Analytica Chimica Acta: X 1 (2019) 100005

Contents lists available at ScienceDirect

Analytica Chimica Acta: X

journal homepage: www.journals.elsevier.com/analytica-chimica-acta-x

Automated supervised learning pipeline for non-targeted GC-MS data analysis

Kimmo Sirén [a, b], Ulrich Fischer [a], Jochen Vestner [a, *]

[a] Institute for Viticulture and Oenology, DLR Rheinpfalz, Breitenweg 71, D-67435, Neustadt, Germany
[b] Department of Chemistry, University of Kaiserslautern, Erwin-Schroedinger-Strasse 52, D-67663, Kaiserslautern, Germany

ARTICLE INFO

Article history:
Received 15 November 2018
Received in revised form
21 December 2018
Accepted 2 January 2019
Available online 10 January 2019

Keywords:
Metabolomics
Chemometrics
Tensor decomposition
Machine learning
Classification
Exploratory data analysis

ABSTRACT

Non-targeted analysis is nowadays applied in many different domains of analytical chemistry such as metabolomics, environmental and food analysis. Conventional processing strategies for GC-MS data include baseline correction, feature detection, and retention time alignment before multivariate modeling. These techniques can be prone to errors and therefore time-consuming manual corrections are generally necessary. We introduce here a novel fully automated approach to non-targeted GC-MS data processing. This new approach avoids feature extraction and retention time alignment. Supervised machine learning on decomposed tensors of segmented chromatographic raw data signal is used to rank regions in the chromatograms contributing to differentiation between sample classes. The performance of this novel data analysis approach is demonstrated on three published datasets.
© 2019 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

High-throughput approaches combined with non-targeted analysis are increasingly used in several different research disciplines such as food and environmental sciences, systems biology and metabolomics. In contrast to targeted analysis methods, the non-targeted approaches are inherently more holistic and therefore provide more comprehensive overview of the sample composition. These approaches are more hypothesis generating as they do not rely on *a priori* defined set of compounds but take known and unknown compounds into account.

The main objective of non-targeted studies is to discover molecules, which distinguish between groups or classes of samples, such as biomarkers from large sparse datasets [1–3]. As high-throughput approaches result in large sample sizes, this objective can be seen as a machine learning classification problem. For LC-MS data open source software such as XCMS [4], MS-DIAL [5] and many others are usually applied [6]. GC-MS data analysis commonly relies on vendor software which is often not applicable or compatible to data formats of other vendors [7]. The freely available software tools used in metabolomics have been recently listed [6]. Due to the popularity of XCMS for LC-MS, many have applied it for GC-MS, though however, the manual parameter optimisation remains vague [8]. Common data analysis approaches for non-targeted GC-MS data use feature detection in single ion chromatograms of individual samples in order to extract quantitative information resulting in concatenated data frames such as peak tables [1]. Typically downstream approaches such as multivariate analysis techniques e.g. unsupervised principal component analysis (PCA) and supervised partial least squares discriminant analysis (PLS-DA) are often subsequently used to analyze and visualize these feature tables. Automated feature detection however remains troublesome due to coelution, retention time shifts and potential erroneous peak picking and/or peak assignment. Relevant information can get lost by using intensity thresholds to differentiate between analytical signals and noise. Moreover, feature detection and alignment of chromatograms are difficult to automate as algorithm settings have to be optimized and validated resulting in more hands-on time. This also reduces reproducibility of the studies.

Several alternative approaches to conventional non-targeted GC-MS data analysis have been developed [9–23]. These approaches aim at better extraction of information and underlying patterns by using the GC-MS raw data signals (retention time × mass-to-charge ratio) as chromatographic fingerprints for modeling. In this way, feature detection is circumvented. Some of

* Corresponding author.
E-mail address: jochen.vestner@dlr.rlp.de (J. Vestner).

https://doi.org/10.1016/j.acax.2019.100005
2590-1346/© 2019 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Abbreviations

PARAFAC	Parallel Factor Analysis
PLS-DA	Partial Least Squares Discriminant Analysis
PCA	Principal Component Analysis
SSCP	Sum of Squares and Cross Product
t-SNE	t-Distributed Stochastic Neighbor Embedding
TIC	Total Ion Chromatogram
RMS-noise	Root Mean Square Noise consensus score

these alternative approaches are avoiding retention time alignment by implementing transformation of raw data signal to particular distance matrices [9,10,13,16–19]. A recent clustering approach [22] to extract features from single samples has been updated to fit MZMine 2 [24] workflow. Additionally, other recent full workflow approaches rely on baseline and time shift corrections or other preprocessing methods [21,25–27]. Multiway approaches to GC-MS data [9,16,18,20] such as PARAFAC has been evaluated against multivariate curve resolution by alternating least squares (MCR-ALS) which are the two most well-known mixture resolving methods. [28]. Recent approaches using alternative segmentation of the GC-MS raw data signal and subsequent feature se lection have been proposed [20,23] These approaches rely on normalized data [23] or focus on single samples [20]. However, many recent approaches that use multivariate statistics or machine learning still continue to rely on preprocessed peak tables instead of using the raw signals for modeling [29–34]. Vestner et al. [10] based their data analysis approach on avoiding classical feature extraction and alignment by segmenting chromatograms along the retention time axis and transforming the two dimensional chromatographic segments to sums of squares and cross-product (SSCP) matrices of the mass channels. In this study we expand this idea further by using supervised learning methods to employ the neglected power of sample sizes and *a priori* knowledge of group memberships on SSCP matrices of automatically chosen segments. We minimize hands-on time by focusing on ranking segments in the chromatogram. Ultimately, the ranking is established on the ability of individual segments to differentiate between classes. Downstream analysis of segments contributing to class differentiation is similar to conventional approaches and can include visual exploration or more sophisticated deconvolution of mass spectra of the relevant regions can be performed if needed [19,35]. Moreover, chemical compounds in the important segments can subsequently be tentatively identified based on retention indices and (resolved) mass spectra.

The main objective of this research was the development of a fast and fully automated approach avoiding time consuming optimization and parameter tuning. The performance of this novel supervised learning approach is demonstrated on three published datasets by comparing the approach to the published results.

2. Material and methods

2.1. Theory

An overview of the workflow is presented in Fig. 1. The workflow is based on following ideas: Segmentation of the whole dataset, construction of a classifier model for the transformed raw data of each segment separately, and the ranking of segments according to their importance to discriminate between classes of samples using model performance metrics. Retention time alignment is circumvented by an automated segmentation of chromatograms along the retention time axis. Subsequent modeling of the full GC-MS raw

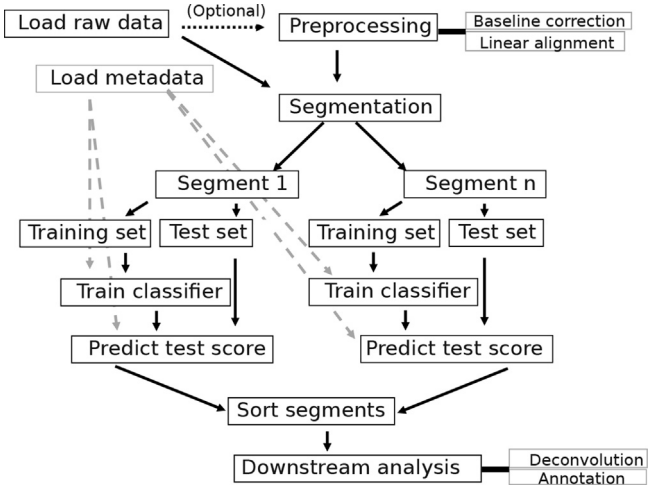

Fig. 1. Flowchart of the non-targeted GC-MS data analysis workflow. Gray arrows mark stages where metadata is needed. The whole analysis is automatic and does not require manual work steps or parameter tuning.

data signals of each individual segment makes feature extraction unnecessary. For each segment, chromatographic raw data signals of all samples are transformed into SSCP matrices following the approach described by Vestner et al. [10]. These SSCP matrices for all samples are merged into a tensor of order three, and decomposed using tensor decomposition. Next, a supervised machine learning pipeline based on gradient boosted tree classification is used on the decomposed tensor. The performance of prediction models are evaluated in order to establish the ranking of the segments. Classification metrics provide information on the importance of each segment on the differentiation among sample classes.

In detail, automated segmentation of chromatograms is achieved by finding local minima of summed total ion chromatograms (TICs) of all samples. With regular injection precision of GC-MS instrumentation, baseline correction and retention time alignment are generally not needed. Larger retention time shifts in GC-MS analysis indicate instability of the system and should rather be solved by optimizing the instrument performance instead of correcting data. However, linear alignment of chromatograms can be used to minimize inter-segment shifting of peaks. Whereas intra-segment shifting is dealt with by transforming the raw data signal matrix (retention time $\times$ mass-to-charge ratio) of each segment and sample to SSCP matrices. For each segment, the SSCP matrices of all samples are stacked together into tensors of order three. Each of these tensors is then individually decomposed using tucker decomposition [36]. Ultimately, the tensor decomposition sample loadings can be interpreted as a representation of the major source of variation among samples in each segment. The tensor decomposition core sizes are optimized to retain 99% variation by selecting enough ranks. The sample dimension loadings are then used to solve the supervised learning classification problem. Thus, a supervised classification model (classifier) is trained based on predefined classes for each segment individually resulting in as many models as segments. Principally, any classifier could be used but tree ensemble models have been found to retain higher order interactions between features and no inter-sample scaling of inputs is required [37].

Finally, the ranking of the segments is attained by evaluating the performance of each model (segment) using custom classification metrics. This ranking of the models according to their importance reflects the amount of information relevant in each segment - thus allowing extraction of the most relevant segments. We follow

established definitions for classification metrics in this paper [38]. The metrics used here are chosen to be applicable to evaluate both binary and multi-class classification models. Classical metrics such as precision, recall and F-score are generally regarded to be insufficient to estimate the performance of classification [39]. Requirements for metrics are as follows: First, individual classes should be highly recoverable. Second, model stability needs to be evaluated. A similar approach has been suggested to classify transcriptomics data using Precision-Recall (PR) curves and Receiver Operating Characteristics (ROC) area under curves (AUC) [40]. PR curves present more informative picture of model performance even though PR and ROC spaces are deeply connected. Therefore, if a curve dominates in ROC space it also dominates in PR space [41]. In order to control class imbalances, micro-average scores are used, so that by the aggregation of all class contributions the average is calculated. The multi-class ROC and AUC calculations were based on Provost and Domingos [42] approach as described by Fawcett [38]. Micro average of AUC and the maximum average value of the PR curve area are used in two dimensional fashion as described by Carbonero-Ruz et al. [43].

After ranking, the most important segments discriminating between sample classes can be further downstream analyzed to extract more chemically meaningful information. Moreover, Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) [44] or clustered heat maps based on similarity and dissimilarity matrices can be used to visually investigate the relations between segments and samples. Additionally, t-SNE can be useful in order to retain both global and local structure which is not possible with traditional dimensionality reduction techniques such as PCA [44]. Subsequently, further downstream processing could include deconvolution of mass spectra using e.g. PARAFAC2 as recently suggested by Tian et al. [45] followed by metabolite annotation process.

2.2. Methods

The here described approach is entirely written in Python 3 programming language [46]. GC-MS files were imported as NetCDF-format [47]. Neither baseline correction, nor linear alignment were applied. For the segmentation, a consensus chromatogram is obtained by summing the TICs of all samples and a consensus score of the root mean square noise (RMS-noise) is estimated. More specifically, the RMS-noise consensus score is estimated by calculating the square root of the summed intensities of each sample and taking the mean value of all the samples scaled to a percentage. Local minima are found by comparing the potential candidate point to the following data points while taking into account the RMS-noise. Note that the lower the RMS-noise value, the more minima are found. A modified script from https://gist.github.com/sixtenbe/1178136 was applied.

In every segment for each sample, a SSCP matrix was calculated. Subsequently, the SSCP matrices of each sample were concatenated to a tensor of order 3 for each segment. These tensors were individually decomposed using tucker decomposition via higher order orthogonal iteration [48,49]. Sample loadings were used to build a supervised learning classification problem. In addition, for model validation, leave-one-out cross-validation was chosen due to the relative low amount of samples in one of the dataset (Table 1).

For the statistical learning pipeline, a tree booster classifier XGBOOST was employed [37]. The default model parameters of the algorithm were retained without further fine-tuning to showcase the approach. The following parameters were used: learning_rate = 0.1, max_depth = 3, min_child_weight = 1, n_estimators = 100, reg_alpha = 0, reg_lambda = 1, scale_pos_weight = 1. Probabilistic function softmax was used as an objective function. Evaluation of the model cycles was set with multiple logistic loss criteria for multiple classes as in logistic regression for the training loss.

Table 1
Description of the datasets used.

Dataset	Samples	Number of classes	Project	Reference
Wine	39	3	Not publicly available	[10]
Urea	160	4	MTBLS71	[55]
Rice	79	20	MTBLS288	[56]

In order to rank and choose segments for further downstream analysis, ROC, AUC and PR curves were calculated from the true and the predicted probabilities of classes for every segment [50]. Segments with micro-average of ROC curve threshold and maximum class PR curve value higher than 0.9 were collected and further evaluated. For downstream examples the learning pipeline was rerun keeping only these selected segments. The parameters for t-SNE were set as follows: perplexity 10, early exaggeration 2, and learning rate 10. Acceleration of the computation was done using Barnes-Hut approach and initializing with dimensionality reduction with PCA [51]. The multidimensional data was visualized using Matplotlib and Seaborn plotting libraries [52,53].

2.3. Application

The fully automatized non-targeted GC-MS data analysis approach reported herein was tested with three different and independent, already published datasets. Two data sets are publicly available through the Metabolights database [54]. Using published data for the verification of the approach has two main advantages. Firstly, the outcome of the data analysis can be directly compared with already published findings. Secondly, multiple datasets from different domains of analytical chemistry can be used for confirming the applicability of the new approach in these domains.

For an overview of the datasets see Table 1. The first dataset consists of GC-MS fingerprinting analysis of volatile constituents of Cabernet Sauvignon wines submitted to different fermentation scenarios. The second dataset derives from a metabolomics study on the impact of urease treatment of urine samples prior to GC-MS [55] analysis. The third dataset originates from a metabolomics study on rice focusing on metabolic shifts during rice grain development [56]. Our workflow, shown in Fig. 1, was applied to all datasets in the same way. The only interactions between the user and the software are: providing the target classes and optionally controlling the segmentation with desired level of RMS-noise and how many data points are included in the segmentation algorithm. The workflow is therefore completely free of user interaction during the data analysis. For the workflow comparison the segmentation parameters were kept default as following: data points: 15; RMS-noise: 1.

2.4. Impact of segmentation on algorithm performance

In the following, the accuracy of finding relevant segment is assessed. In this regards, the impact of the number of segments on the performance of the approach was investigated for all three datasets without linear alignment as follows: Segment sizes were varied for the automated segmentation by changing the amount of RMS-noise (RMS-noise ε {100, 10, 1, 0.1}). All segments with micro-average AUC score higher than 0.9 were selected. Additionally, cophenetic correlation coefficients were calculated to estimate the quality of the hierarchical clustering dendrogram by preserving the pairwise correlation distances with average clustering method

between the original data [57,58].

For the urea dataset the same approach was followed with a robust linear retention time alignment using openCV module [59]. Ultimately, the same dataset was also further downstream analyzed by deconvolution of the important segments using PARAFAC2 algorithm of the multiway package [60] in R (version 3.4.1) [61]. Random forest based [62] feature selection wrapper algorithm [63] was used to retain most relevant deconvoluted peaks contributing to the separation between the four classes.

3. Results and discussion

A comparison of the outcome of our new approach to published results for the three data sets is presented in the following. The major focus of our new algorithm lies on the very fast elucidation of differentiation between sample classes and the identification of segments in the chromatogram which are responsible for differences between classes. However, to provide confirmation of the importance of the found segments, a more detailed downstream analysis for the wine data set is provided. Subsequently, with the rice and the urea data set two further examples of the performance of the new approach are provided focussing on the comparison of the results of class differentiation of our approach to the published data.

3.1. Wine dataset

The wine dataset consists of 39 chromatograms of wines fermented with three different commercial yeast products. Number of samples for each class were 20, 12 and 7 for Clos, Rbs and Vrb respectively [10]. No baseline correction or linear alignment was performed. TICs of all samples were summed to create a reference chromatogram for the segmentation, which resulted in 78 segments. Overall ranking of all segments is visualized by plotting the segment scores of the two metrics micro average of AUC and the maximum average value of the PR curve area against each other (Fig. S1).

After the ranking of the modeled segments according to the ability to differentiate between the three classes, 10 segments exceeded the cut-off values of the model evaluation metrics of 0.9 and were therefore selected for further investigation. A visual inspection of the TIC overlays of these segments provides clear evidence that these segments contain peaks, which clearly contribute to the differentiation of sample classes (Figs. S2–11). Examples of rejected and accepted segments are shown in Fig. 2. Segment loadings of the tucker decomposition of accepted segments are visualized using t-SNE. The iteration was stopped after no further progress was made with an error of 0.0474 after 119250 iterations. All three classes are clearly separated (Fig. 3). The class Rbs differentiates along the first dimension whereas the Vrb and Clos differentiate along the second dimension (Fig. 3). Moreover, an equivalent result figure is also found by using the approach described in the original work [10] which the reader is referred to.

Our proposed workflow is easily adjustable to further downstream processes such as PARADISE [64]. Moreover, in order to investigate deeper on the actual chemical compounds as potential biomarkers, PARAFAC2 deconvolution and annotation of deconvoluted mass spectra to NIST library of all peaks in the 10 accepted segments was performed as described in the reference method [10]. In total 15 peaks were found (Table S1). A comparison of the results with the original work reveals that 13 compounds were found by both studies while one unknown compound (LRI 1263) was not found with proposed supervised learning method. Note, that this unknown compound is contained in segment 31 which has still very high metrics scores of 0.89 and 0.78 for average precision score and micro average of ROC curve, respectively. Interestingly, two new compounds were found with the here proposed supervised learning method, which were not considered in the original paper (Table S1). These two compounds were tentatively identified as ethyl heptanoate and linalool. Particularly, linalool is known to be an important wine aroma constitute and the concentration of linalool can be strongly influenced by the yeast strain used during fermentation [65].

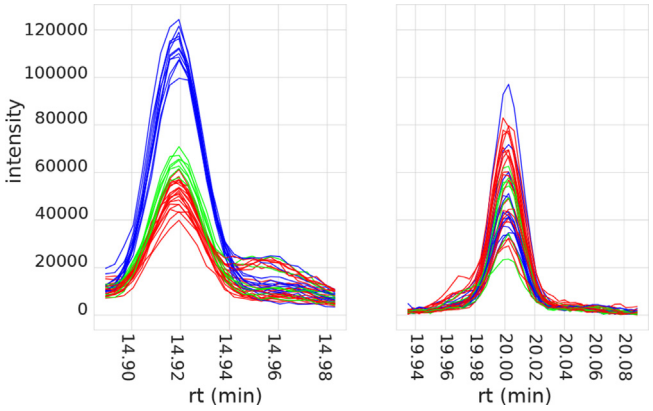

Fig. 2. TICs of a good and a poor segment found by the classification model of the wine dataset [10]. Good classification happens when one or multiple classes are clearly differentiated from the rest. For poor classification no separation between classes is observed. Colors correspond to the different commercial wine yeast starter cultures. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.2. Urea dataset

The urease pretreatment dataset was downloaded from MetaboLights [54] with identifier MTBLS71 [55]. The experiment consists of 160 samples, with two separate classes: female-male; and no pretreatment - urease pretreated, yielding 4 classes in total. Although, this could be seen as a multi-labelled classification task, a multi-class approach was done in accordance with our approach.

TICs of all samples were summed to create a reference chromatogram which was used for segmentation resulting in 95 segments. 14 segments were retained for downstream analysis. Tucker decomposed loadings of the chosen segments were visualized using t-SNE with an error of 0.4928 after 300000 iterations. As the data was more noisy, the perplexity value was set to 4 from default 10 to emphasize more local effects [55]. The main separation was observed among urease pretreated and the control, while there seems to be a trend dividing the female-male class especially in the urease pretreated class (Fig. 4). The reference method gives a corresponding result figure as described in the original work [55] which the reader is referred to. Additionally, the visualization shows clearly that the experiment was done with biological duplicates. These findings are in accordance with the original work, where the main driving factor of variability was identified as the pretreatment while variability related to gender was only observed in lower-resolution [55].

3.3. Rice dataset

The rice cultivar dataset was downloaded from MetaboLights [54] with identifier MTBLS288 [56] and consists of 79 samples. Two sets of classes are defined: grain cultivars (4 classes) and development measured in days (5 classes). In regards to the original

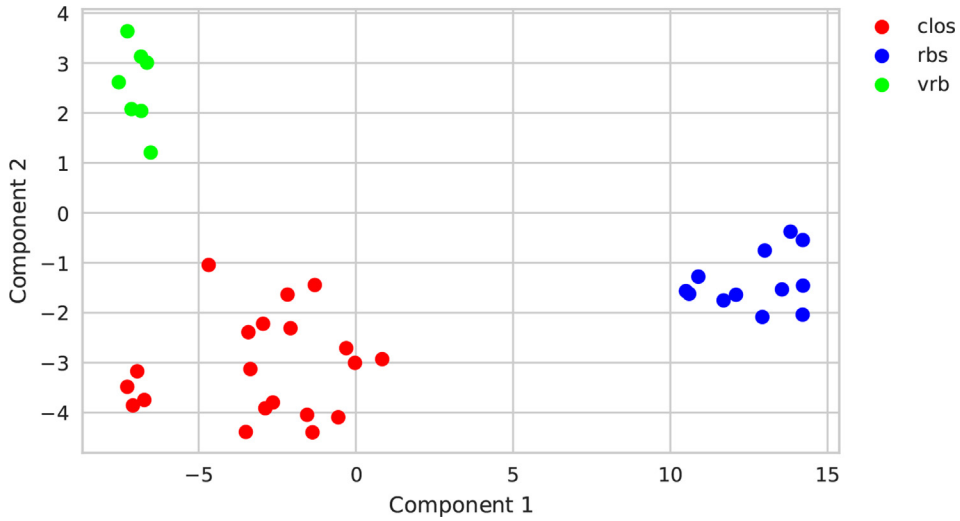

Fig. 3. t-SNE visualization of the wine dataset [10]. Tucker decomposed sample loadings of the ten highest scored segments are projected with two component t-SNE visualization. The three commercial wine yeast products are clearly separated from each other. Colors correspond to the different commercial wine yeast products. Rbs: commercial yeast product 1, Clos: commercial yeast product 2, Vrb: commercial yeast product 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

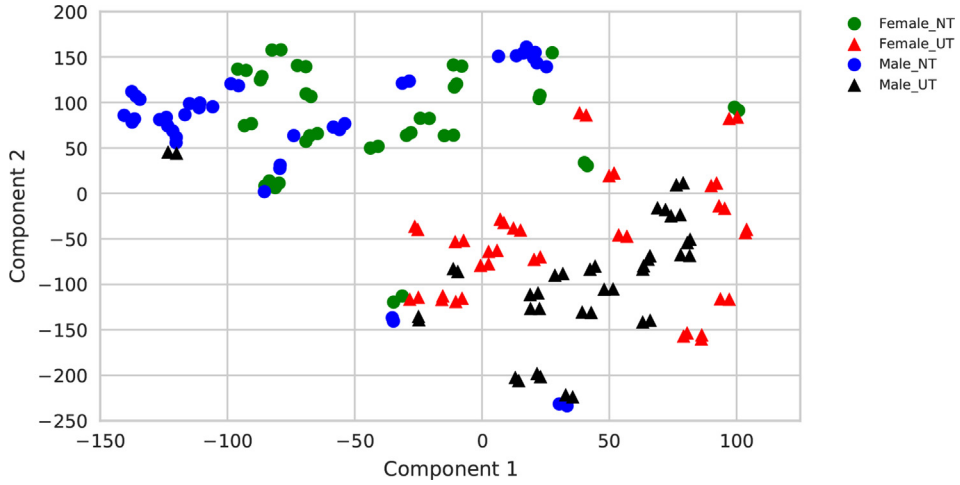

Fig. 4. t-SNE visualization of the urea dataset [55]. Tucker decomposed sample loadings of the 14 highest scored segments are projected with two component t-SNE visualization. Highest variation in the dataset is between the urease pretreated and no pretreated. Additionally, urease pretreated male and female classes seem to be different. NT: no pretreatment; UT: urease pretreatment.

authors work we chose to combine these two to create 20 classes with each having 4 samples. For the group 14_Nip only 3 samples were found due to some inconsistency in the public repository. No baseline correction or linear alignment was performed. TICs of all samples were summed to create a reference chromatogram which was used to for segmentation resulting in 134 segments. In total 38 segments were retained after running the pipeline. A comparably large number of important chemical compounds were also observed by the original authors [56].

Further, the sample loadings of the tucker decomposition of the retained segments were visualized with t-SNE. The iteration was stopped after no further progress was made with an error of 0.0697 after 105750 iterations. From the t-SNE representation it is evident that the rice grain development days form a gradient (Fig. 5). Early stage development samples (7–14 days) are seemingly more separated than the late stage samples (28–42 days). In the latter the cultivar effect seems to be more dominant (Fig. 5). These findings follow the same pattern as reported in the original work, in which the resolved metabolic features were presented through PCA [56] with a results figure which the reader is referred to. By taking overfitting into consideration, this example verifies that even for small sample sizes per class, it is still possible with our fully automated approach to extract the coincident and meaningful information as reported in the original work.

3.4. Impact of segmentation on algorithm performance

In order to investigate the effect of the robust segmentation on the results of the approach, four different segmentation scenarios were tested. Differing numbers of segments were achieved by varying the amount of RMS-noise for the detection of minima in the reference chromatogram. For all datasets it was found that the number of chosen segments is reaching a plateau once a relevant number of segments is reached (Tables S2–S5). For the urea dataset the segments with metric values larger than 0.9 after running our pipeline were deconvoluted with PARAFAC2 to obtain more

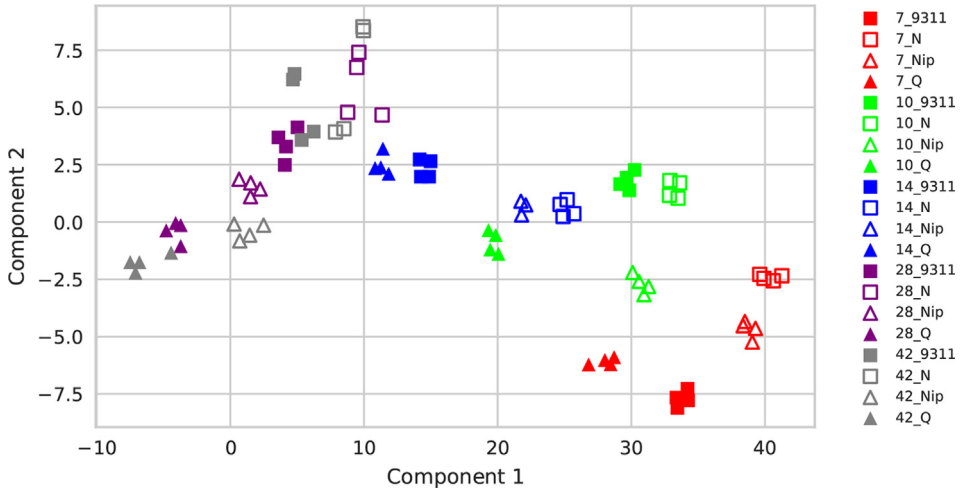

Fig. 5. t-SNE visualization of the rice dataset [56]. Tucker decomposed sample loadings of the 30 highest scored segments are projected with two component t-SNE visualization. The first component of the t-SNE distinguishes rice grains from different developmental stages. The early stage development (7-14) variation is higher than the late stage development (28−42), where the cultivar effect seems to be more dominant. The numbers in the sample names correspond to grain development measured in days. The labels 9311, N, Nip and O correspond to the cultivars 9311, Nongken 58, Nipponbare and Qingfengai, respectively.

information on the performance during downstream analysis. The selected deconvoluted peaks also reach a maximum number (Table S5). Furthermore, clustering performance, determined with cophenetic correlation coefficient, follows the same behavior as described above. Robust linear alignment leads to a small increased number of important deconvolute compounds (Table S5).

Smaller segment intervals are generally preferable as they are easier to deconvolute and to investigate visually. Large intervals with low number of segments are robust but also prone to lose information. While the classifier works on raw data signal (taking interactions into account) the evaluation of the classifier performance becomes less pronounced for complex segments, which in turn leads to lower segment selection metrics. For exploratory data analysis purposes segment selection parameter thresholds can be set lower in order to increase the amount of false positives, which can be identified and removed during the downstream analysis process. Small intervals with many segments lead however not necessarily to more important information in the downstream analysis as demonstrated by the cophenetic correlation table for the urea dataset (Table S5).

The size of segments is also crucial for the tensor decomposition and downstream processing, as the rank of the segment tensors is related to speed and accuracy of the model performance [66]. Smaller segments are easier to deconvolute for instance with PARAFAC2 to obtain more reliable and valid models [67]. Other interesting strategies for segmentation of chromatograms could be further included in our approach. Our data analysis strategy could also be used with other established methods for feature detection instead of using transformed chromatogram segments [4,22,67,68], albeit the advantages of modeling decomposed raw data segments would of course be lost. Using our approach with other established strategies for feature selection might not only increase resolution of the approach but makes the model applicable to other chromatographic techniques and to high resolution mass spectrometry.

3.5. Algorithmic efficiency and user interaction

Besides the optional control of the segmentation with desired level of RMS-noise and the inspection of results no further user interaction is needed for our approach. The calculation time of the algorithm depends on the total scan numbers in the chromatogram and the number of samples. The computations were performed on a computer equipped with an quad core Intel Core i7-6700 CPU with 3.4 Ghz. Total computation time including all segments for the three data sets took 1 min, 12.7 min and 20.7 min for the wine, urea and rice data set, respectively. Downstream analysis including repetition of the algorithm with only important segments took additional 0.1 min, 3.3 min and 2.5 min, respectively. These facts make our approach a very fast and strong tool to find class differentiation in non-targeted GC-MS data and reducing the time spent on total data analysis tremendously. Through greater efficiency and a shorter analysis route, specific important regions in the chromatograms can be discovered for any further investigation.

4. Conclusions

In this paper we introduce a novel, automated workflow for non-targeted GC-MS exploratory data analysis using supervised machine learning exploiting the power of sample size and knowledge of class memberships. Remarkably, with only segmented chromatographic raw data signals, meaningful information can be extracted in an automated way allowing the user productively focus on the downstreaming of only important regions in the chromatograms, which are responsible for class separation.

Our approach is reducing manipulation of data during the data analysis progress such as peak picking, deconvolution and retention time alignment. We have shown that our approach is able to reproduce the results of three published datasets. The key benefit of this automatized workflow is to speed up data analysis and facilitate differentiation between sample groups straight from the chromatographic raw data signals allowing the user to focus on the inspection of the relevant regions. Classification is therefore directly achieved, whereas additional tools for downstream analysis are necessary for the identification of biomarkers. The proposed workflow is freely available and can be found at https://github.com/kkpsiren/vesi/.

Declaration of interests

We have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Work by KS was funded by the Horizon 2020 Programme of the European Commission within the Marie Skłodowska-Curie Innovative Training Network "MicroWine" (grant number 643063). Work by JV was funded by the funding initiative "Plant Breeding Research for the Bioeconomy" under the National Research Strategy BioEconomy 2030 of the Federal Ministry of Education and Research (BMBF) and the Federal Ministry of Food and Agriculture (BMEL) of the Federal Republic of Germany (grant number 031B0197).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.acax.2019.100005.

References

[1] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: a review, Anal. Chim. Acta 914 (2016) 17–34.
[2] J. Boccard, J.-L. Veuthey, S. Rudaz, Knowledge discovery in metabolomics: an overview of MS data handling, J. Separ. Sci. 33 (2010) 290–304.
[3] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Metabolomics by numbers: acquiring and understanding global metabolite data, Trends Biotechnol. 22 (2004) 245–252.
[4] R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinf. 9 (2008) 504.
[5] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, M. Arita, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, Nat. Methods 12 (2015) 523–526.
[6] R. Spicer, R.M. Salek, P. Moreno, D. Cañueto, C. Steinbeck, Navigating freely-available software tools for metabolomics analysis, Metabolomics 13 (2017) 106.
[7] M.-E.P. Papadimitropoulos, C.G. Vasilopoulou, C. Maga-Nteve, M.I. Klapa, Untargeted GC-MS metabolomics, in: G.A. Theodoridis, H.G. Gika, I.D. Wilson (Eds.), Metabolic Profiling: Methods and Protocols, Springer New York, New York, NY, 2018, pp. 133–147.
[8] M.L. Santoru, C. Piras, A. Murgia, V. Palmas, T. Camboni, S. Liggi, I. Ibba, M.A. Lai, S. Orrù, S. Blois, A.L. Loizedda, J.L. Griffin, P. Usai, P. Caboni, L. Atzori, A. Manzin, Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients, Sci. Rep. 7 (2017) 9523.
[9] J.M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC-MS problems with PARAFAC2, Trends Anal. Chem. 27 (2008) 714–725.
[10] J. Vestner, G. de Revel, S. Krieger-Weber, D. Rauhut, M. du Toit, A. de Villiers, Toward automated chromatographic fingerprinting: a non-alignment approach to gas chromatography mass spectrometry data, Anal. Chim. Acta 911 (2016) 42–58.
[11] J.H. Christensen, J. Mortensen, A.B. Hansen, O. Andersen, Chromatographic preprocessing of GC–MS data for analysis of complex chemical mixtures, J. Chromatogr., A 1062 (2005) 113–123.
[12] M. Cocchi, C. Durante, M. Grandi, D. Manzini, A. Marchetti, Three-way principal component analysis of the volatile fraction by HS-SPME/GC of aceto balsamico tradizionale of modena, Talanta 74 (2008) 547–554.
[13] M. Daszykowski, R. Danielsson, B. Walczak, No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis–mass spectrometry, J. Chromatogr., A 1192 (2008) 157–165.
[14] N.A. Sinkov, J.J. Harynuk, Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling, Talanta 83 (2011) 1079–1087.
[15] D. Ballabio, T. Skov, R. Leardi, R. Bro, Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques, J. Chemom. 22 (2008) 457–463.
[16] J. Jaumot, N. Escaja, R. Gargallo, C. González, E. Pedroso, R. Tauler, Multivariate curve resolution: a powerful tool for the analysis of conformational transitions in nucleic acids, Nucleic Acids Res. 30 (2002) e92.
[17] M. Daszykowski, B. Walczak, Methods for the exploratory analysis of two-dimensional chromatographic signals, Talanta 83 (2011) 1088–1097.
[18] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309.
[19] L.G. Johnsen, J.M. Amigo, T. Skov, R. Bro, Automated resolution of overlapping peaks in chromatographic data: chromatographic data analysis, J. Chemom. 28 (2014) 71–82.
[20] X. Domingo-Almenara, A. Perera, J. Brezmes, Avoiding hard chromatographic segmentation: a moving window approach for the automated resolution of gas chromatography-mass spectrometry-based metabolomics signals by multivariate methods, J. Chromatogr., A 1474 (2016) 145–151.
[21] X. Domingo-Almenara, J. Brezmes, M. Vinaixa, S. Samino, N. Ramirez, M. Ramon-Krauel, C. Lerin, M. Díaz, L. Ibáñez, X. Correig, A. Perera-Lluna, O. Yanes, eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC/MS-based metabolomics, Anal. Chem. 88 (2016) 9821–9829.
[22] A. Smirnov, W. Jia, D.I. Walker, D.P. Jones, X. Du, ADAP-GC 3.2: graphical software tool for efficient spectral deconvolution of gas chromatography-high-resolution mass spectrometry metabolomics data, J. Proteome Res. 17 (2018) 470–478.
[23] L.A. Adutwum, R.J. Abel, J. Harynuk, Total ion spectra versus segmented total ion spectra as preprocessing tools for gas chromatography - mass spectrometry data, J. Forensic Sci. 63 (2018) 1059–1068.
[24] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinf. 11 (2010) 395.
[25] L. Han, Y.-M. Zhang, J.-J. Song, M.-J. Fan, Y.-J. Yu, P.-P. Liu, Q.-X. Zheng, Q.-S. Chen, C.-C. Bai, T. Sun, Y.-B. She, Automatic untargeted metabolic profiling analysis coupled with Chemometrics for improving metabolite identification quality to enhance geographical origin discrimination capability, J. Chromatogr., A 1541 (2018) 12–20.
[26] Y.-J. Yu, H.-Y. Fu, L. Zhang, X.-Y. Wang, P.-J. Sun, X.-B. Zhang, F.-W. Xie, A chemometric-assisted method based on gas chromatography-mass spectrometry for metabolic profiling analysis, J. Chromatogr., A 1399 (2015) 65–73.
[27] A. Trimigno, L. Münger, G. Picone, C. Freiburghaus, G. Pimentel, N. Vionnet, F. Pralong, F. Capozzi, R. Badertscher, G. Vergères, GC-MS based metabolomics and NMR spectroscopy investigation of food intake biomarkers for milk and cheese in serum of healthy humans, Metabolites 8 (2018) 26.
[28] H. Nikpour, M. Mousavi, H. Asadollahzadeh, Qualitative and quantitative analysis of Teucrium polium essential oil components by GC-MS coupled with MCR and PARAFAC methods, Phytochem. Anal. 29 (2018) 590–600.
[29] C. Chen, J. Husny, S. Rabe, Predicting fishiness off-flavour and identifying compounds of lipid oxidation in dairy powders by SPME-GC/MS and machine learning, Int. Dairy J. 77 (2018) 19–28.
[30] S. Taghadomi-Saberi, S. Mas Garcia, A. Allah Masoumi, M. Sadeghi, S. Marco, Classification of bitter orange essential oils according to fruit ripening stage by untargeted chemical profiling and machine learning, Sensors 18 (2018) 1922.
[31] A. Acharjee, Z. Ament, J.A. West, E. Stanley, J.L. Griffin, Integration of metabolomics, lipidomics and clinical data using a machine learning method, BMC Bioinf. 17 (2016) 440.
[32] Q. Yang, L. Xu, L.-J. Tang, J.-T. Yang, B.-Q. Wu, N. Chen, J.-H. Jiang, R.-Q. Yu, Simultaneous detection of multiple inherited metabolic diseases using GC-MS urinary metabolomics by chemometrics multi-class classification strategies, Talanta 186 (2018) 489–496.
[33] A. Smolinska, A.-C. Hauschild, R.R.R. Fijten, J.W. Dallinga, J. Baumbach, F.J. van Schooten, Current breathomics–a review on data pre-processing techniques and machine learning in metabolomics breath analysis, J. Breath Res. 8 (2014) 027105.
[34] X. Wang, D. Wang, J. Wu, X. Yu, J. Lv, J. Kong, G. Zhu, R. Su, Metabolic characterization of myocardial infarction using GC-MS-based tissue metabolomics, Int. Heart J. 58 (2017) 441–446.
[35] Z. Lai, H. Tsugawa, G. Wohlgemuth, S. Mehta, M. Mueller, Y. Zheng, A. Ogiwara, J. Meissen, M. Showalter, K. Takeuchi, T. Kind, P. Beal, M. Arita, O. Fiehn, Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics, Nat. Methods 15 (2018) 53–56.
[36] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.
[37] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2016, pp. 785–794.
[38] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 861–874.
[39] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Sattar A., Kang B. (eds) AI 2006: Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science, vol. 4304. Springer, Berlin, Heidelberg.
[40] J. Ambroise, A. Robert, B. Macq, J.-L. Gala, Transcriptional network inference from functional similarity and expression data: a global supervised approach, Stat. Appl. Genet. Mol. Biol. 11 (2012). Article 2.
[41] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference, 2006. https://www.biostat.wisc.edu/~page/rocpr.pdf. (Accessed 6 July 2018).
[42] P.D. Foster Provost, Well-Trained PETs: Improving Probability Estimation Trees, 2000. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.309. (Accessed 6 July 2018).
[43] M. Carbonero-Ruz, F.J. Martínez-Estudillo, F. Fernández-Navarro, D. Becerra-Alonso, A.C. Martínez-Estudillo, A two dimensional accuracy-based measure for classification performance, Inf. Sci. 382–383 (2017) 60–80.
[44] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[45] K. Tian, L. Wu, S. Min, R. Bro, Geometric search: a new approach for fitting PARAFAC2 models on GC-MS data, Talanta 185 (2018) 378–386.
[46] P. Software Foundation, Python Language Reference, 2017, Version 3.6. 3.
[47] E. Jones, T. Oliphant, P. Peterson, {SciPy}: Open Source Scientific Tools for {Python}, 2001. http://www.scipy.org.

[48] T. Kolda, B. Bader, Tensor decompositions and applications, SIAM Rev. 51 (2009) 455–500.
[49] J. Kossaifi, Y. Panagakis, A. Anandkumar, M. Pantic, TensorLy: Tensor Learning in Python, CoRR, 2018. http://arxiv.org/abs/1610.09555.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Others, scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[51] L. van der Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (2014) 3221–3245.
[52] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95.
[53] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D.C. Gemperline, T. Augspurger, Y. Halchenko, J.B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M.L. Williams, C. Evans, C. Fitzgerald, Brian, C. Fonnesbeck, A. Lee, A. Qalieh, mwaskom/seaborn: v0.8.1 (September 2017), 2017, https://doi.org/10.5281/zenodo.883859.
[54] K. Haug, R.M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J.L. Griffin, C. Steinbeck, MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data, Nucleic Acids Res. 41 (2013) D781–D786.
[55] B.-J. Webb-Robertson, Y.-M. Kim, E.M. Zink, K.A. Hallaian, Q. Zhang, R. Madupu, K.M. Waters, T.O. Metz, A statistical analysis of the effects of urease pre-treatment on the measurement of the urinary metabolome by gas chromatography-mass spectrometry, Metabolomics 10 (2014) 897–908.
[56] C. Hu, T. Tohge, S.-A. Chan, Y. Song, J. Rao, B. Cui, H. Lin, L. Wang, A.R. Fernie, D. Zhang, J. Shi, Identification of conserved and diverse metabolic shifts during rice grain development, Sci. Rep. 6 (2016) 20942.
[57] X. Liu, X.-H. Zhu, P. Qiu, W. Chen, A correlation-matrix-based hierarchical clustering method for functional connectivity analysis, J. Neurosci. Methods 211 (2012) 94–102.
[58] D. Müllner, fastcluster: fast hierarchical, agglomerative clustering routines for R and Python, J. Stat. Software (2013). https://www.jstatsoft.org/article/view/v053i09/v53i09.pdf.
[59] G. Bradski, The opencv library, Dr. Dobb's Journal of Software Tools (2000), 2000.
[60] N.E. Helwig, Multiway: Component Models for Multi-Way Data, 2018. https://CRAN.R-project.org/package=multiway.
[61] R Core Team, A Language and Environment for Statistical Computing, 2017. https://www.R-project.org/.
[62] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (2002) 18–22. http://CRAN.R-project.org/doc/Rnews/.
[63] M.B. Kursa, W.R. Rudnicki, Others, feature selection with the boruta package, J. Stat. Software 36 (2010) 1–13.
[64] L.G. Johnsen, P.B. Skou, B. Khakimov, R. Bro, Gas chromatography–mass spectrometry data processing made easy, J. Chromatogr., A 1503 (2017) 57–64.
[65] A. Rapp, H. Mandery, Wine aroma, Experientia 42 (1986) 873–884.
[66] I. Oseledets, D. Savostianov, E. Tyrtyshnikov, Tucker dimensionality reduction of three-dimensional arrays in linear time, SIAM J. Matrix Anal. Appl. 30 (2008) 939–956.
[67] B. Khakimov, R.J. Mongi, K.M. Sørensen, B.K. Ndabikunze, B.E. Chove, S.B. Engelsen, A comprehensive and comparative GC-MS metabolomics study of non-volatiles in Tanzanian grown mango, pineapple, jackfruit, baobab and tamarind fruits, Food Chem. 213 (2016) 691–699.
[68] N. Dalmau, C. Bedia, R. Tauler, Validation of the regions of interest multivariate curve resolution (ROIMCR) procedure for untargeted LC-MS lipidomic analysis, Anal. Chim. Acta 1025 (2018) 80–91.