Chromatographia
https://doi.org/10.1007/s10337-023-04301-z

ORIGINAL

A Complete Analysis Pipeline for the Processing, Alignment and Quantification of HPLC–UV Wine Chromatograms

Alan Ianeselli[1] · Edoardo Longo[2] · Simone Poggesi[3] · Marco Montali[1] · Emanuele Boselli[2]

Received: 21 September 2023 / Accepted: 10 November 2023

Abstract

Elucidating the chemistry of wine would help defining its quality, chemical and sensory characteristics and optimise the wine-making processes. High-performance liquid chromatography coupled with UV–Vis spectroscopy (HPLC–UV–Vis) is a common analysis method used to obtain the molecular profile of wine samples. We propose a complete procedure for the analysis of wine chromatograms. Data are pre-processed using standard methods of down-sampling, smoothing and baseline subtraction. Multiple samples are then merged in a three-dimensional tensor, decomposed using parallel factor analysis (PARAFAC2) into three factors: (i) one reduced (rank-one) chromatogram per sample, (ii) an estimate of the samples' spectral UV–Vis profile and (iii) an estimate of the samples' concentrations. If the decomposition is performed on a single peak of the tensor, the second and third factors correspond to the representative wavelength spectrum and to the relative concentrations of the samples, respectively. Otherwise, when multiple peaks are analysed, further processing is required. In the latter case, the decomposed rank-one chromatograms are peak-detected and aligned, clustered and integrated. A table containing the concentration of the peaks at different retention times is obtained. The pipeline proposed in this study is a guideline for a quantitative and reproducible chemical analysis of wine, or other samples, via the HPLC–UV–Vis method.

Keywords Chemometrics · Parallel factor analysis · Food chemistry · Data analysis · HPLC–UV–Vis · Wine

Introduction

To understand the chemistry of vinification at a molecular level, it is necessary to explore the chemical complexity of wine samples, also at different stages of the wine-making process [1–3]. Factors, such as grape variety, regionality, storage conditions and wine-making practices, must be taken into account [4–7].

The characterisation of the determinants of wine qualities is a challenging process, which requires advanced studies of correlation between the chemical profiles and sensory data, the modelling of a multitude of chemical parameters, and the complete characterisation of the environmental effects during the wine-making process [8–11]. Such association studies require a large mole of data at the chemical level, which are often generated in laboratories using advanced experimental instruments, for example HPLC–UV–Vis, mass spectrometry and NMR [12, 13]. The data analysis steps are very time-consuming and are currently the bottleneck of most of the studies in the field [14]. With this paper, we aim at developing a data analysis pipeline for HPLC–UV–Vis data, which combines standard processing techniques with advanced mathematical methods, in order to simplify the large mole of data and obtain the most important characteristics of each sample.

The HPLC–UV–Vis raw data are high-dimensional (absorbance intensity for n° samples × retention

✉ Alan Ianeselli
alan.ianeselli@unibz.it

✉ Marco Montali
marco.montali@unibz.it

✉ Emanuele Boselli
emanuele.boselli@unibz.it

Edoardo Longo
edoardo.longo@unibz.it

Simone Poggesi
s.poggesi@massey.ac.nz

[1] Faculty of Engineering, Free University of Bozen-Bolzano, Bolzano, Italy

[2] Oenolab, Faculty of Science and Technology, Free University of Bozen-Bolzano, Bolzano, Italy

[3] Food Experience and Sensory Testing (Feast) Lab, Massey University, Palmerston North 4410, New Zealand

Published online: 17 January 2024

time × wavelength), and therefore contain a very large amount of information, making data analysis and visualisation challenging. Chromatograms sometimes present problematics, such as high noise from the instrument, inconsistent retention times shifts between the samples, overlapping peaks, and high baseline level [15–17]. Even though there are many studies that describe the analysis of chromatograms and its issues, clearly defined steps and a generally shared analysis pipeline is still missing [18–21].

Here, we present an analysis procedure for a complete and accurate quantification of the chemical profile of wine chromatograms (but also applicable to samples of different nature). Note that all the data shown in the figures have been obtained from real HPLC–UV–Vis data of Pinot Blanc samples (Figs. 1, 2, 4) or from calibration standards (phenols at known concentrations, Fig. 3), measured as indicated in "Materials and methods".

General Description of the Analysis Pipeline

The raw data coming out of an HPLC–UV–Vis instrument are multi-dimensional (Fig. 1a). Each sample contains the absorbance intensity (in arbitrary units) over time (e.g. minutes) and for different wavelengths (for example, from 200 to 800 nm). The raw chromatograms (also called elution profiles) are often noisy, depending on the sensibility of the instrument. The solvent can also give unwanted contributions to the absorbance, increasing the baseline level and modifying the shape of the overall chromatogram. At this step, pre-processing data analysis techniques are required to address such issues. Chromatograms are down-sampled (eventually, when resolution is too high), the signals smoothed, and the baseline signal is subtracted from the data.

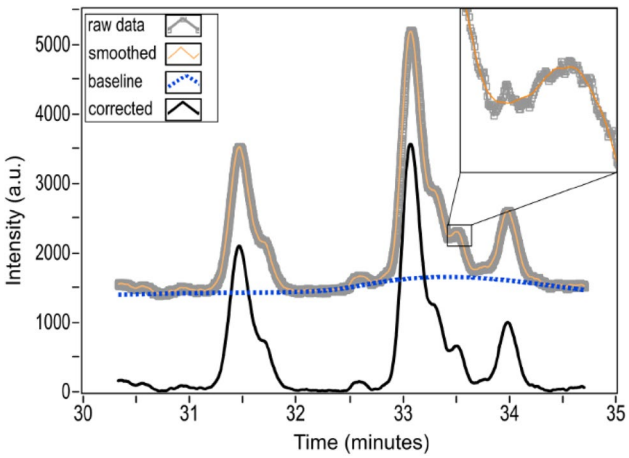

Fig. 2 Smoothing and baseline subtraction. The raw HPLC–UV–Vis data are smoothed with the Savitzky–Golay filter, using a polynomial function of order = 3 and a window size of 51. The baseline is estimated using the baseline asymmetric least squares smoothing algorithm with smoothness = 1e+7 and asymmetry = 2e−5. In this plot we show the elution profile at 260 nm, but the process is repeated for every wavelength

The pre-processed chromatograms of multiple samples are then merged to create a tensor (Fig. 1b). A factor decomposition technique called PARAFAC 2 is used to decompose the high-order array (i.e. the tensor) into a smaller number of factors. One factor is a bi-dimensional array containing a reduced (mono-dimensional) elution profile per sample. The other factors are two vectors. When the analysis is performed on single peaks (i.e. single chemical species) one vector directly represents the samples' UV–Vis spectrum, and the other represents the

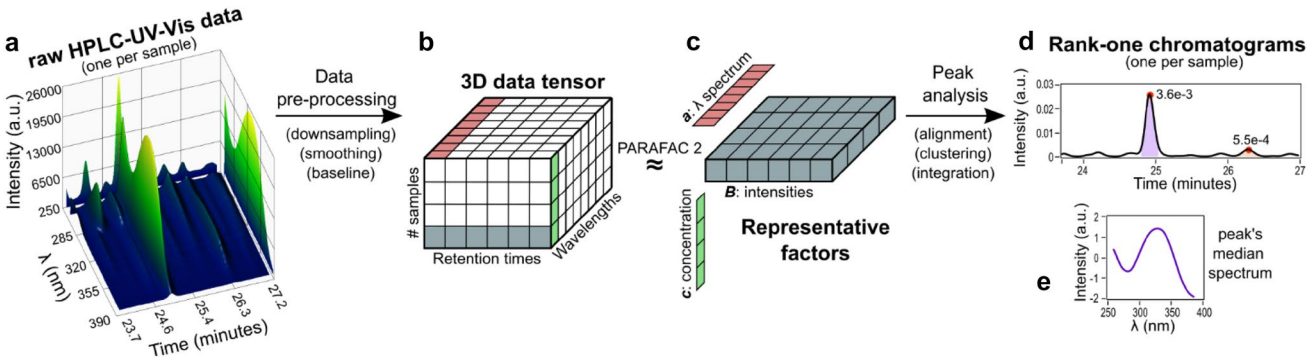

Fig. 1 General description of the analysis pipeline. The procedure consists of data pre-processing, tensor decomposition and peak analysis. **a** The raw HPLC–UV–Vis data of each sample are three-dimensional. They undergo pre-processing steps that down-sample the data, remove the noise and subtract the signal baseline. **b** The data from different samples are merged to create a high-order tensor. Parallel factor analysis (PARAFAC 2) decomposes the tensor into a smaller number of factors, representative of the samples' UV–Vis spectrum, elution profiles, and concentration. **d** The reduced (rank-one) elution profiles of multiple peaks undergo further analysis to align, cluster, fit, integrate and normalise the peaks corresponding to each chemical species

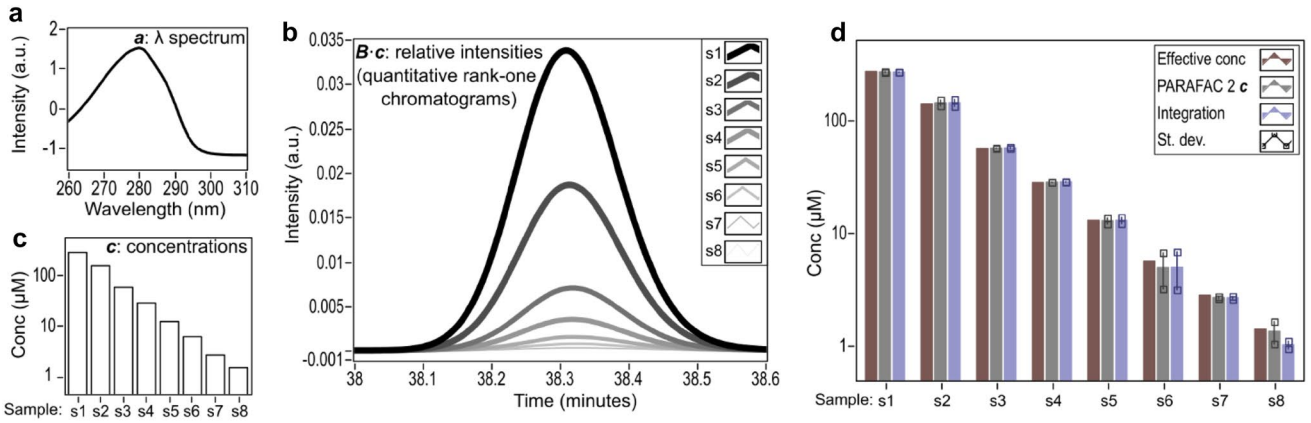

Fig. 3 Decomposition of epicatechin calibration standards using PARAFAC2. When single peaks are analysed, the factors obtained through tensor decomposition directly represent the constituent elements of each dimension of the tensor dataset. **a** Factor ***a***: vector corresponding to the representative λ spectrum of epicatechin in the range 260–310 nm. The data have been SNV corrected. **b** Factor ***B***: array containing the representative elution profiles of epicatechin for each sample. In this display, the array ***B*** has been multiplied with the vector ***c*** in order to restore the relative magnitude of each signal. **c** Factor ***c***: vector corresponding to the concentration of epicatechin between the samples. Data are shown in μM. **d** Comparison of the quantification of the calibration standards for epicatechin by parallel factor analysis or by peak integration. Standard deviation was calculated from duplicates

concentration profile of the molecule across the samples (Fig. 1c).

Otherwise, when multiple peaks are analysed, further processing of the peaks is required. The position of the peaks is detected, signals are aligned and peaks are clustered according to their retention times. The number of peaks of each sample and the median UV–Vis spectrum are calculated. After peak fitting, integration and normalisation, the area (i.e. relative concentration) of each peak is quantified and a peak table is obtained (Fig. 1d, e).

The next steps require a qualitative evaluation from an expert eye to assign each peak to a specific molecule or class of molecules, based on the information that have been obtained: elution profile, retention time, concentration, spectrum.

Results and Discussion

Data Pre-processing

The HPLC–UV–Vis raw data have been smoothed with the Savgol filter algorithm [22–24]. It calculates a polynomial fit of low order for successive subsets of adjacent data points (of arbitrary window size), as it moves across the signal. The filter estimate at the centre of each window is obtained by the polynomial fit at that central point. The free parameters of this method are the polynomial order (set to 3, in our case) and the window size (which we varied between 15 and 151, depending on the noise of each individual sample). We have smoothed the elution profile over the time axis, for every wavelength. The results of the smoothing, for a single wavelength, are shown in Fig. 2 (grey and orange lines and zoomed insert).

The mean absolute percentage error (MAPE) metric [25], calculated between the original and the smoothed data, was used to evaluate the performance of the smoothing process. A MAPE between 0.3% and 1.0% was an indicator of a good smoothing performance. A very low MAPE (< 0.2%) indicates overfitting so an unsuccessful smoothing, whilst a too high MAPE (> 1.5%) indicates under-fitting and implies that the original shape of the data has been drastically modified (Supplementary section 1).

Other smoothing methods would yield similar results. For example, moving average [26], exponential smoothing [26], smoothing spline [27], fast Fourier transform [28], low-pass filter [29] are only some of the common available methods for data smoothing.

Then, we performed the subtraction of the baseline from the smoothed signals, again over the time axis for every wavelength. We have used the baseline asymmetric least squares smoothing algorithm [30] (AsLS), which estimates the baseline by second derivative constrained weighted regression. There are two parameters: asymmetry and smoothness. Since there is not a defined way to obtain the best values of the parameters, one has to check by trial and error. A visual inspection is often sufficient to get good values. In our case, good results were obtained with a smoothness value in the range 1e+7–1e+6, and an asymmetry value between 2e−4 and 2e−5, depending on the individual samples. Results are shown in Fig. 2 (blue dashed line). Supplementary section 1 shows how different parameter combinations affect the baseline subtraction performance.

Some other baseline subtraction techniques that could have been used alternatively are, for example, least squares fit [31], multivariate background correction [32], temporal median filter [33], and kernel density estimation [34].

Tensor Decomposition

The smoothed and baseline subtracted data of each sample are then merged to create a 3-way tensor (a multidimensional matrix) of absorbance intensity for n° samples (n) × retention time (t) × wavelength (w) (Supplementary section 2). At this point, we decomposed the tensor using PARAFAC2 [35–38], in order to break it into its constituent elements (factors). The decomposition process can be formulated as follows:

$$\boldsymbol{X}_{n,t,w} \approx \sum_{i=1}^{r} \boldsymbol{a}_i \times \boldsymbol{B}_i \times \boldsymbol{c}_i$$

where $\boldsymbol{X}_{n,t,w}$ is the 3-way tensor, r is the rank of the decomposition, $\boldsymbol{a}$ is a vector of length w, $\boldsymbol{B}$ is an array of shape $n \times t$, and $\boldsymbol{c}$ is a vector of length n.

To test the method, we have initially applied the PARAFAC2 tensor decomposition method on a set of calibration data of phenols at known concentration, in duplicate. This has allowed us to gain a deep understanding on what $\boldsymbol{a}$, $\boldsymbol{B}$ and $\boldsymbol{c}$ really represent in the context of our dataset. Results are shown in Fig. 3, for the phenol molecule epicatechin. Note that the following considerations are valid only when one single peak of the tensor is analysed at a time.

Factor $\boldsymbol{a}$ is a vector of length w, and contains one intensity value per wavelength, in our case in the range of 260–310 nm. It corresponds to the absorbance λ spectrum of the overall elution profile across the samples, which ultimately corresponds to the representative spectrum of the molecule (epicatechin). It is shown in Fig. 3a after SNV (standard normal variate) correction [39].

Factor $\boldsymbol{B}$ is an array of values of shape $n \times t$, and contains one signal per sample. Every signal corresponds to the elution profile across the wavelengths, for each sample. This corresponds to the representative rank-one elution profile of the molecule (epicatechin). The elution profiles are shown in Fig. 3b, multiplied with the vector $\boldsymbol{c}$ in order to restore the relative magnitude of each signal (as discussed next).

Factor $\boldsymbol{c}$ is a vector of length n, and contains one intensity value per sample (in our case eight samples). Every value corresponds to the quantification of the overall elution profile across the wavelengths. Ultimately, this corresponds to the concentration of the molecule (epicatechin) in each sample. Results (in µM) are shown in Fig. 3c.

According to the formula shown before, it is possible to reconstruct the original tensor from the decomposition factors $\boldsymbol{a}$, $\boldsymbol{B}$ and $\boldsymbol{c}$. This can be done in order to estimate the accuracy of the decomposition method. In this way, we could estimate the error of our model by calculating the MAPE between the original and the back-calculated tensors. In the case of epicatechin shown in Fig. 3, the average MAPE resulted to be 0.8%. Similar errors have been obtained when analysing the calibration data of the other phenols (data not shown).

We investigated if the decomposition procedure is quantitative and if it maintains the relative concentrations between the samples. Calibration standards of five different phenols (gallic acid, protocatechuic acid, catechin, epicatechin, coumaric acid) with known initial concentrations have been measured by HPLC–UV–Vis, in duplicate. Data have been pre-processed and decomposed following the procedure presented before. As control, we compared the decomposition analysis with the standard quantification of the peaks by integration.

Results are shown in Fig. 3d for epicatechin. Effective initial (known) concentrations are shown in brown bars. Quantification by PARAFAC2 (vector $\boldsymbol{c}$) is shown in grey bars. As previously explained, the factor $\boldsymbol{c}$ is a vector that directly contains the estimated concentrations of the reduced (rank-one) elution profiles, one for each sample. The blue bars correspond to the control quantification by peak integration. The PARAFAC2 $\boldsymbol{c}$ follows the effective concentrations very accurately, and with similar uncertainty (black lines, standard deviation) to the peak integration method.

Table 1 compares the accuracy of parallel factor analysis vs integration, by means of MAPE from the effective (known) concentration. The quantification of the peaks by PARAFAC2 $\boldsymbol{c}$ has an average MAPE of 6.9% from the effective known concentration, which is a fairly acceptable error. Moreover, it is substantially identical to the MAPE obtained through the classical peak integration method (7.0%). Note that the MAPE from the effective

Table 1 Accuracy of the quantification methods

Molecule	Retention time (min)	MAPE from effective conc (%)	
		PARAFAC2	Integration
Gallic acid	17.1	7.4	7.3
Protocatechuic acid	24.8	8.0	7.5
Catechin	33.0	5.6	5.8
Epicatechin	38.3	6.7	7.5
Coumaric acid	45.1	6.8	6.8

The calibration standards of five phenols at known concentrations have been quantified by PARAFAC2 or integration. The last two columns indicate the MAPE from the effective known concentrations. Note that the value shown here also implicitly includes other error sources, such as the manual pipetting errors during sample preparation and the accuracy of the instrument (HPLC–UV–Vis)

concentration for both methods are likely to be overestimated, since they also implicitly include laboratory error sources during the preparation of the samples (e.g. pipetting errors) and the intrinsic uncertainty of the measurement instruments.

This results indicate that peak analysis by PARAFAC2 can be successfully employed as a quantification method for the molecular analysis of HPLC–UV–Vis chromatograms.

Peak Detection, Alignment and Clustering

When multiple peaks at different retention times are to be analysed, further processing is required. This is because, in this case, the vectors $\boldsymbol{a}$ and $\boldsymbol{c}$ cannot simply correspond to the representative λ spectrum and concentration profile, respectively, as we discussed in the previous section for single peaks. The matrix $\boldsymbol{B}$, instead, still contains the representative elution profile of each sample. The further analysis of the peak is then performed on these reduced chromatograms (matrix $\boldsymbol{B}$). Their reduced dimensionality (rank-one) strongly simplifies the analysis procedure.

The matrix $\boldsymbol{B}$, therefore, contains one rank-one chromatogram per row (i.e. per sample), which correspond to the non-aligned chromatograms of each sample (as shown in Fig. 4a). In order to cluster the peaks of the components by retention time, the chromatograms have to be aligned first. To properly align them, it is important to first calculate the position of the peaks. Peak positions were identified by a simple comparison of the neighbouring values, and were defined as any position whose direct neighbours have a smaller amplitude [40]. A manually defined height point was set to exclude the peaks below a specific amplitude, in order to ignore the fluctuations at the bottom of the chromatogram. For the peak detection to be successful, the pre-processing steps of data smoothing and baseline subtraction were essential. Otherwise, false positive peaks could be detected, or the location of the peak could be slightly off.

Due to instrumental inaccuracies, the elution profiles of different samples might have unpredictable elution time shifts, which can range from seconds to even minutes (e.g. Fig. 4a). It is therefore necessary, when possible, to align the signals in order to be able to classify the molecules based on their respective elution time.

The elution profiles have been aligned with the *msalign* algorithm [41–43], which aligns peaks in signals to reference peaks. It builds a synthetic Gaussian at the reference positions (given as input parameter). The signals are then shifted until the cross-correlation between the original signal and the synthetic signal is maximised. The process is then repeated for each sample. For this step, we therefore used the peak positions obtained in the previous step, to give as input parameter for the *msalign* algorithm. Results are shown in Fig. 4a, b.

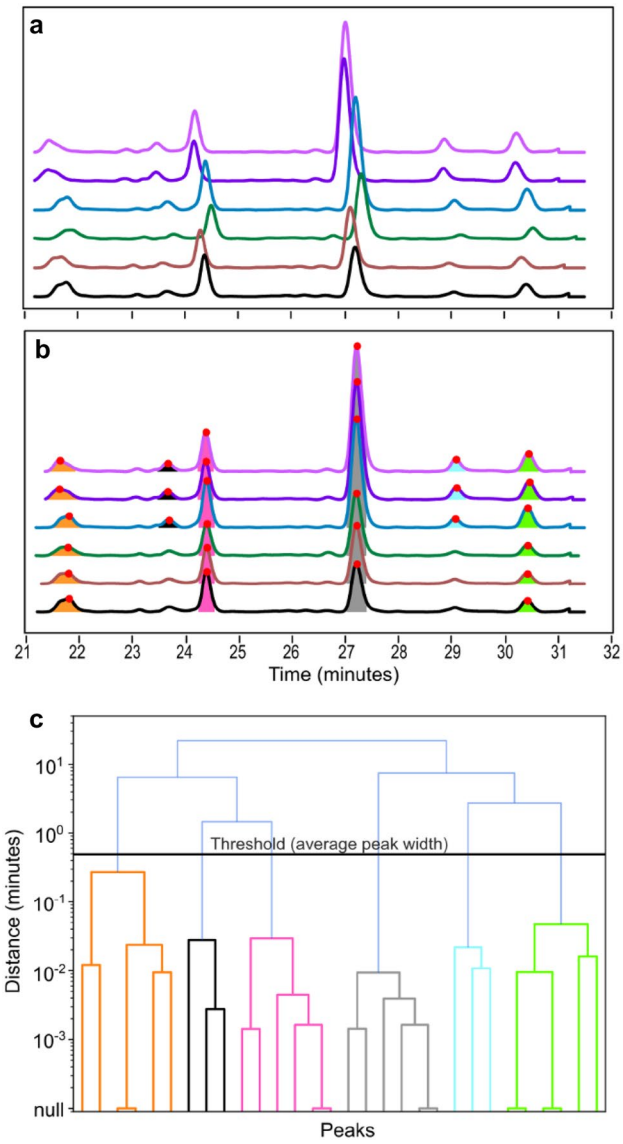

Fig. 4 Alignment and clustering of the peaks. **a** Misaligned rank-one chromatograms. Note that the data shown here have been artificially shifted randomly. **b** Aligned chromatograms via the *msalign* algorithm, based on the peaks previously identified (red dots). **c** Dendrogram of the agglomerative clustering of the peaks after alignment. The peaks belonging in the same cluster are shown with the same colour. The horizontal black line corresponds to the distance threshold utilised for cluster identification (average peak width)

In order to automatically group the peaks by retention time, in a unsupervised manner, we clustered the aligned peak positions using the algorithm of hierarchical agglomerative clustering [44]. Starting from individual instances, agglomerative clustering connects the nearest pairs of clusters, until all clusters are merged into one big cluster which contains all the objects [45]. The result is a tree-based

Table 2 Integrated peaks

Peak (min)	Area (a.u.)					
	r1	r2	r3	r4	r5	r6
21.7	5.7e−4	4.8e−4	4.0e−4	4.6e−4	4.7e−4	4.1e−4
23.7	–	–	–	2.1e−4	2.1e−4	2.2e−4
24.4	7.9e−4	7.1e−4	6.4e−4	9.3e−4	7.7e−4	7.9e−4
27.2	1.2e−3	1.4e−3	1.5e−3	2.6e−3	2.9e−3	3.0e−3
29.1	–	–	–	2.2e−4	3.0e−4	2.9e−4
30.4	3.1e−4	2.7e−4	2.8e−4	5.2e−4	4.7e−4	4.8e−4

Quantification of the peaks by integration, for the chromatograms r1–r6 (columns 2–7, colour code) shown in Fig. 4b. The six peaks are ordered by retention time. Peaks that are absent in a specific sample are marked with the "–" symbol

representation called dendrogram. We applied the clustering algorithm in an unsupervised manner i.e. without specifying the number of clusters to be found. The distance between the pairs was set to Euclidean. The linkage distance threshold (the distance threshold at or above which clusters will not be merged) was set to the average width of the peaks, calculated on all the peaks that were found in the previous step. Figure 4c presents the dendrogram. The dendrogram was then cut to a specific height, again the average width of the peaks, in order to obtain the number of clusters and the cluster labels.

Peak Fitting and Integration

After the clustering of the retention times, the concentration of the chemical species can be quantified by peak integration. First, each peak is fitted with a Gaussian function of the form α*exp(−(x − μ)^2/(2*σ^2)), where μ and σ represent the mean and the standard deviation, respectively, and α indicates the amplitude of the Gaussian function. The peak is then integrated at a confidence interval of 95% (2σ). Overlapping peaks require additional analysis, because they cannot be automatically resolved and clustered. They can be fitted by multiple Gaussians and then separately integrated, as indicated in the Supplementary section 3.

At the end, we obtain a peak table containing the concentration of each peak ordered and clustered by retention time, as shown in Table 2. The median spectrum of each peak (after SNV correction) can give further information for the identification of the specific chemical species.

Conclusions

HPLC–UV–Vis data are large and complex and require numerous steps of analysis before being able to extract qualitative and quantitative observations. Initial pre-processing steps, such as smoothing and baseline subtraction, are necessary to remove instrumental artefacts (e.g. noise) and improve the data quality. Then, PARAFAC 2 decomposition was used to reduce the dimensionality of the data, yielding one chromatogram per sample, thus strongly facilitating the successive analyses. Moreover, the decomposition factors were representative of the sample concentrations and absorbance spectrum. After the alignment of the signals and the detection of the peaks, the agglomerative clustering algorithm granted an automatic, unsupervised grouping of the peaks based on their retention time. This facilitated the identification of the molecular species corresponding to each peak, which could then be integrated and quantified.

The data analysis pipeline presented here is a complete procedure for the analysis of chromatograms of wine or other samples, measured via the HPLC–UV–Vis instrument. It is also not exclusive to a particular class of chemical compounds or analytical methods. It consists of standard routines for data processing as well as advanced mathematical methods and algorithms for data decomposition and clustering. It has been demonstrated that this pipeline quantifies the concentration of calibration standards with good accuracy and reproducibility. We hope that this paper can simplify this type of analysis and help other scientists and analysts, in order to set new analysis standards and enhance the reproducibility of the results in the field.

Materials and Methods

The HPLC separation was carried out in gradient mode according to a published procedure [7] on an ODS column (Eurosphere II, C18 stationary phase, 250 × 4.6 mm × 5 µm, Knauer, LabService Analytica, Anzola dell'Emilia, Bologna, Italy) using a Nexera X2 UHPLC (Shimadzu, Milano, Italy) equipped with a UV–Vis PDA detector (sampling rate 12.5 Hz, time constant = 0.320 s, scan range = 200–800 nm, 1.2 nm slit width) and a fluorescence detector (FLD, 10 Hz sampling rate, λ_{ex} = 276 nm, λ_{em} = 316 nm, with 1 × gain) used in series. The mobile phases were: solvent A 0.1% formic acid in degassed milliQ water; solvent B 0.1%

formic acid in acetonitrile. The gradient programme was: 0–2.5 min 1% B, 2.5–50 min 1–25% B, 50–51 min 25–99% B, 51–55 min 99% B, 55–56 min 99–1% B, 56–60 min 1% B. A constant 0.7 $mL \cdot min^{-1}$ flow rate was applied.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10337-023-04301-z.

Acknowledgements We thank Paolo Silos Labini and Dr. Paola Lecca from the Free University of Bozen-Bolzano for the support in the preparation of the parallel factor analysis routine.

Author contributions AI wrote the paper, analysed the results and wrote the analysis software. EL and SP provided the data and made the HPLC–UV–Vis experiments. AI, EL and SP determined the analysis pipeline, the quantification and validation procedures. All authors designed the project, set the methods and identified the relevance of the research.

Funding Open access funding provided by Libera Università di Bolzano within the CRUI-CARE Agreement. The presented work was carried out as part of the project "Wine Identity Card" (acronym; WineID; grant number: TN201A; ID 2019 call, funder: Free University of Bozen/Bolzano).

Data availability The data that support the findings of this study are available from the corresponding author, AI, upon reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Arapitsas P, Speri G, Angeli A, Perenzoni D, Mattivi F (2014) The influence of storage on the "chemical age" of red wines. Metabolomics 10:816–832
2. Robinson J, Harding J, Vouillamoz J (2013) Wine grapes: a complete guide to 1368 vine varieties, including their origins and flavours. Penguin Books Ltd
3. Waterhouse AL, Sacks GL, Jeffery DW (2016) Understanding wine chemistry. Wiley, Hoboken, NJ. https://doi.org/10.1002/9781118730720
4. Poggesi S, Merkytė V, Longo E, Boselli E (2022) Effects of microvibrations and their damping on the evolution of pinot noir wine during bottle storage. Foods 11:2761
5. Crandles M, Wicks-Müller M, Schuessler C, Jung R (2016) The effect of simulated transportation conditions on the chemical, physical and sensory profiles of Müller-Thurgau and Scheurebe wines. J Food Sci Eng 6:177–196
6. Soares De Andrade RH et al (2013) Anthocyanic composition of Brazilian red wines and use of HPLC-UV-Vis associated to chemometrics to distinguish wines from different regions. Microchem J 110(3):256–262
7. de Matos AD et al (2020) Pinot blanc: impact of the winemaking variables on the evolution of the phenolic, volatile and sensory profiles. Foods 9:499
8. Poggesi S et al (2022) Fusion of 2DGC-MS, HPLC-MS and sensory data to assist decision-making in the marketing of international monovarietal Chardonnay and Sauvignon blanc wines. Foods 11:3458
9. Vinci G, Maddaloni L, Prencipe SA, Ruggieri R (2021) Natural contaminants in wines: determination of biogenic amines by chromatographic techniques. Int J Environ Res Public Heal 18:10159
10. Eunicia M, Skyszygfrid R, Vitri T, Caren V (2022) Modeling red wine quality based on physicochemical tests: a data mining approach. Formosa J Multidiscip Res 1:89–110
11. Guerrini L et al (2022) Kinetic modeling of a Sangiovese wine's chemical and physical parameters during one-year aging in different tank materials. Eur Food Res Technol 248:1525–1539
12. Önal A, Tekkeli SEK, Önal C (2013) A review of the liquid chromatographic methods for the determination of biogenic amines in foods. Food Chem 138:509–515
13. Guasch J, Busto O (2000) Wine: gas and liquid chromatography. In: Encyclopedia of separation science. Elsevier. pp 4490–4498. https://doi.org/10.1016/B0-12-226770-2/01181-9
14. Khakimov B, Gürdeniz G, Engelse SB (2015) Trends in the application of chemometrics to foodomics studies. Acta Aliment 44:4–31
15. Sousa PFM, de Waard A, Åberg KM (2018) Elucidation of chromatographic peak shifts in complex samples using a chemometrical approach. Anal Bioanal Chem 410:5229
16. Windig W, Phalp JM, Payne AW (1996) A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. Anal Chem 68:3602–3606
17. Gil García MD et al (1997) Resolution of overlapping peaks in HPLC with diode array detection by application of partial least squares calibration to cross-sections of spectrochromatograms. Anal Chim Acta 348:177–185
18. Di Guida R et al (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12:93
19. Brereton RG (2013) The evolution of chemometrics. Anal Methods 5:3785–3789
20. Beisken S, Earll M, Portwood D, Seymour M, Steinbeck C (2014) MassCascade: visual programming for LC-MS data processing in metabolomics. Mol Inform 33:307
21. Liggi S et al (2018) KniMet: a pipeline for the processing of chromatography–mass spectrometry metabolomics data. Metabolomics 14:52
22. Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36:1627–1639
23. Gallagher NB. Savitzky-Golay smoothing and differentiation filter
24. scipy.signal.savgol_filter—SciPy v1.9.3 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html
25. de Myttenaere A, Golden B, Le Grand B, Rossi F (2016) Mean absolute percentage error for regression models. Neurocomputing 192:38–48
26. Nugus S (2009) Smoothing techniques. In Financial planning using excel. pp 47–58. https://doi.org/10.1016/B978-1-85617-551-7.00004-5

27. Craven P, Wahba G (1978) Smoothing noisy data with spline functions—estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403
28. Kosarev EL, Pantos E (1983) Optimal smoothing of 'noisy' data by fast Fourier transform. J Phys E 16:537
29. Kaiser JF, Reed WA (2008) Data smoothing using low-pass digital filters. Rev Sci Instrum 48:1447
30. Paul HC, Eilers HFMB (2005) Baseline correction with asymmetric least squares smoothing. Leiden Univ Med Cent Rep 1:5
31. Statham PJ (1977) Deconvolution and background subtraction by least-squares fitting with prefiltering of spectra. Anal Chem 49:2149–2154
32. Boelens HFM, Dijkstra RJ, Eilers PHC, Fitzpatrick F, Westerhuis JA (2004) New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. J Chromatogr A 1057:21–30
33. Lo BPL, Velastin SA (2001) Automatic congestion detection system for underground platforms. In: Proc. 2001 Int. Symp. Intell. Multimedia, Video Speech Process. ISIMP 2001. pp 158–161. https://doi.org/10.1109/ISIMP.2001.925356
34. Elgammal A, Harwood D, Davis L (2000) Non-parametric model for background subtraction. Lect Notes Comput Sci 1843:751–767
35. Yu H, Bro R (2021) PARAFAC2 and local minima. Chemom Intell Lab Syst 219:104446
36. Kiers HAL, ten Berge JMF, Bro R (1999) PARAFAC2—part I. A direct fitting algorithm for the PARAFAC2 model. J Chemom 13:175–294
37. Demonstration of PARAFAC2 — TensorLy: Tensor Learning in Python. http://tensorly.org/dev/auto_examples/decomposition/plot_parafac2.html
38. Kossaifi J, Panagakis Y, Anandkumar A, Pantic M. tensorly. decomposition.Parafac2. http://tensorly.org/stable/modules/generated/tensorly.decomposition.Parafac2.html
39. Barnes RJ, Dhanoa MS, Lister SJ (2016) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl Spectrosc 43:772–777
40. Signal processing (scipy.signal)—SciPy v1.9.3 Manual. https://docs.scipy.org/doc/scipy/reference/signal.html
41. Monchamp P, Andrade-Cetto L, Zhang JY, Henson R (2007) Signal processing methods for mass spectrometry. In: System bioinformatics: an engineering case-based approach. Artech House Publishers. pp 101–124
42. Kumar K, Espaillat A, Cava F (2017) PG-Metrics: a chemometric-based approach for classifying bacterial peptidoglycan data sets and uncovering their subjacent chemical variability. PLoS ONE 12:e0186197
43. Migas L (2022) msalign: signal calibration and alignment by reference peaks. https://github.com/lukasz-migas/msalign
44. sklearn.cluster.AgglomerativeClustering—scikit-learn 1.2.0 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
45. Nielsen F (2016) Hierarchical clustering. In: Introduction to HPC with MPI for data science. Springer, Cham, pp 195–211. https://doi.org/10.1007/978-3-319-21903-5_8