PARADISe -
A ground-breaking tool to treat complex GC-MS datasets

MIKAEL AGERLIN PETERSEN and Rasmus Bro
University of Copenhagen, Department of Food Science, Rolighedsvej 26, DK 1958 Frederiksberg C

Abstract

A new approach to treatment of complex GC-MS datasets is introduced. The approach is based on PARAFAC2 modelling but does not require extensive coding and in-depth mathematical knowledge due to the new 'PARAFAC2 based Deconvolution and Identification System' (PARADISe). PARADISe can, in a user-friendly way, perform all the necessary steps in treatment of GC-MS data. It is demonstrated how PARADISe can efficiently quantify peaks, resolve co-elution, improve identification and save significant amounts of time.

Introduction

Modern GC-MS systems combined with efficient sampling techniques produce chromatograms with a large number of peaks of which many are not well-resolved. Well-designed experiments and screening investigations include many samples and replicates. The result is unavoidably heavy workload on the investigator to treat this data and extract the chemical information. Many approaches have been used from simple analysis of total ion chromatograms over single-ion techniques to different kinds of deconvolution techniques. They all have significant draw-backs: most are very time-consuming, results can be user-dependent to different degrees, and for almost all techniques, chromatograms are treated independently of each other. Furthermore, many approaches can only handle moderately overlapping peaks and often experience problems with low signal-to-noise peaks. Non-detects remain an issue as well.

Here, a completely different approach using the so-called PARAFAC2 modelling (PARAllel FACtor analysis 2) is demonstrated. Until now, PARAFAC2 modelling has only been available for mathematical users and has required extensive coding for efficient use [1]. An integrated approach called PARAFAC2 based Deconvolution and Identification System (PARADISe) has, however, become available. The solution is user-friendly, extremely time-saving, and produces reliable results that are less user-dependent. It is developed by a group of chemometricians around the 'Chemometrics and Analytical Technology' group at Department of Food Science, University of Copenhagen, and is freely available.

PARADISe benefits from the ability of PARAFAC2 to resolve co-eluting chromatographic peaks for all investigated chromatograms simultaneously [2]. It overcomes the limitation of PARAFAC2 which only works on time intervals, by assisting the user in defining appropriate intervals in the chromatograms, and it can thus perform all the necessary steps from visualization of data to generation of a final table of identified compounds for an entire set of chromatograms.

The steps in an analysis of a set of chromatograms by PARADISe are:

- Conversion of datafiles to AIA format
- Open/import files in PARADISe
- Inspect raw data (zoom/pan, search in NIST, exclude samples…)
- Define intervals

B. Siegmund & E. Leitner (Eds): Flavour Sci., 2018, Verlag der Technischen Universität Graz
DOI: 10.3217/978-3-85125-593-5-89,

- Calculate PARAFAC2 models
- Evaluate models (decide number of components)
- Tag relevant compounds
- Make report

In the following, examples are given to compare data treatment of real datasets done with a commonly used vendor software (Agilent ChemStation) and with PARADISe. It will be demonstrated how the techniques perform with regard to integration/baseline-modelling, deconvolution, peak identification, and user's time-consumption.

Experimental

Chromatograms from datasets exhibiting typical challenges were selected from recent projects carried out in our lab. The chromatograms were from different food products and were all obtained using dynamic headspace sampling in combination with thermal desorption (Perkin Elmer Turbomatrix ATD 650) gas chromatography mass spectrometry (7890A GC-system interfaced with a 5975C VL MSD with Triple-Axis detector from Agilent Technologies, Palo Alto, California) as described by Fjaeldstad et al. [3]. The chromatograms were treated using Agilent's software ChemStation (MSD ChemStation E.02.02.1431) and using PARADISe, a software package developed by Johnsen et al. [4] and available from *http://models.life.ku.dk/paradise* (PARADISe version 1.1.6).

Results and discussion

Example 1

This is a simple case to demonstrate the basic features in PARAFAC2 modelling as carried out in PARADISe. The raw data is the time interval from 3.71 to 3.99 minutes taken from 40 chromatograms. Part of the task in using PARADISe is to determine how many components need to be used. There are several utilities for this in the software and some are explained in Example 3. Figure 1 shows how a PARAFAC2 model with 2 components can separate the raw data into two 'phenomena' or components: Component 1 which includes mass fragments of typical background noise (air, water a.o.) and component 2 which mainly includes the mass fragments 43 and 86 (see Figure 2).

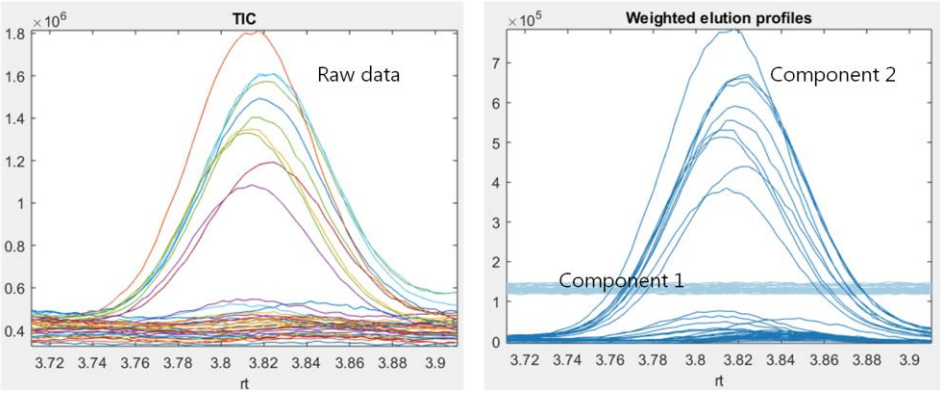

Figure 1: Total Ion Chromatograms (TIC) from the interval 3.71 - 3.99 min from 40 chromatograms and weighted elution profiles from a PARAFAC2 model with two components

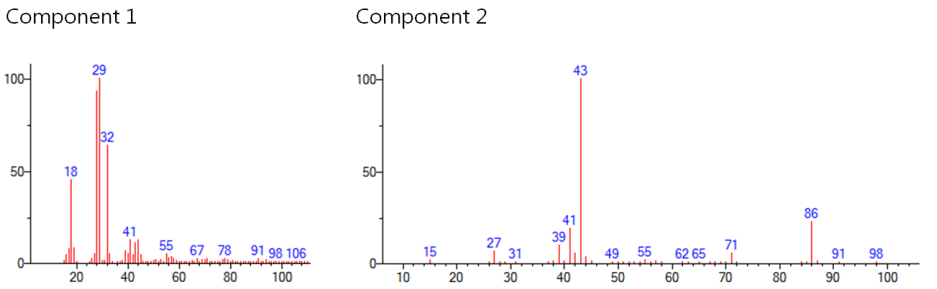

Figure 2: Patterns of mass fragments (=mass spectra) constituting component 1 and 2 in Figure 1

The mass fragments of component 2 do actually make up a mass spectrum, and when searched in the NIST database, it was identified as 3-methyl-2-butanone. It is seen that the PARAFAC2 model eliminates the need for integration of peaks. Instead the background is modelled and separated into its own component(s), in this case component 1, so component 2 exclusively represents 3-methyl-2-butanone. Even background noise that changes in intensity and in composition throughout the interval can be modelled, but may then require more than one component.

The PARAFAC2 model extracts one mass spectrum for each component by combining information from all chromatograms. This results in a mass spectrum of higher quality and better match factors are most often experienced. Finally, the PARAFAC2 model creates a concentration profile which is a list of the peak areas in all the chromatograms included. It should be noted that minor retention time shifts (for example as those seen most clearly in the weighted elution profiles in Figure 1) are handled by the model without problems. It is also worth noting that the PARAFAC2 model does not assume any particular shape (e.g. Gaussian or Lorentzian) of the elution profiles. The shape is solely determined by the data.

Example 2

This example demonstrates a more complex situation, see Figure 3. The figure shows two coeluting peaks which were expected to be 2- and 3-methylbutanal. PARAFAC2 modelling did, however, reveal that 6 ‘phenomena’ or components could be found in the interval, see Figure 4.

The first component is representing ethyl acetate, but it is only a small remain (or ‘tail’) not belonging to this time interval. Component 2 and 3 represent rather small peaks that were hidden behind 2- and 3-methylbutanal in the TIC, but could still be identified with high match factors as vinyl isopentylether and 2-methyl-2-propanol. Component 4 models background noise. So, in addition to performing a near perfect separation, and thus quantification and identification, of 2- and 3-methylbutanal, two hidden peaks were identified and quantified with high reliability.

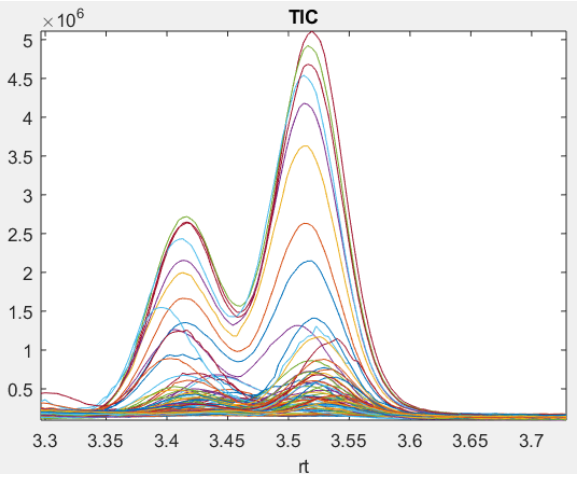

Figure 3: Total Ion Chromatograms (TIC) from the interval 3.30 - 3.75 min from 80 chromatograms

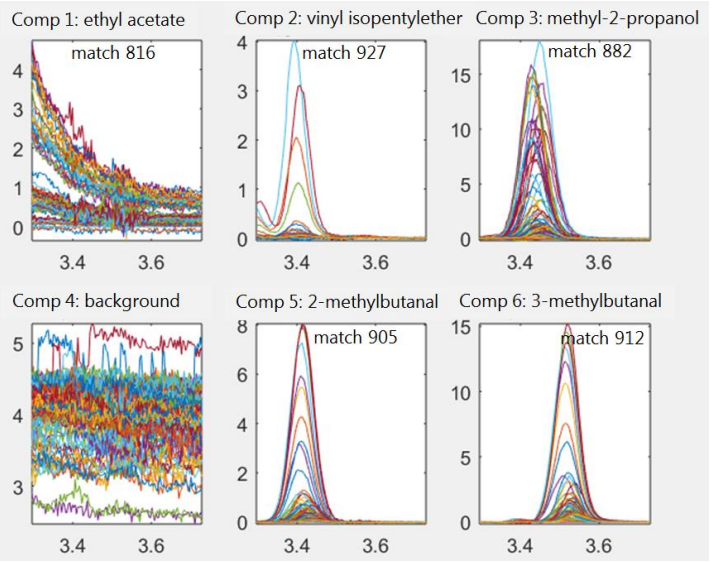

Figure 4: Weighted elution profiles (not overlaid) from a 6 component PARAFAC2 model applied to the data shown in Figure 3. Identifications and match factors from search in the NIST database are also shown.

Example 3

This example shows how the appropriate number of components is determined and how decisions on number of components affect the data obtained. The same 40 chromatograms and the same retention time interval as in example 1 are used, supplemented by data from the interval 3.73 - 3.92 min which include the compound 2-butanone. Figure 5 and 6 show peak areas of the two compounds from five selected samples. The peak areas were calculated from the TIC's using standard integration settings in ChemStation and by applying 1, 2, 3 and 4 component PARAFAC2 models in PARADISe.

To determine the appropriate number of components in the models, PARADISe includes two diagnostics: Fit and core consistency. Fit will normally increase with increasing number of components while core consistency tends to decrease. Both values should be as high as possible (range: 0-100). Fit and core consistency are included in the figures. The numbers indicate that 2 and 3 components could both be reasonable. When several models are appropriate it is often useful to select the one with most components in order to extract as many chemical pieces of information as possible.

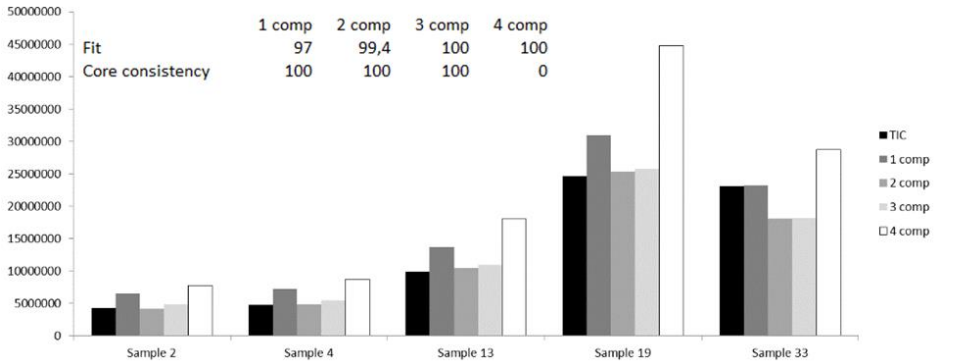

Figure 5: Peak areas of 2-butanone in five selected samples. The peak areas were calculated from TIC's using standard integration settings in ChemStation and by applying 1, 2, 3 and 4 component PARAFAC2 models.

2-Butanone (Figure 5) is a medium sized peak. A 3 component model would be the choice since it has fit and core consistency values of 100. The 2 component model works almost equally well, but the 4 component model is obviously wrong, having a core consistency of 0. The 1 component model gives too high peak areas because the background noise is not modelled by a separate component but is included in component 1. The TIC data from ChemStation fits the 2 and 3 component models well. The reason for the discrepancy in sample 33 is a coelution which is not resolved by ChemStation (and neither by the 1 component model).

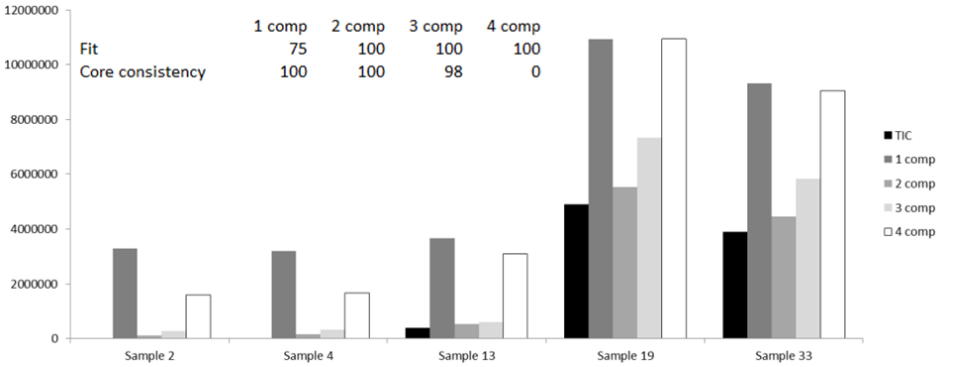

Figure 6: Peak areas of 3-methyl-2-butanone in five selected samples. The peak areas were calculated from TIC's using standard integration settings in ChemStation and by applying 1, 2, 3 and 4 component PARAFAC2 models.

3-Methyl-2-butanone (Figure 6) has very small peaks in some of the samples. A 2 component model would be the choice since it has fit and core consistency values of 100.

The 3 and especially the 4 component models have lower core consistency and are therefore less appropriate. The TIC data from ChemStation fits the 2 component model well except in sample 2 and 4 where the peak is too small to be integrated by the ChemStation software.

This example shows that the diagnostics fit and core consistency give good guidance in determining the correct number of components. Even when the guidance is not clear (as for 2-butanone) the two possible selections (2 or 3 components) result in almost equal peak areas. Furthermore, it is demonstrated that PARADISe does not depend on integration settings, but gives areas of all peaks independent of their size, and that the peak areas reported by PARADISe are practically equal to those obtained when well separated TIC peaks are integrated in ChemStation. Note, that even in samples without a certain chemical present, it will still be quantified. All chemicals are quantified in all samples and hence, there is no issue with below limit of detection.

Time consumption

To go through the steps mentioned in the introduction, a user of PARADISe will typically spend a few minutes to convert and import files. Time used for inspecting raw data depends mostly on the data. Defining intervals can be done within 30 min for an experienced user. The calculation of PARAFAC2 models is very time consuming (few hours to more than a day) but will be carried out by the computer unattended. Evaluating the models and tagging compounds may take up to a couple of hours depending on the complexity of the chromatograms, and finally the report is created within few minutes. In total, the typical time consumption will be 2-3 hours for an experienced user – almost independent on the number of chromatograms included.

Conclusion

It is concluded that treatment of large datasets with PARADISe results in extraction of more information, the information is more reliable, and user's time-consumption when treating datasets with numerous complex samples/chromatograms is dramatically reduced.

References

1. Kiers H.A.L., ten Berge J..MF. and Bro R. (1999), Journal of Chemometrics, 13, 275-294.
2. Amigo J.M., Popielarz M.J., Callejón R.M., Morales M.L., Troncoso A.M., Petersen M.A. and Toldam-Andersen T.B. (2010), Journal of Chromatography A, 1217, 4422–4429
3. Fjaeldstad A., Petersen M.A. and Ovesen T. (2017), Chem. Percept., 10, 42–48
4. Johnsen L.G., Skou P.B., Khakimov B. and Bro R. (2017), Journal of Chromatography A, 1503, 57–64