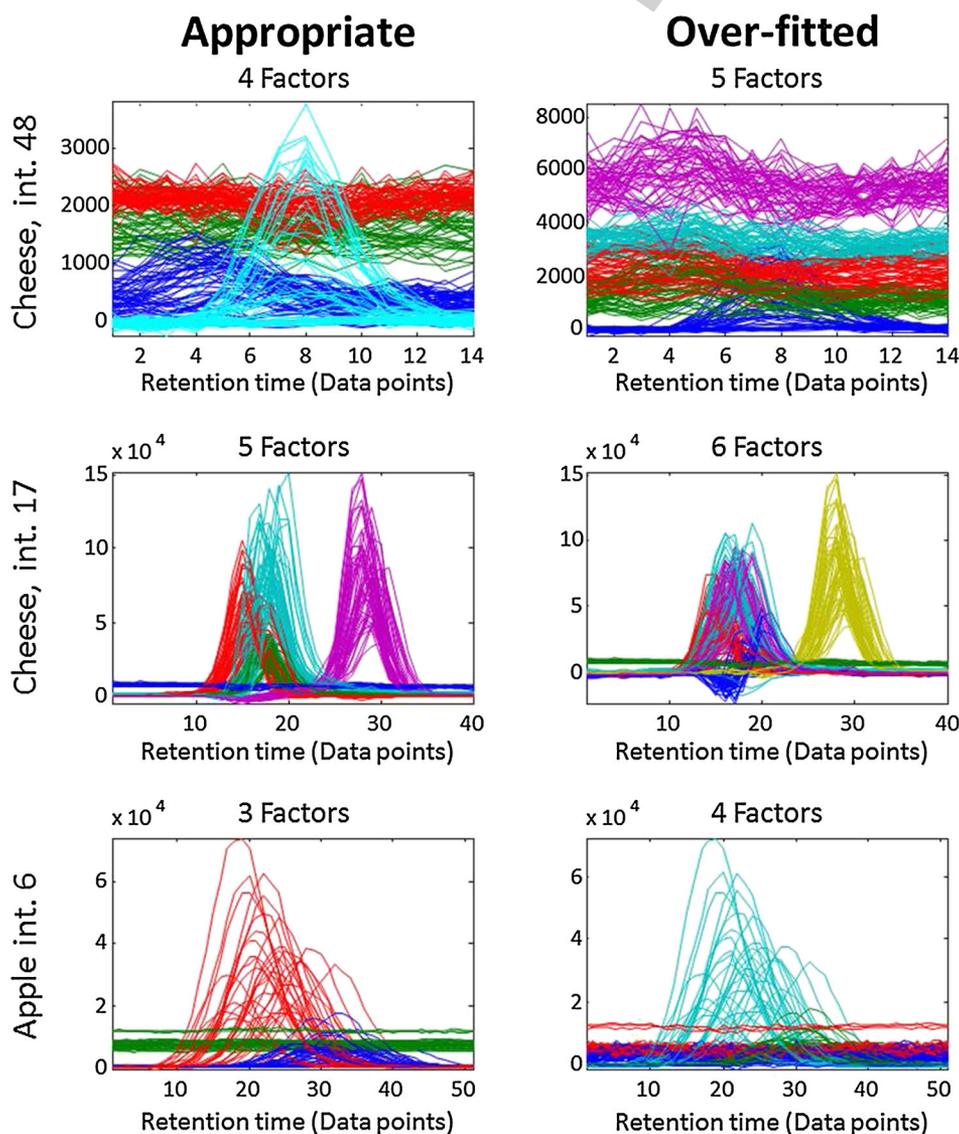


Figure 8. Right: examples of models correctly classified as being the first over-fitted model by the PLS-DA model. Left: the models with one factor less, concluded to be the appropriate models. The colors in the figure refer to the different components.

Table I. Cases of incongruence between manual and automated evaluation from the test set with a total of 80 intervals

	Dataset	Interval	# Factors	Comments
Predicted as not over-fit by the automated approach	Apple	19	3	Also appropriate
		22	4	PARAFAC2 inappropriate
		24	3	Also appropriate
	Cheese	2	2	PARAFAC2 inappropriate
		22	3	Also appropriate
		38	3	PARAFAC2 inappropriate
		31	3	Also appropriate
	Aroma	46	3	Also appropriate
		1	2	Also appropriate
		5	2	PARAFAC2 inappropriate
7		2	Also appropriate	
8		2	Also appropriate	
Predicted as over-fit by the automated approach as over-fit	Apple	11	3	PARAFAC2 inappropriate
		1	5	Misclassified
		13	3	Misclassified
	Cheese	22	4–5	PARAFAC2 inappropriate
		5	4	Misclassified
		10	3	Misclassified
		12	4–6	Misclassified
		14	4	PARAFAC2 inappropriate
		16	2–3	PARAFAC2 inappropriate
		20	3	Misclassified
58	3	PARAFAC2 inappropriate		
Wine	65	2	Misclassified	
	30	4	Misclassified	
	38	3	Also appropriate	
		41	2–4	PARAFAC2 inappropriate

4.2. Intervals offering several solutions

In cases where the classification might be correct even though it differs from the manual classification, there exists more than one number of factors for which it could be argued that the model is appropriate. An example of this could be interval 24 from the apple data. In Figure 10, the PARAFAC2 models with two and three factors, respectively, are illustrated. In the model with two factors (left side), there is still a small amount of systematic behavior in the residuals, indicating that an additional factor should be included in the model. However, if an additional factor is included (right side), there is an increase in negative values in the spectral profile, indicating that too many factors have been included. By constraining the model with non-negativity in spectral profiles and concentration, the model with three factors no longer has indications of being over-fitted. This indicates that the model with three factors indeed is not over-fitted as also suggested by the classification model. A more thorough investigation was conducted by calculating new models on the same interval with different starting points. These new models indicated that the obtained model is representing a local minimum.

Another example of a model with seemingly conflicting results is from the cheese dataset, interval 22. Upon thorough inspection of the model and the raw data, there are indications that this three-factor model may have been correctly classified as not over-fitted and that it is the initial manual assessment that is wrong. The spectral profile obtained by the model (middle

plot, Figure 11) shows that there are some masses that separate

the two peaks. If these masses are inspected in the raw data (leftmost plot in Figure 11), it can be seen that there actually seems to be a shift, indicating that two chemical compounds are eluting. However, the two compounds are poorly resolved (as also shown by the modeled elution time profiles; leftmost plot in Figure 11), and this might be the reason that PARAFAC2 has some difficulties in modeling them. It is interesting that apparently, the results of a properly trained expert system can outperform a skilled analyst.

4.3. Misclassified

The incorrectly predicted models can be divided into two sub-categories: misclassified as not over-fitted and misclassified as over-fitted. However, there are no models that fall into the first category, indicating that the actual specificity of the classification model is very high. In the second category, there are eight models.

In the cheese dataset, interval 5, there is a small amount of tailing, which is not caught in the model identified as the correct one. However, neither the obtained spectral profile nor the concentration profiles (not shown) are affected by this. From all practical aspects, this will not affect any further interpretation of data. In the apple dataset, interval 13, a small peak is missing in the model identified as being appropriate. However, this peak is of such small magnitude that it does not affect the obtained concentration profiles and mass spectra, which are very similar in the two models. The same situation is observed for cheese,

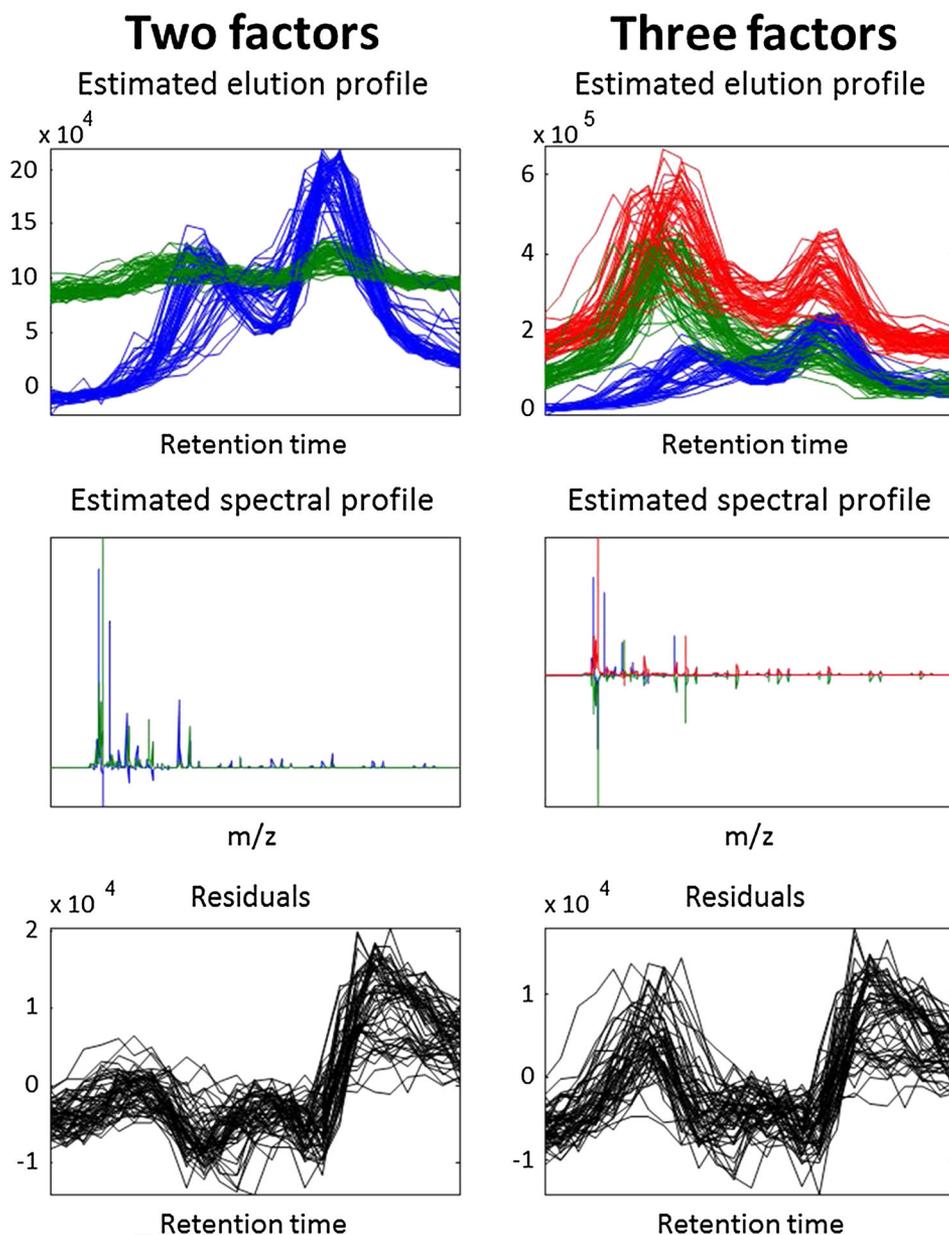


Figure 9. Interval 2 from the cheese data modeled with two and three factors, respectively. The model with two factors seems to have too few factors, whereas the model with three seems to have too many factors included. The colors in the figure refer to the different components; for example, blue in elution and spectral profile refers to the same component.

interval 20. In this case, the missing peak has a higher magnitude, but nevertheless, the compounds are modeled meaningfully and very similar to those in the appropriate model (illustrated in Figure 12). Furthermore, both spectra and concentration profiles (not shown) of the main peak (which is present in both models) are practically unaffected. This means that further interpretation will not be affected besides the fact that a compound is missed.

The misclassifications of the remaining five intervals (apple, interval 1; cheese, intervals 10, 12, and 65; wine, interval 30) result in models that deviate more severely from the appropriate model.

Summing up, these observations actually show that over 90% of the models that are identified as correct with the suggested approach are describing real underlying chemical information (assuming that such a model can be found using PARAFAC2). This is to our best belief a significant improvement compared

to the alternatives offered by the manufacturer's software. We have shown that the suggested automated approach can indeed be used to identify models that are describing the underlying chemistry. It may be interesting to compare automatically determined PARAFAC2 models to solutions obtained with alternative methods for quantification of overlapping signals in order to determine which methods are most suitable for different situations.

Additionally, it should be mentioned that the additional computation time needed for calculating the seven diagnostics included in the final PLS-DA model is insignificant compared to the time needed for calculation of the PARAFAC2 models. The step in the automated procedure that is most time consuming is the correction of the sign indeterminacies, and also, this step is considerably less time consuming than the calculation of the PARAFAC2 models.

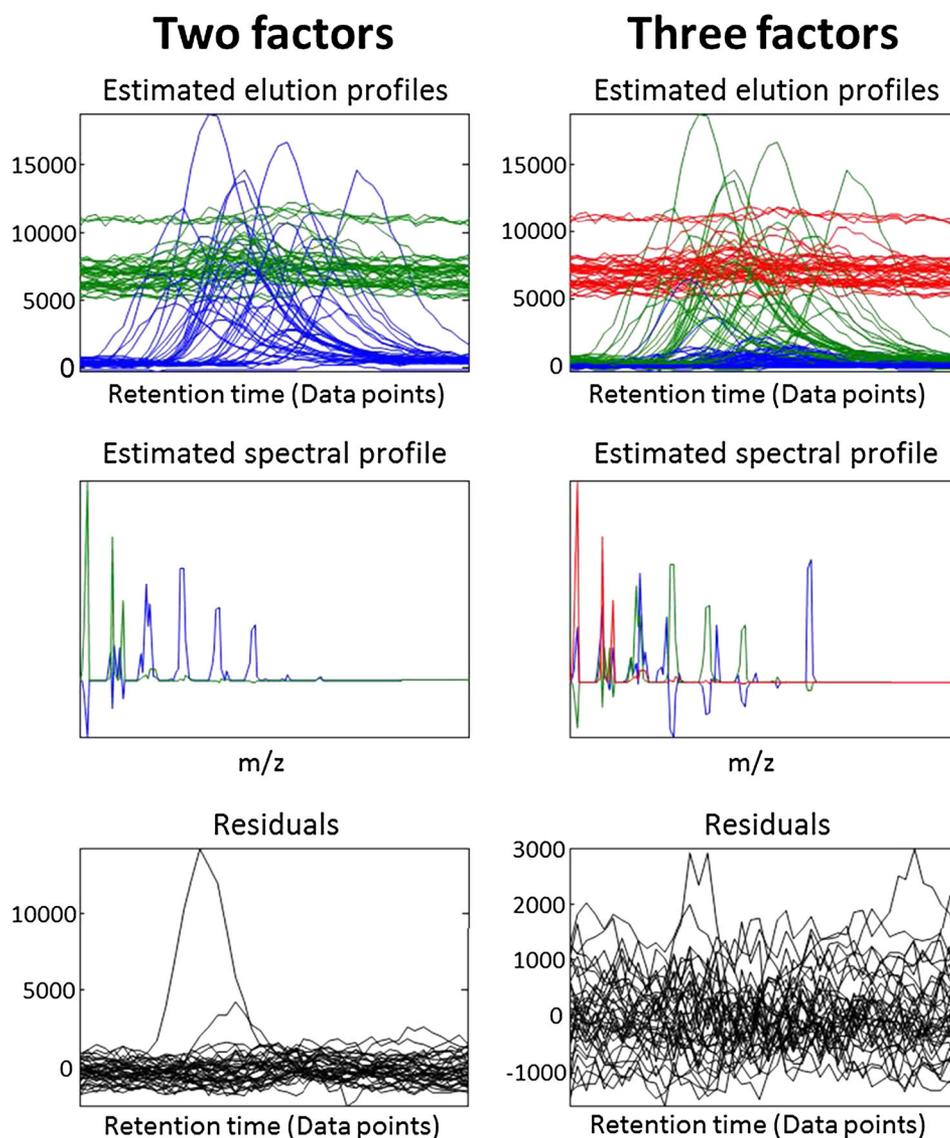


Figure 10. Interval 24 from the apple data modeled with two and three factors, respectively. The model with two factors seems to have too few factors because there is still some systematic behavior in the residuals, but otherwise, it seems like an appropriate model. The model with three factors seems to have too many factors included, because of the increase in negative values in the spectra profile; otherwise, it is a nice model. The colors in the figure refer to the different components; for example, blue in elution and spectral profile refers to the same component.

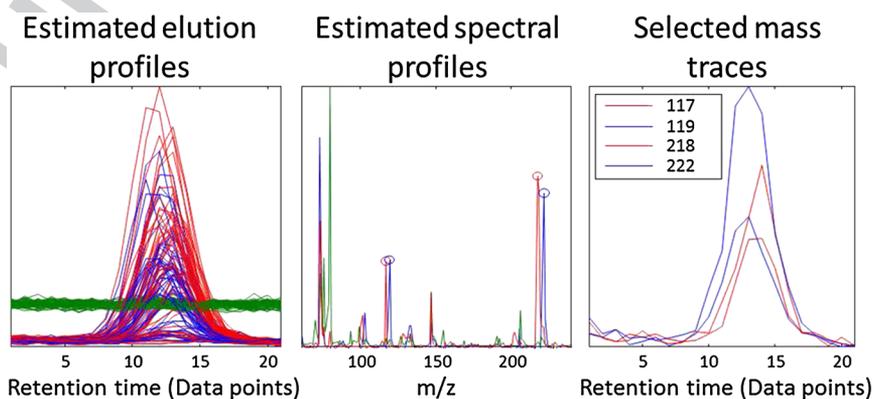


Figure 11. Illustration of the elution time profiles (left) and mass spectral profiles (middle) obtained from cheese, interval 22, modeled with three factors. The rightmost plot shows the mass traces of the masses, which separates the two peaks (indicated with circles in the middle plot). The colors in the figure refer to the different components; for example, blue in elution and spectral profile refers to the same component.

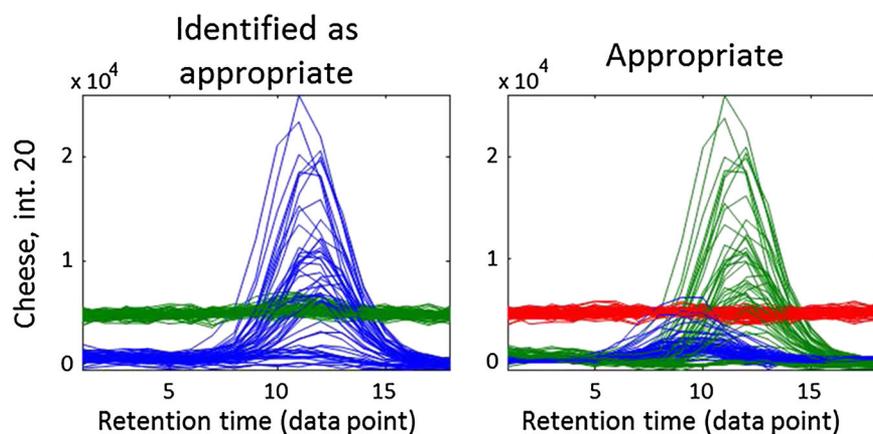


Figure 12. Illustration of the difference in elution time profiles between the model appointed as being appropriate and the actually appropriate model (cheese data, interval 20). The colors in the figure refer to the different components.

5. CONCLUSION

By the usage of different variable selection techniques, we have found that a PLS-DA model based on seven quality criteria can be used to automate selection of the number of components in PARAFAC2. The seven quality criteria that turned out to be most important for the classification are as follows:

- (1) core consistency,
- (2) change in the negative area in the elution profile,
- (3) change in negative correlation between spectra,
- (4) logarithm to the number of iterations,
- (5) the positive correlation between spectra,
- (6) the relative change in how systematic the residuals are (indicated with Durbin Watson),
- (7) the relative change in the correlation between the TIC from raw data and the TIC from the obtained model.

The obtained PLS-DA model was evaluated on an independent test set. This validation showed that over 90% of the models that were determined to be appropriate by the automated procedure describe the underlying chemistry. This makes the approach very useful for handling huge amounts of data, without the need of a skilled chemometrician to manually evaluate every model. It has been shown that the developed method can handle very different kinds of GC-MS data, and optimization should therefore only be necessary if the model is applied on other types of data (e.g., liquid chromatography-MS).

A MATLAB routine has been written, which calculates PARAFAC2 models and finds valid models using the approach described in this paper. The function, which is called AutoChrome, is available at www.models.life.ku.dk (April 2013). If the prediction of the number of compounds is combined with additional tools [2], identification can be accomplished using the open source software OpenChrom [21] combined with the NIST mass spectral library.

REFERENCES

1. Skov T, Bro R. Solving fundamental problems in chromatographic analysis. *Anal. Bioanal. Chem.* 2008; **390**: 281–285.
2. Murphy KR, Wenig P, Parcsi G, Skov T, Stuetz RM. Characterizing odorous emissions using new software for identifying peaks in chemometric models of gas chromatography-mass spectrometry datasets. *Chemom. Intellig. Lab. Syst.* 2012; **118**: 41–50.
3. Amigo JM, Skov T, Bro R. ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics. *Chem. Rev.* 2010; **110**: 4582–4605.
4. Bro R, Andersson CA, Kiers HAL. PARAFAC2—part II. Modeling chromatographic data with retention time shifts. *J. Chemom.* 1999; **13**: 295–309.
5. Amigo JM, Skov T, Bro R, Coello J, Maspocho S. Solving GC-MS problems with PARAFAC2. *Trends Anal. Chem.* 2008; **27**(8): 714–725.
6. Amigo JM, Popielarz MJ, Callejón RM, et al. Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *J. Chromatogr. A* 2010; **1217**(26): 4422–4429.
7. Khakimov B, Amigo JM, Bak S, Engelsen SB. Plant metabolomics: resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *J. Chromatogr. A* 2012; **1266**: 84–94.
8. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an “exploratory” multimodal factor analysis. *AUCLA Working Pap. Phonetics* 1970; **16**: 1–84.
9. Bro R, Kiers HAL. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* 2003; **17**: 274–286.
10. Kamstrup-Nielsen MH, Johnsen LG, Bro R. Core consistency diagnostic in PARAFAC2. *J. Chemom.* 2013; **5**: 99–105.
11. Ballabio D, Todeschini R. Infrared Spectroscopy for Food Quality Analysis and Control. *Academic Press* 2009: 83–102.
12. Ståhle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J. Chemom.* 1987; **1**(3): 185–196.
13. Carroll JD, Chang J. Analysis of individual differences in multidimensional scaling via an n -way generalization of “Eckart-Young” decomposition. *Psychometrika* 1970; **35**(3): 283–319.
14. Cattell RB. “Parallel proportional profiles” and other principles for determining the choice of factors by rotation. *Psychometrika* 1944; **9**(4): 267–283.
15. Sidiropoulos ND, Bro R. On the uniqueness of multilinear decomposition of N -way arrays. *J. Chemom.* 2000; **14** (3): 229–239.
16. Kiers HAL, ten Berge JMF, Bro R. PARAFAC2—part I. A direct fitting algorithm for the PARAFAC2 model. *J. Chemom.* 1999; **13**: 275–294.
17. Bro R, Leardi R, Johnsen LG. Solving the sign-indeterminacy for multi-way models. *J. Chemom.* 2013; **27**: 70–75.
18. Timmerman ME, Kiers HAL. Three-mode principal components analysis: choosing the numbers of components and sensitivity to local optima. *Br. J. Math. Stat. Psychol.* 2000; **53**(1): 1–16.
19. Ballabio D, Skov T, Leardi R, Bro R. Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques. *J. Chemom.* 2008; **22**(8): 457–463.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
20. Kanani H, Chrysanthopoulos PK, Klapa MI. Standardizing GC-MS metabolomics. *J. Chromatogr. B* 2008; **871**(2): 191–201.
21. Hoggard JC, Synovec RE. Parallel factor analysis (PARAFAC) of target analytes in GC×GC-TOFMS data: automated selection of a model with an appropriate number of factors. *Anal. Chem.* 2007; **79**(4): 1611–1619.
22. Andersen CM, Bro R. Variable selection in regression—a tutorial. *J. Chemom.* 2010; **24**(11–12): 728–737.
23. Wenig P, Odermatt J. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. *BMC Bioinf.* 2010; **11**(405).

Q6 66

Q7 67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

UNCORRECTED PROOF

Author Query Form

Journal: Journal of Chemometrics

Article: cem_2575

Dear Author,

During the copyediting of your paper, the following queries arose. Please respond to these by annotating your proofs with the necessary changes/additions.

- If you intend to annotate your proof electronically, please refer to the E-annotation guidelines.
- If you intend to annotate your proof by means of hard-copy mark-up, please refer to the proof mark-up symbols guidelines. If manually writing corrections on your proof and returning it by fax, do not write too close to the edge of the paper. Please remember that illegible mark-ups may delay publication.

Whether you opt for hard-copy or electronic annotation of your proofs, we recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

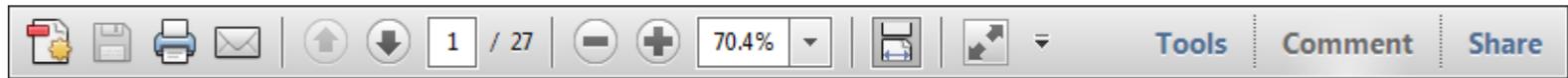
Query No.	Query	Remark
Q1	AUTHOR: Please check the suitability of the suggested short title.	
Q2	AUTHOR: Please provide a short abstract (maximum of 80 words) summarizing the key findings presented in the paper for Table of Content (TOC) entry.	
Q3	AUTHOR: Please provide the details (city, state [if USA], and country [other than USA]) of Hewlett Packard.	
Q4	AUTHOR: Please define VIP and iPLS.	
Q5	AUTHOR: Please list all authors.	
Q6	AUTHOR: Reference "23" is not cited in the text. Please indicate where it should be cited; or delete from the reference list and renumber the references in the text and reference list.	
Q7	AUTHOR: If this reference has now been published in print, please add relevant page information.	

USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

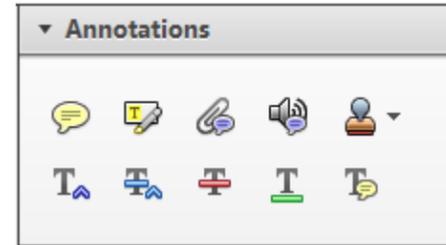
Required software to e-Annotate PDFs: Adobe Acrobat Professional or Adobe Reader (version 7.0 or above). (Note that this document uses screenshots from Adobe Reader X)

The latest version of Acrobat Reader can be downloaded for free at: <http://get.adobe.com/uk/reader/>

Once you have Acrobat Reader open on your computer, click on the [Comment](#) tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the [Annotations](#) section, pictured opposite. We've picked out some of these tools below:



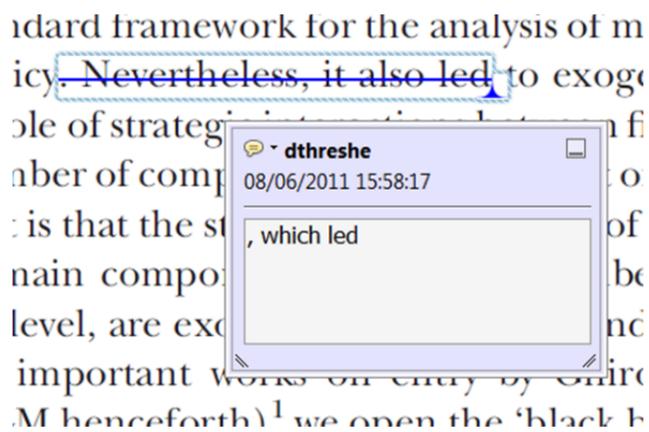
1. Replace (Ins) Tool – for replacing text.



Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the [Replace \(Ins\)](#) icon in the Annotations section.
- Type the replacement text into the blue box that appears.



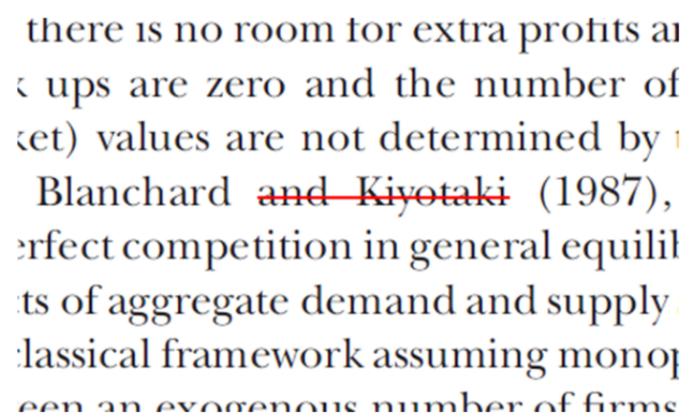
2. Strikethrough (Del) Tool – for deleting text.



Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the [Strikethrough \(Del\)](#) icon in the Annotations section.



3. Add note to text Tool – for highlighting a section to be changed to bold or italic.

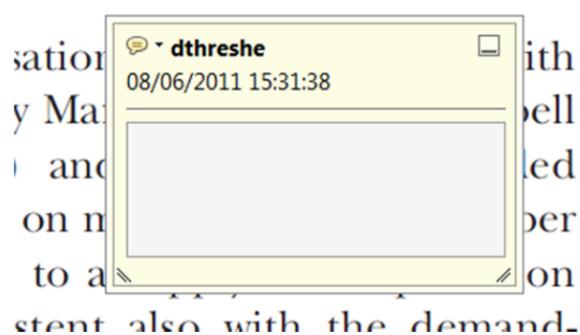


Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the [Add note to text](#) icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.

dynamic responses of mark ups
ent with the **VAR** evidence



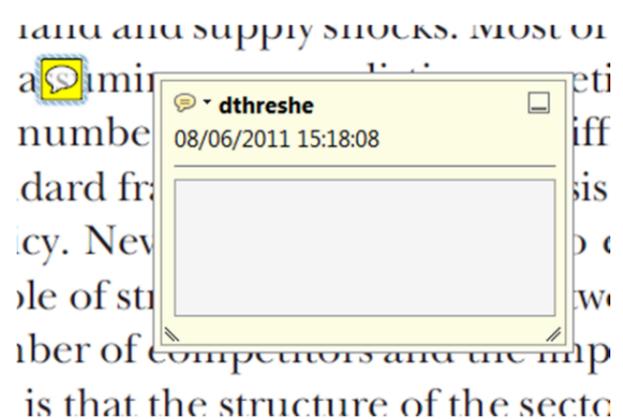
4. Add sticky note Tool – for making notes at specific points in the text.



Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the [Add sticky note](#) icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.



USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

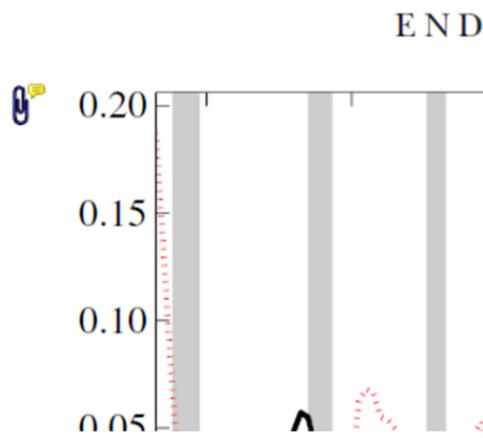
5. Attach File Tool – for inserting large amounts of text or replacement figures.



Inserts an icon linking to the attached file in the appropriate place in the text.

How to use it

- Click on the [Attach File](#) icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.



6. Add stamp Tool – for approving a proof if no corrections are required.

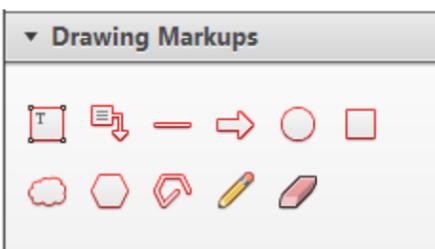


Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the [Add stamp](#) icon in the Annotations section.
- Select the stamp you want to use. (The [Approved](#) stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).

of the business cycle, starting with the
 on perfect competition, constant return
 production. In this environment goods
 extra profits and the number of firms
 he number of firms is determined by
 determined by the model. The New-Key
 otaki (1987), has introduced product
 general equilibrium models with nomi
 ed and supply shocks. Most of this literat

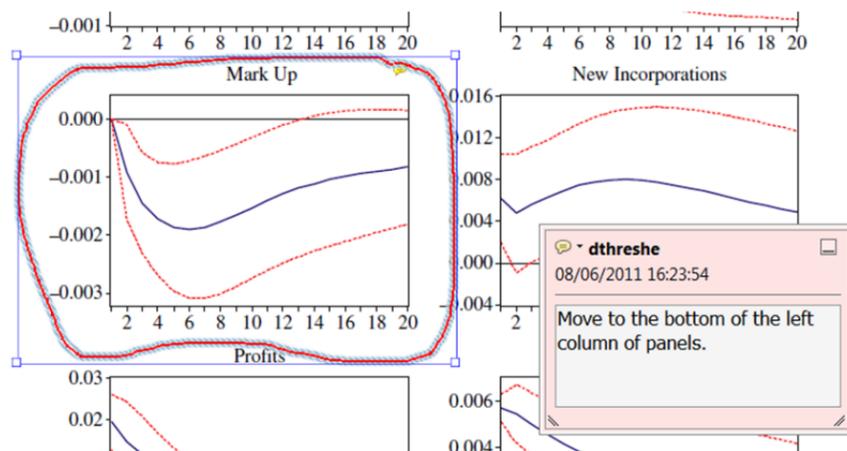


7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

How to use it

- Click on one of the shapes in the [Drawing Markups](#) section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.



For further information on how to annotate proofs, click on the [Help](#) menu to reveal a list of further options:

