Using clustering as pre-processing in the framework of signal unmixing for exhaustive exploration of archaeological artefacts in Raman imaging.

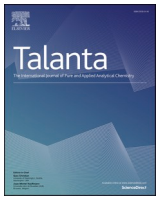
M. Offroy, Mario Marchetti, T.H. Kauffmann, P. Bourson, Ludovic Duponchel, Laurent Savarese, Jean-Michel Mechling

Talanta 274 (2024) 125955

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta



Using clustering as pre-processing in the framework of signal unmixing for exhaustive exploration of archaeological artefacts in Raman imaging

Marc Offroy [a,*], Mario Marchetti [b,f], Thomas H. Kauffmann [c], Patrice Bourson [c], Ludovic Duponchel [d], Laurent Savarese [e], Jean-Michel Mechling [f]

[a] Université de Lorraine, CNRS, LIEC, F-54000, Nancy, France
[b] Université Gustave Eiffel, MAST, FM2D, IFSTTAR, 14-20 Boulevard Newton, Cité Descartes, Champs sur Marne, F-77447, Marne La Vallée Cedex 2, France
[c] Université de Lorraine, CentraleSupelec, LMOPS, F-57000, Metz, France
[d] Université de Lille, CNRS, UMR 8516, LASIRE, 59000, Lille, France
[e] Centre de Recherches Archéologiques de Ruscino, Ville de Perpignan, Chercheur Associé UMR 5140 TESAM, France
[f] Université de Lorraine, CNRS, IJL, F-54000, Nancy, France

ARTICLE INFO

Handling editor: Prof. J.-M. Kauffmann

Keywords:
Raman imaging
Archaeology
Chemometrics
Biogenic materials
Threshold-based clustering algorithm
OPA
MCR-ALS
Signal unmixing

ABSTRACT

Analytical chemistry on archaeological material is an essential part of modern archaeological investigations and from year to year, instrumental improvement has made it possible to generate data at a high spatial and temporal frequency. In particular, Raman spectral imaging can be successfully applied in archaeological research by its simplicity of implementation to study past human societies through the analysis of their material remains. This technique makes it possible to simultaneously obtain spatial and spectral information by preserving sample integrity. However, because of the inherent complexity of the samples in Archaeology (e.g. seniority, fragility, lack or full absence of any information about its composition), chemical interpretation can be difficult at first glance. Indeed, specific problems of spectral selectivity related to unexpected chemical compounds could appear due to their state of conservation. Furthermore, detecting minor compounds becomes challenging as major components impose their contributions in the acquired spectra. Therefore, a relevant chemometric approach has been introduced in this context to characterize distinct spectral sources in a Raman imaging dataset of an archaeological specimen – a mosaic fragment. The fragment was unearthed during the *Ruscino* archaeological dig on the outskirts of Perpignan, France. It dates back to the *oppidum* period. The aim is to extract selective spectral information from pixel clustering analysis in order to enhance the initial optimisation step within the Multivariate Curve Resolution and Alternating Least-Squares (MCR-ALS) algorithm, a well-known signal unmixing technique. The underlying principle of the MCR-ALS is that the acquired spectra can be expressed as linear combinations of pure spectra of all individual components present in the chemical system under study. Sometimes it can be difficult to obtain the desired results through the algorithm, particularly if initial estimates of spectral or concentration profiles are inaccurate due to complex signals, noise or lack of selectivity, resulting in rank deficiency (i.e. a poor estimation of the total number of pure signals). For this reason, an innovative threshold-based clustering algorithm, combined with multiple Orthogonal Projection Approaches (OPA), has been developed to improve matrix rank investigation and thus the initialisation step of the MCR-ALS approach before optimisation. The effective analysis of Raman imaging data for an archaeological mosaic played a crucial role in uncovering significant chemical information about a particular biogenic material. This insight sheds light on the origins of mortar manufacture during the *oppidum* period.

1. Introduction

Several reviews set out the basic principles of analytical techniques and their practical application to archaeological research [1,2], such as radiocarbon dating (or other isotopic forms) [3,4], spectroscopies [5,6] or elemental fingerprinting to deduce the elemental composition of artefacts [7,8], and also chromatography to identify small organic molecules [9,10]. Each analytical technique has certain advantages or

* Corresponding author.
E-mail address: marc.offroy@univ-lorraine.fr (M. Offroy).

https://doi.org/10.1016/j.talanta.2024.125955
Received 7 December 2023; Received in revised form 14 March 2024; Accepted 18 March 2024
Available online 26 March 2024

limitations depending on the sample analysed and the scientific question posed at the outset about a specific type of human activity. In addition, interpreting a significant amount of data from a wide range of available analytical methods can be a very complex task if an archaeological sample is to be fully inspected. The modern concept of archaeology is no longer just about finding or hunting for artefacts, but about conducting carefully designed, empirical and data-driven research to uncover the human behaviour and social processes behind the artefacts. The sample studied in this research comes from a ditch fill found in the archaeological excavation at *Ruscino* on the outskirts of Perpignan, France. The site was occupied by the Romans from the second half of the 2nd century BC to the end of the 1st century AD (i.e. *oppidum* period). This archaeological artefact is a fragment of a mosaic apparently fixed to a stone with mortar (see material and methods). The reality and origin of the mortar are enigmatic and have required detailed research and characterisations.

Raman spectral imaging was chosen because it does not require any specific preparation of samples [11], and therefore preserves its intrinsic nature, which can be important in archaeology [2] specifically for further experimental analysis. In addition, spectroscopic imaging is a powerful technique for visualizing the spatial and spectral information of complex and heterogeneous samples [12]. Micro-Raman spectral imaging instruments use a confocal microscope coupled to a spectrometer with a data collection system. The microscope focuses and collects scattered (or unabsorbed) photons from a specific mode of vibration. The spatial distribution for a particular chemical compound in the sample under consideration is usually deduced by a simple signal integration. In fact, the aim of hyperspectral imaging is to isolate a specific wavelength (i.e. selectivity) for a particular chemical component in order to deduce its spatial distribution [13]. However, this classical approach has several drawbacks that scientists need to bear in mind. Firstly, it is necessary to know a priori all the pure compounds in the sample being analysed. If this assumption is not verified, it would be possible to choose a non-selective spectral area and therefore overestimate the concentrations of a compound (i.e. the number of pixels generated for a zone of interest could be false). In other words, such conditions will generate biased chemical maps therefore not representative of the 'analytical' reality of the investigated sample. Secondly, it is impossible to identify a truly selective wavelength when there is a strong spectral overlap due, for example, to large bandwidths and/or the sample complexity. Finally, it would be difficult to detect, identify and produce chemical maps for unexpected compounds. Despite these difficulties, this classical signal integration approach is still widely used in hyperspectral imaging, as it provides rapid answers in the vast majority of cases. Nevertheless, the use of chemometrics is essential when the sample is complex and can benefit from hyperspectral imaging due to the amount of data that can be generated by the instrument (i.e. one spectrum per pixel) [14,15]. Chemometrics also has a number of pre-processing techniques for correcting certain variations in the data caused by chemical and/or physical interference, such as fluorescence emissions in Raman spectroscopy. In fact, the shading of the Raman signal by fluorescence is a limitation of Raman spectroscopy, particularly for studies of ancient materials [16]. The use of Raman spectroscopy to identify and study chemical compounds present in art objects, archaeology and conservation science has greatly increased in recent years. This is the case for pigment characterization [17,18], conservation-induced weathering/degradation processes of cultural heritage [17,19], palaeontology/prehistoric art [20,21], or forensic applications and authenticity research [22,23]. However, at the same time, the use of chemometrics within these fields is rare and when it is applied, it is limited to classical mathematical approaches such as Principal Component Analysis (PCA) [22,24], hierarchical clustering [25,26], Gaussian/Lorentzian regression [27,28] or other conventional statistical tools [29]. Nevertheless, the richness of archaeological samples and our desire to delve deeper into our inquiries now necessitate the development and/or the implementation of much more advanced methods [14,30] capable of addressing the scientific issues of both communities.

In this article, the major and minor spectral responses of materials present in the mosaic sample will be clearly identified after developing and implementing a signal unmixing pipeline based on clustering coupled with Multivariate Curve Resolution and Alternative Least-Square (MCR-ALS) approach [12,14,15]. MCR-ALS is a signal unmixing method based on bilinearity assumption. In other words, it assumes that the measured spectra are linear combinations of spectra of 'pure component' in the system under study. The term 'pure component' can either be a chemical compound or a mixture of chemical compounds whose concentrations correlate with a specific spectral signature. Therefore, the steps of the algorithm include the determination of the number of pure components present in the chemical system by (i) rank analysis methods, (ii) the generation of initial estimates for the concentration or the spectral profiles and (iii) a final iterative optimisation of the extracted profiles under suitable constraints. However, the results of MCR-ALS can sometimes be unsuitable for several reasons: (i) determining matrix rank is not straightforward. Indeed, careful analysis of percentages of variance explained using methods such as Principal Component Analysis (PCA) [31], Singular Value Decomposition (SVD) [32] or the Durbin-Watson criterion [33] depend on the signal-to-noise ratio. An erroneous estimation of the rank can lead to an -under- or -over- estimation of the pure components during the initialisation step of the MCR-ALS algorithm, thus leading to biased chemical images at the end of the optimisation process. (ii) On a mathematical level, the MCR-ALS is an optimisation problem involving a dependence on the starting solutions (initial guess) i.e. the more complex the sample studied, the more difficult it is to find a global minimum [32]. Consequently, it is possible to obtain correct solutions after studying the regression residuals, but these solutions are abysmal chemically speaking due to rotational ambiguities. To minimize this problem, constraints can be added to the MCR-ALS algorithm linked to the system under study [12,14,15,32,33] such as non-negativity of profiles, unimodality, or closure. Another method is to work on the initial selective variables (i.e. the initial estimate of the spectral and/or concentration profiles) for each putative compound in the complex sample [32,34,35] or to use spatial information from spectroscopic images based on local ranks [36–38]. The aim of the proposed chemometric approach is to use both ideas in Raman imaging to overcome, or at least to be less dependent on, potential flaws due to MCR-ALS.

2. Materials and methods

2.1. Description of the archaeological sample

The archaeological sample studied here corresponds to a mosaic fragment seemingly fixed to a stone i.e. sandy chalk (Fig. 1). The layer of tesserae is grouted with a very fine calcitic coating (also defined as 'bedding') and a few traces of tile mortar are present below, corresponding to the mosaic 'nucleus'. The top of the mosaic appears to be the perfect continuity of the sandy chalk and detailed observations of the interface show a succession of very thin layers that gradually fill in the volume between the tesserae and the stone. These layers, which could be interpreted as a mortar used to fix the stone on the mosaic, are no more than 500 μm thick. The reality and origin of this mortar are enigmatic and have required detailed investigations and characterizations. The blue box in Fig. 1 corresponds to the area analysed by the Raman spectral imaging instrument, which includes the mosaic, the glue, the sandy chalk and the bedding mortar.

2.2. Raman hyperspectral imaging

The Raman hyperspectral imaging dataset (Fig. S1 in supplementary material) is acquired by using the LabRAM Evolution HR confocal scanning spectrometer manufactured by Horiba Jobin Yvon Scientific

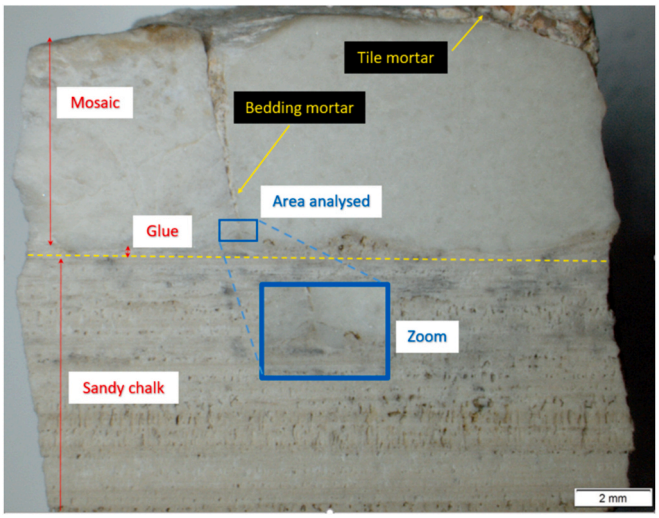


Fig. 1. Overview of the archaeological sample found in the *Ruscino* archaeological excavation from the periphery of Perpignan in South of France. The site was occupied by the Romans from the second half of the 2nd century BC to the end of the 1st century AD (i.e. *oppidum period*).

Company. The spectrometer is confocally coupled to an Olympus BX 40 high stability microscope equipped with a × 50 objective (NA = 0.5, WD = 10.6 mm). This instrument is equipped with a holographic grating of 1800 grooves/mm, giving a spectral resolution that can be as low as 0.5 cm^{-1}. A 1024 × 512 pixels charge-coupled device (CCD) camera cooled at −60 °C is used as detector. Raman backscattering is obtained with an excitation wavelength of 785 nm (25 mW at the sample) supplied by a solid-state laser. The sample is placed under the microscope on a motorized XY stage from Marzhauser Wetziar Company, which offers resolution in 0.1 μm steps in both the X and Y directions. The z-axis is controlled by an autofocus equal to a maximum of 50 μm. The region of interest on the archaeological mosaic fragment was specifically chosen with the aim of including both the adhesive mortar at the junction between the stone and the tesserae, and the original bedding mortar present between the tesserae (Fig. 1, blue rectangle). The experimental acquisition conditions for this 905 × 395 μm^2 area were optimised to obtain the best signal-to-noise ratio representing a 3D data matrix of size 181 × 79 × 1009 (i.e. pixels per pixels per wavelengths). Due to a point-by-point acquisition system, the step between two consecutive pixels was 5 μm. Acquisition took 8 h for the entire hyperspectral data cube, with a spectral resolution of around 1.7 cm^{-1} in our case. The next step was to use pre-processing to correct the raw data before extracting the relevant chemical information with the proposed approach.

2.3. Spectral data pre-processing

Fig. S2 shows the results of the various pre-processing stages on the collected raw Raman spectra. A typical issue known in Raman spectroscopy is that spectra are sometimes 'contaminated' by peaks, called 'spikes', caused by high-energy cosmic rays hitting the charge-coupled device (CCD) detector used to collect Raman photons. Their signals are associated with very narrow bandwidth peaks present at random positions in the spectrum. These spikes are problematic because they can interfere with subsequent analyses, particularly if multivariate analysis of the data is required. This is why one of the first steps in processing Raman spectra is dedicated to their removal. The software called LabSpec used to acquire Raman spectra had an initial filtering of the accumulated spikes known as *single peaks*. Even if, it eliminates most of the spikes, this correction is not ideal due to the choice of a window size to find the spikes (in our case equal to 3). As a consequence, several Principal Component Analysis (PCA) were applied to locate spikes-contaminated spectra or some other artefact signals (e.g. Raman spectra with saturation) and therefore to correct the signals by replacing them with the median of the spectra of the surrounding pixels. Fig. S3 shows some examples of 'corrupted' pixels found by the study of the scores and loadings from the different PCA.

The second pre-processing step consists in correcting the intense background commonly associated with fluorescence contributions. The intense, irregularly shaped baselines that change from pixel to pixel need to be corrected in order to extract current, unbiased chemical information from the spectra. The weighted Least-Squares (WLS) method, also called Asymmetric Weighted Least-Squares (AsLS) method, removes fluorescence contributions [14]. This pre-processing algorithm was originally proposed by Eilers et al. to subtract baseline shifts in chromatography. It is based on a recursive local fit of the entire spectrum to a baseline obtained using a Whittaker smoother [39].

Finally, a Multiplicative Scatter Correction (MSC) [40] is performed to remove the light-scattering effects in spectra. This correction is achieved by regressing a measured spectrum against a reference spectrum, in this case the median spectrum, and then correcting the measured spectrum using the slope of this fit. All pre-processing used in this study was carried out using the PLS toolbox v8.52 (Eigenvector Inc.) in the Matlab environnement (Natick, Massachusetts: The MathWorks Inc).

2.4. The proposed signal unmixing pipeline

The chemometric approach presented here is based on the Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) algorithm developed by R. Tauler and A. De Juan [33,38,41] with an image segmentation step founded on a threshold-based clustering algorithm. This segmentation step allows several spectral data matrices to be obtained from the pre-processed data. Each data matrix resulting from the clustering is then analysed using the Orthogonal Projection Approach (OPA) [31,35,42] to extract the best initial spectra for use in the MCR-ALS optimisation step. The three different steps of this data analysis pipeline will be described below.

2.4.1. STEP#1: image segmentation using a threshold-based clustering algorithm

Firstly, each spectrum at a given pixel is normalized by the sum of the absolute value of all its wavelengths (i.e. L1 normalization). As a consequence, it returns a vector with an area under the curve equal to 1. The aim is then to find the best estimate of the median image from the cube of pre-processed data on which the segmentation will be performed. All pixels must have the same weight, even if certain chemical contributions are less representative in terms of intensity. Pixel normalization is imperative here before running a clustering algorithm because of the use of a distance metric called Euclidean distance. Secondly, an unsupervised hierarchical clustering approach based on a threshold clustering algorithm is used, as with the 'mean-shift' methodology [43] or a modified 'k-means' approach [44]. This is a centroid-based algorithm that works by updating candidate pixels for a centroid group (i.e. a cluster) with a threshold criterion, and then finding all the given pixel regions present in the image until the number of iterations is reached. Table 1 presents the threshold-based algorithm implemented with MATLAB R2016a. The main advantage of this method lies in the automatic search for an optimal number of clusters, which naturally influences the quality of data partitioning. Furthermore, it is potentially less sensitive than classical clustering methods [45]. The convergence of the algorithm towards a unique solution is ensured by a large number of iterations. Furthermore, depending on the application, this unsupervised algorithm can easily be adapted by changing the threshold criterion. This key parameter of this algorithm defines the size of the search window and influences the maximum distance over which points are moved with respect to a random centroid that is updated regularly. It can therefore have a significant impact on the clustering

Table 1
The threshold-based clustering algorithm.

Initialisation: Select an element from the given dataset as the centroid (in this case, the average of the intensities of the pixels in the median image). This element is considered to be the 'seed' of a cluster.

WHILE it is true

→ Calculate the Euclidean distance between the elements (pixels) and the centroid selected earlier. The objective is to find for each unclassified element (pixels) its distance to the existing cluster (remember that in the first iteration this is all the pixel intensities of the median image).

→ Calculate the threshold (here a standard deviation, **equation (1)**):

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{n}^{i=1}(x_i - c)^2} \quad (1)$$

With x_i the elements (pixel intensities) and c the selected centroid.

→ If the distance is less than the threshold, assign the element to this cluster. Recalculate the centroid of the cluster as the average of all selected elements and if no such cluster can be found after examining all current clusters, you can assign the new centroid as the 'seed' for the cluster.

→ If the distance of new cluster to another one is less than the threshold, merge the two nearby clusters and recalculate the distances between the cluster.

→ If all the elements have been assigned to either cluster, the algorithm **STOPS**.

UNTIL the number of iterations has reached its maximum

→ As the end of the previous step, the centroids of clusters are sorted and the difference between two consecutive centroids is calculated in order to reject cluster centroids that are less than the inter-cluster distance.

Result: You have access to the centroids and their clusters

results. The function use to define it, is therefore important (e.g. standard deviation, Gaussian density, etc.), but that's what makes it a such flexible tool. As well as being sensitive to bandwidth choices, it can also be computationally expensive as the amount of data increases.

At the end of STEP#1, similar pixels are grouped into j different clusters. It is then possible to work directly on the spectral information of these identified clusters because in hyperspectral imaging, a continuous spectrum is measured for each pixel [41]. Each spectral matrix found from similar pixels is noted here $\mathbf{X_j}$ ($m \times n$) and corresponds to m spectra measured at regular time intervals due to the Raman mapping and n columns are the wavelengths (Raman shift in cm^{-1}).

2.4.2. STEP#2: initialisation with multiple orthogonal projection approach (OPA)

OPA is applied to each spectral matrix $\mathbf{X_j}$ deduced from the j clusters in order to extract their most dissimilar spectra. The OPA algorithm uses a dissimilarity criterion based on Gram-Schmidt orthogonalisation [31,35,42]. The number of columns or rows selected for each matrix $\mathbf{X_j}$ corresponds to the number of chemical components present in a cluster of the archaeological sample and is estimated during rank evaluation with PCA and with the dissimilarity criteria of an OPA. In this paper, the OPA searches for the least correlated spectra with the highest average intensity over the whole spectral range (i.e. the direction of the spectra in matrix $\mathbf{X_j}$). In our case, the independent evaluation of the ranks of the $\mathbf{X_j}$ matrices is not critical for the proposed multivariate curve resolution since the initial spectra are normalized to be carefully selected visually before the ALS optimisation. Thus, an overestimation of these 'local' ranks will not lead to an overestimation of the 'global' rank of the pre-processed matrix, as all the selected initial spectra at the end are different from each other in terms of wavelength selectivity. In addition, spectral contributions that do not represent relevant chemical information are not to be selected.

Usually, in the initialisation step of the MCR-ALS algorithm, a 'pure' variable-based method [34,46] called SIMPLISMA, acronym for SIMPLe-to-use Interactive Self-modelling Mixture Analysis, is carried out [12,14,15]. The assumption of SIMPLISMA is that each spectral component of the mixture has a variable that has a specific contribution for a given particular compound, and that this variable has zero intensity for all other spectral components of the mixture, i.e. the notion of 'pure' component. This ideal case is practically never encountered and this method is therefore based on variables closest to the 'pure' variables [46]. On the other hand, the assumption of OPA is that the purest spectra in the data matrix are mutually more dissimilar than the corresponding spectra of the mixture, which can then be applied to detect significant changes in successive recorded spectra. As a consequence, when spectra are very similar due to baseline shifts and/or spectral overlap (i.e. selectivity problems), as in our case in Raman imaging due to the complexity of the sample, OPA is preferable [47,48]. The main objective of multiple OPA is to find the right number of spectral components before ALS optimisation and also, the best initial spectra. This step is crucial since it conditions future solutions of the MCR-ALS algorithm [32]. As with any optimisation method, the MCR-ALS depends on the estimate of the initial matrix which itself depends on the estimate of the global rank of the pre-processed data matrix. The OPA calculations were implemented in MATLAB R2016a.

2.4.3. STEP#3: Alternating Least square (ALS) optimisation

The MCR-ALS algorithm is used to decompose a matrix $\mathbf{D}$ into the pure distribution maps $\mathbf{C}$ and the pure spectra $\mathbf{S^T}$ of all constituents present in the analysed archaeological sample. This decomposition is carried out according to the bilinear model presented in **Equation (2)** [12,14,15,33,38,41]. A bilinear model of k pure components can reproduce the preprocessed matrix $\mathbf{D}$ containing n mixed spectra collected at λ spectral channels with $n \times k$ concentration profiles and their $k \times \lambda$ related spectral profiles.

$$\mathbf{D} = \mathbf{CS^T} + \mathbf{E} \quad (2)$$

The error contribution of the model is represented by the residual matrix $\mathbf{E}$ ($n \times \lambda$). The Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS), developed by R. Tauler et al. [33,38,41], is an

iterative curve resolution method used to recover the underlying spectroscopic bilinear model in **Equation (2)**. The ALS step involves the operations $\mathbf{C}_{opt} = \mathbf{DS}(\mathbf{S}^{T}\mathbf{S})^{-1}$ and $\mathbf{S}^{T}_{opt} = (\mathbf{C}^{T}\mathbf{C})^{-1}\mathbf{CD}$ respectively which are alternatively refined under given constraints. The end of the iterative process takes place when the product of the resolved concentration profiles and spectra reproduce the original data **D** without significant variation between consecutive iterations. Both the explained variance (r^2) and the lack of fit (*Lof*), defined as follows in **Equations 3** and **4**, are calculated to determine the fit quality of the MCR-ALS model:

$$r^2(\%) = 100 \times \left(1 - \frac{\sum_i \sum_j e_{i,j}^2}{\sum_i \sum_j d_{i,j}^2}\right) \quad (3)$$

$$Lof(\%) = 100 \times \sqrt{\frac{\sum_i \sum_j e_{i,j}^2}{\sum_i \sum_j d_{i,j}^2}} \quad (4)$$

where $e_{i,j}$ are the elements of **E** matrix and $d_{i,j}$ are the elements of the raw dataset **D**. Subindexes *i* and *j* refer to the pixel and the wavelength number, respectively. As a result, the ALS optimisation is applied here to the pre-processed spectral data (i.e. a Raman spectra 2D matrix of size 14299 × 1009) with a non-negativity constraint added to concentration and spectral profiles. A normalization constraint is then used, but only for spectra. These constraints are applied in order to exclude some solutions during the MCR-ALS optimisation and thus reduce the rotational ambiguity inherent to these types of matrix decomposition calculations [24,31]. The MCR-ALS Matlab code is available free on the website: http://www.mcrals.info/.

3. Results and discussion

The results of the three different steps presented above for the multivariate curve resolution analysis will be described. STEP#1 is the image segmentation based on the threshold-based algorithm and is performed on the median image (181x79, Fig. 2A) deduced from the pre-processed data cube (181x79x1009). The result is five clusters (i.e. j = 5) (Fig. 2B) containing respectively 4 pixels (cluster n°1, blue colour), 574 pixels (cluster n°2, light blue), 4819 pixels (cluster n°3, light green), 6979 pixels (cluster n°4 orange) and 1923 pixels (cluster n°5, dark red). The aim here is to extract the best spatial information from the surface of the archaeological artefact studied. Five spectral matrices with respective dimensions 4 × 1009 for $\mathbf{X_1}$, 574 × 1009 for $\mathbf{X_2}$, 4819 × 1009 for $\mathbf{X_3}$, 6979 × 1009 for $\mathbf{X_4}$ and 1923 × 1009 for $\mathbf{X_5}$ are then deduced from this clustering step (Fig. 2C). Interestingly, the spectral matrices found by segmentation are related to the signal-to-noise ratio of the measurements. Similar data points (pixels) found by the threshold-based algorithm clustering appear to have an equivalent signal-to-noise ratio value from the spectral information point of view. Indeed, the higher the number of a cluster is, the lower the signal-to-noise ratio of the spectral matrix is.

STEP#2 is then performed to obtain (by the end of this step) the final matrix of the initial spectra to be used in the ALS optimisation. To do so, multiple OPA independently decompose each matrix $\mathbf{X_j}$ (here with j = 1, …, 5) by searching for the most dissimilar spectra. Here, the explained variance of the PCA and the dissimilarity criteria of the OPA, are studied respectively. The aim is to estimate a 'local' rank for each $\mathbf{X_j}$ matrix. Finally, a rigorous study of each spectral component extracted by multiple OPA from each $\mathbf{X_j}$ matrix is carried out by comparing them after normalization. This is done in order to select the best 'pure' compounds as an initial matrix before ALS optimisation.

To understand the methodology, the simple case of cluster No. 1 can be taken, as there are only four spectral signals from the pre-processed

Table 2
Study of the matrix $\mathbf{X_1}$ with PCA and OPA.

PCA		OPA	
Component number	Explained variance (%)	Spectrum number	Dissimilarity criteria (× 10^8 u.a)
1	81.01	**1**	1.497
2	10.91	**2**	1.080
3	7.927	**3**	0.235
4	0.151	**4**	1.317

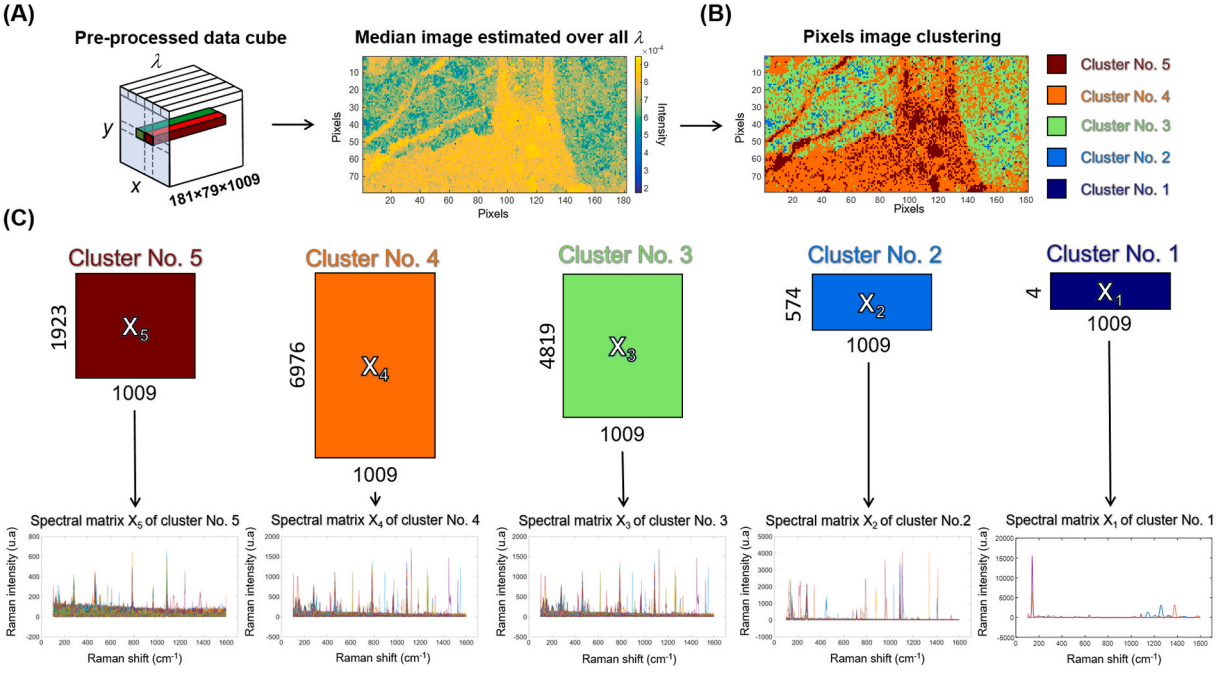


Fig. 2. STEP#1 results. (A) The median image estimated from the pre-processed data cube, (B) the result of the segmentation of the median image with the threshold-based clustering algorithm implemented. Five clusters of pixels of respective dimensions were found: 4 pixels (blue), 574 pixels (light blue), 4819 pixels (light green), 6979 pixels (orange) and 1923 pixels (dark red).

data after STEP#1. Table 2 shows the percentage of variance explained by each four PCA components and the dissimilarity criteria calculated over the four spectra by the OPA. As observed, the first three components explain 99.99% of the variance of the $\mathbf{X_1}$ matrix, which means that the first three components have the most relevant information in this matrix. It is therefore possible to estimate that the 'local' rank of the $\mathbf{X_1}$ matrix is equal to 3, which is also confirmed by the dissimilarity criteria where spectra 1, 4 and 2 are different from the third, respectively 1.497×10^8, 1.317×10^8, 1.080×10^8 compared to 0.235×10^8.

Therefore, the OPA detects spectrum n°3 as correlating with another spectral signature. Fig. S4 shows the four spectral signatures of the $\mathbf{X_1}$ matrix, where spectra n°3 and n°4 can be seen as having the same spectral signatures. The combined results of PCA and OPA allow us to estimate the 'local' rank of the matrix $\mathbf{X}_1$ as being 3. Consequently, the OPA decomposition of this matrix gives a 'first initialisation matrix' of dimension 3×1009.

Given the number of spectra contained in the $\mathbf{X_1}$ matrix, PCA and OPA were not needed to estimate its 'local' rank and therefore its decomposition. The aim was to explain our reasoning, which could be repeated for matrices $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_4}$ and $\mathbf{X_5}$ where the difficulty is much greater because of the number of spectra. Fig. 3 shows the results for all matrices, using the same reasoning as discussed above. The matrix $\mathbf{X_2}$ has a dimension of 579×1009 which justifies first studying the variance explained by PCA in order to understand which components best summarise the most relevant information and thus propose an estimate of its matrix rank. As Fig. 3B shows, component No. 1 alone explains 90% of the total variance contained in the $\mathbf{X_2}$ matrix. Zooming in, components No. 2 and No. 3 are respectively around 3% and 2% before a plateau around 0.5% appears after component No. 4. It is therefore difficult to estimate the 'local' rank of this matrix directly because of component No. 1. In fact, the matrix rank of the latter is more than likely not equal to 1 and the values associated with the variances would tend to indicate a rank deficiency. However, the study of the dissimilarity criterion on the $\mathbf{X_2}$ matrix shows several different spectral contributions, in particular for dissimilarity values greater than 2×10^7 where ten spectra appear to be very different. It is therefore possible to consider a selection of 10 spectra for the $\mathbf{X_2}$ matrix. For matrix $\mathbf{X_3}$, the values of the variance explained by components also make it difficult to estimate its rank (Fig. 3C). Despite several changes in slope observed between components No. 3 and No. 7, these are nonetheless small. On the other hand, the graph of dissimilarities shows that from a value greater than 1.5×10^7, seven spectral contributions appear to be very different, allowing us to select seven spectra for matrix $\mathbf{X_3}$. For matrices No. 4 and No. 5, the task is more difficult because the dissimilarity peaks are very close to the noisy signal. We noted earlier that it was possible to classify the spectral matrices found by clustering in descending order of signal to noise ratio (Fig. 2C) i.e. $\mathbf{X_1} > \mathbf{X_2} > \mathbf{X_3} > \mathbf{X_4} > \mathbf{X_5}$. Fig. 5 shows that the proportion of variance explained by the first component of $\mathbf{X}_4$ (panel D) and $\mathbf{X}_5$ (panel E) matrices falls to around 70% and 50% respectively. It is even harder to estimate their ranks. The dissimilarity values decrease drastically (they are now 10^6 order of magnitude) and are closer to each other than to those of matrices $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_3}$. However, it is possible to find different spectral contributions for dissimilarity values greater than 5.5×10^6 and 1.25×10^6 allowing the selection of six and three spectra respectively for matrices $\mathbf{X_4}$ and $\mathbf{X_5}$.

In summary, following the PCA and OPA study of the $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_4}$ and $\mathbf{X_5}$ matrices, an initial set of spectra is selected for each of them and is of respective size: 3×1009, 10×1009, 7×1009, 6×1009 and 3×1009. It is not yet possible to use all these contributions as a single initial matrix prior to ALS optimisation. Indeed, it is possible that from one matrix $\mathbf{X_j}$ to another, there are identical and/or noisy spectral contributions. It is quite possible, for example, that spikes have not been completely corrected or that spectra have been distorted by pre-processing of the raw data.

It is therefore essential to carefully examine and compare each of the spectral components, selecting those that are as 'pure' as possible. Fig. 4 shows the ten spectra selected. As the result, the overall matrix rank of the pre-processed matrix has been estimated as being equal to 10 using this method. All that remains is to carry out STEP#3 with the ALS optimisation of the pre-processed data matrix.

Figs. 5 and 6 show the results of the ten chemical compounds extracted from the ALS optimisation and referred to respectively as the 'pure' spectral and concentration profiles, i.e. $\mathbf{S}_{opt}^{T}$ and $\mathbf{C}_{opt}$. The convergence criterion equal to 0.01 was achieved after 184 iterations. The lack of fit was 24.14% and the percent of explained variance r^2 by the model at the optimum was 94.17%, which are relatively good figures of merit considering our spectral data. As a reminder, the main idea behind the MCR-ALS algorithm is that each pixel or spectrum in Raman imaging can be described as a linear combination of a set of pure compound spectra. Therefore, as indicated earlier, the term 'pure compound' can be either a chemical compound or a mixture of chemical compounds with its own spectral signature.

As can be seen in Fig. 5, chemical compound No. 8 is attributed to a non-spectral component which is due to the non-ideal correction of the baseline of the raw data. The fluorescence effect observed in the raw data with a baseline that necessitated correction was significant and, as is often the case in data analysis, the pre-processing did not ensure perfect correction (see Fig. S2, panel A). On the one hand, the extracted chemical components No. 1, No. 2, No. 3, No. 5 and no. 10 are potentially associated with pure compounds. On the other hand, the extracted chemical compounds No. 4, No. 6, No. 7 and No. 9 are mixtures of chemical contributions.

Fig. 6 confirms the first spectral observations (as seen above) due to the ten spatial distributions of the chemical contributions over the area of the archaeological sample analysed by Raman imaging. In addition to the fact that contribution No. 8 is considered as a non-chemical component, the pure chemical compounds maps No. 1, No. 2, No. 3, No. 5 and No. 10 cover the entire surface of the archaeological sample analysed, but with predominantly low pixel intensities and a few pixels of high intensity (Fig. 6, red circles) confirming the presence of truly specific spectral sources. The spatial comparison of both pure chemical compound maps, No. 4 and No. 9, appears to correlate, for example the shape of the cracks observed (Fig. 6, red squares). However, they differ from one pixel to another because of their spectral component. In fact, from the spectral point of view (Fig. 5), only the wavelength present at 1086.5 cm^{-1} can be used to distinguish compound No. 4 from compound No. 5. In addition, spatial observations of the pure chemical maps of compounds No. 6 and No. 7 show a more pronounced visual feature, with a line separating the analysis zone into two distinct parts (Fig. 6, dotted red line). Looking more closely at the spectral contributions of No. 6 and No. 7 (Fig. 5), no real differences in wavelength can be seen, but rather a difference in the Raman shift around 5 cm^{-1}. This discrepancy is not always observed as moving in the same direction: sometimes it must be added, sometimes it must be subtracted. The main difference does appear in the 1080-1090 cm^{-1} spectral range related to $CaCO_3$ and may even indicate the presence of several polymorphs or differences in carbonate chemistry. However, the spatial information is unique which can be explained by the nature of the material used during the *oppidum* period on the archaeological *Ruscino* site.

To confirm this hypothesis, each of the spectral signatures will be identified using the literature. Firstly, the pure chemical compound No. 5 is silicon carbide SiC [49] which is present in our Raman spectral analysis because the archaeological sample was polished prior to analysis by scanning electron microscopy analysis. The wavelengths at 146.9 cm^{-1}, 238.6 cm^{-1}, 503.9 cm^{-1}, 765.5 cm^{-1}, 786.7 cm^{-1} and 967.8 cm^{-1} are selective. Consequently, this pure chemical compound is not present in the pristine sample. Pure chemical compound No. 4, No. 6, No. 7 and No. 9 are biogenic or inorganic magnesian calcites [50,51], respectively calcite with 3.9% of $MgCO_3$, pure calcite, calcite with 9.9% of $MgCO_3$ and calcite with 2% $MgCO_3$. Magnesian calcites are an important mineral component of modern and Pleistocene carbonate

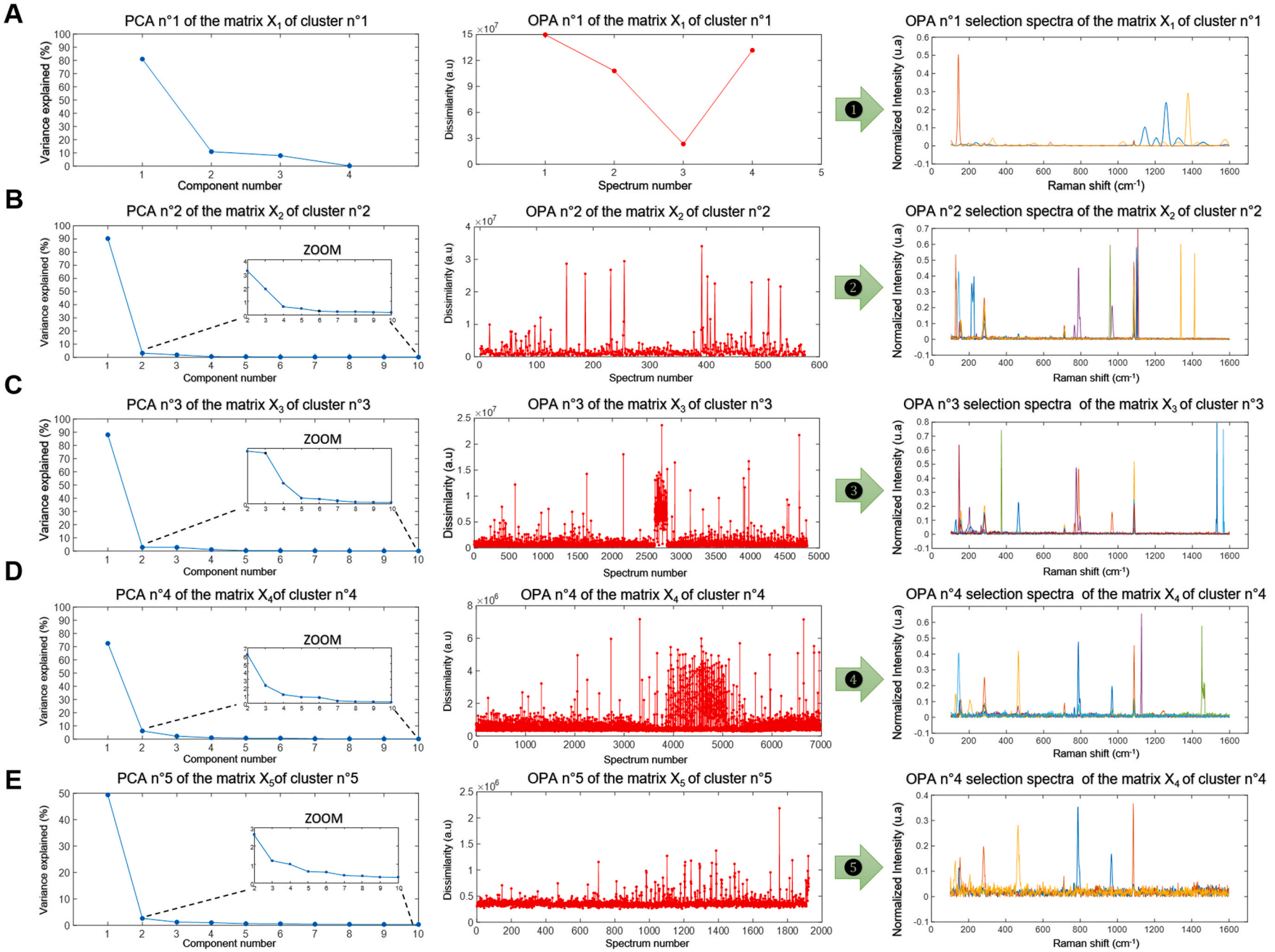


Fig. 3. The panels A, B, C, D and E present the results of the selected spectra by OPA for each matrix $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_4}$ and $\mathbf{X_5}$ respectively.

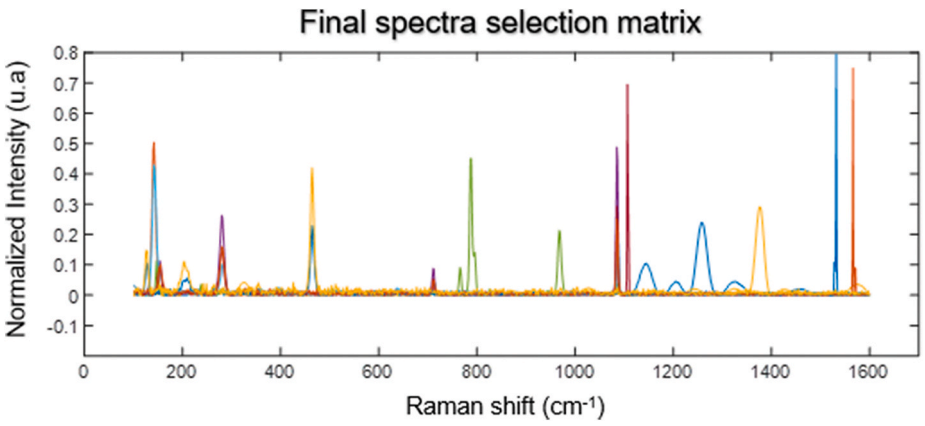


Fig. 4. The 10 spectral contributions to be used as the initial matrix before ALS optimisation.

sediments. These phases are found mainly in the skeletons of marine invertebrates and in cements [52–54]. Consequently, the Raman shift observed between the spectral information can be explain by the percentage of $MgCO_3$ contained in calcite and it remains unclear what the origin of this spectral signature is, e.g. coral, pearl or skeletons. Pure chemical compound No. 1 has specific wavelengths at 302.8 cm^{-1}, 1146.3 cm^{-1}, and 1581.8 cm^{-1} which could be attributed to the red/ocher pigment $Fe(Phen)_2(NCS)_2$ based on Fer(II) ions [55]. Pure chemical compound No. 2 has two specific bands at 142.0 cm^{-1} and 195.5 cm^{-1} from koechlinite [56] and three specific bands from anatase TiO_2 single crystal [57] at 395.4 cm^{-1}, 514.4 cm^{-1} and 635.4 cm^{-1} although very weak. Pure chemical compound No. 10 has specific wavelengths at around 126.8 cm^{-1}, 205.5 cm^{-1}, 393.7 cm^{-1} and 464.2 cm^{-1} which are attributed to xenomorphic quartz inclusion [58]. Finally, the pure chemical compound No. 3 is much more difficult to interpret as it only shows one intense spectral wavelength at 1375.2 cm^{-1} attributed to –NH band and a smaller one at 325.7 cm^{-1} due to the stretch bond –O-C-*O*-. These peaks are characteristic of amino-acids and collagen traces [59,60]. Hypotheses may be supported where biogenic magnesian calcites are found.

4. Conclusion

4.1. The relevance of the proposed chemometric approach

The chemometric methodology presented in this work demonstrates the possibility of characterizing an archaeological sample in Raman imaging using MCR-ALS without any prior knowledge. For this, relevant pixels from the median Raman image of the artefact were first identified and grouped into five clusters after a novel threshold-based unsupervised clustering approach. Afterwards, multiple OPA analysis were applied to find the best initial selective spectral data used in the ALS optimisation. The segmentation image (STEP#1) followed by multiple OPA analysis (STEP#2) facilitated the estimation of the rank of the data matrix and thus gave a better estimate of the initial guess necessary for ALS optimisation. Therefore, the rotation ambiguities were limited and the convergence towards a global minimum was improved. Consequently, the solutions deduced from this optimisation were more relevant and characteristic of the chemical reality.

Indeed, if the classical approach to estimate the rank of a data matrix is compared with the proposed chemometric approach, the results will be very different. Fig. S5 (in supplementary material) shows the eigenvalues (Panel A) and the percentage of explained variance for each of them (Panel B) after decomposition by PCA [42]. Estimating matrix rank is complicated here without a priori by the difficulty of truncating the curves of panel A and B in Fig. S5 at the appropriate number of principal components (red solid and dotted curves). The values of eigenvalues after 7 principal components do not improve significantly and express 80% of the most important information in the pre-processed matrix analysis. Consequently, if the matrix rank is 7, the initial matrix $\mathbf{S}_{ini}^{T}$ would be constructed with fewer or even different spectral contributions. The bilinear decomposition by MCR-ALS would have been less representative of the chemical reality due to underestimating the matrix rank. To confirm this, an initial estimate was made using SIMPLISMA and OPA directly on the pre-processed matrix with a rank equal to 7 (Fig. S6 in supplementary material). The ALS optimisation offers the same solutions between SIMPLISMA or OPA initializations, but not always in the same order, which in itself is not harmful. Nevertheless, important contributions are not detecting with these usual approaches. The pure chemical compound No. 1 found with the proposed signal unmixing pipeline in this article in Fig. 6, is not present in Fig. S6 panel A and B. Consequently, the chemical reality is much better described with our proposed chemometrics approach, because it is less dependent on matrix rank estimation.

In general, it is necessary to -over- or -under- estimate the value of the matrix rank in order to then look at the results of the MCR-ALS optimisation, even if the other risk is to overestimate the compounds present in the sample. However, in addition to concerns about rank deficiencies, the complexity of the signals measured (e.g. low signal-to-noise ratio) can lead to inaccurate initial estimates of spectral profiles or concentrations, and thus to unsuitable solutions being proposed after ALS optimisation. The chemometrics approach presented here endeavours to be less sensitive to, or dependent on, these drawbacks. Thus, it tends towards the most relevant chemical characterization. To confirm this, an initial estimate was made using SIMPLISMA and OPA directly on the pre-processed matrix with a rank equal to 10 which is not possible to justify as explained above (Fig. S7 in supplementary material). A comparison with Fig. 4 shows that some initial spectral contributions are different from those selected by our approach. For example, the spectral contribution around 218 cm^{-1} is present in both the SIMPLISMA (Blue arrow in Fig. S7, panel A) and OPA (Blue arrow in Fig. S7, panel B) estimates but is not selected using our approach (Fig. 4). This contribution is explained by the presence of a spike that could not be corrected properly. In addition, some initial spectral contributions between the two approaches SIMPLISMA and OPA are also different (red arrows around 1100 cm^{-1}, Fig. S7, panel A and B) and this is explained by the difference in the selection criterion.

The initialisation step before ALS optimisation is therefore important. In fact, three potential initial spectral matrices could be used in the ALS (those present in Fig. 4, Fig. S7 panel A and B) with a risk of obtaining unsatisfactory solutions. Fortunately, in our case, regardless of the matrix used for a rank of 10, the optimisation offers the same solutions but not always in the same order as those in Figs. 5 and 6, reflecting the fact that a global minimum is found. Nevertheless, one initialisation seems better than the others. Table 3 shows the optimisation results obtained by ALS with the non-negativity constraints on concentrations

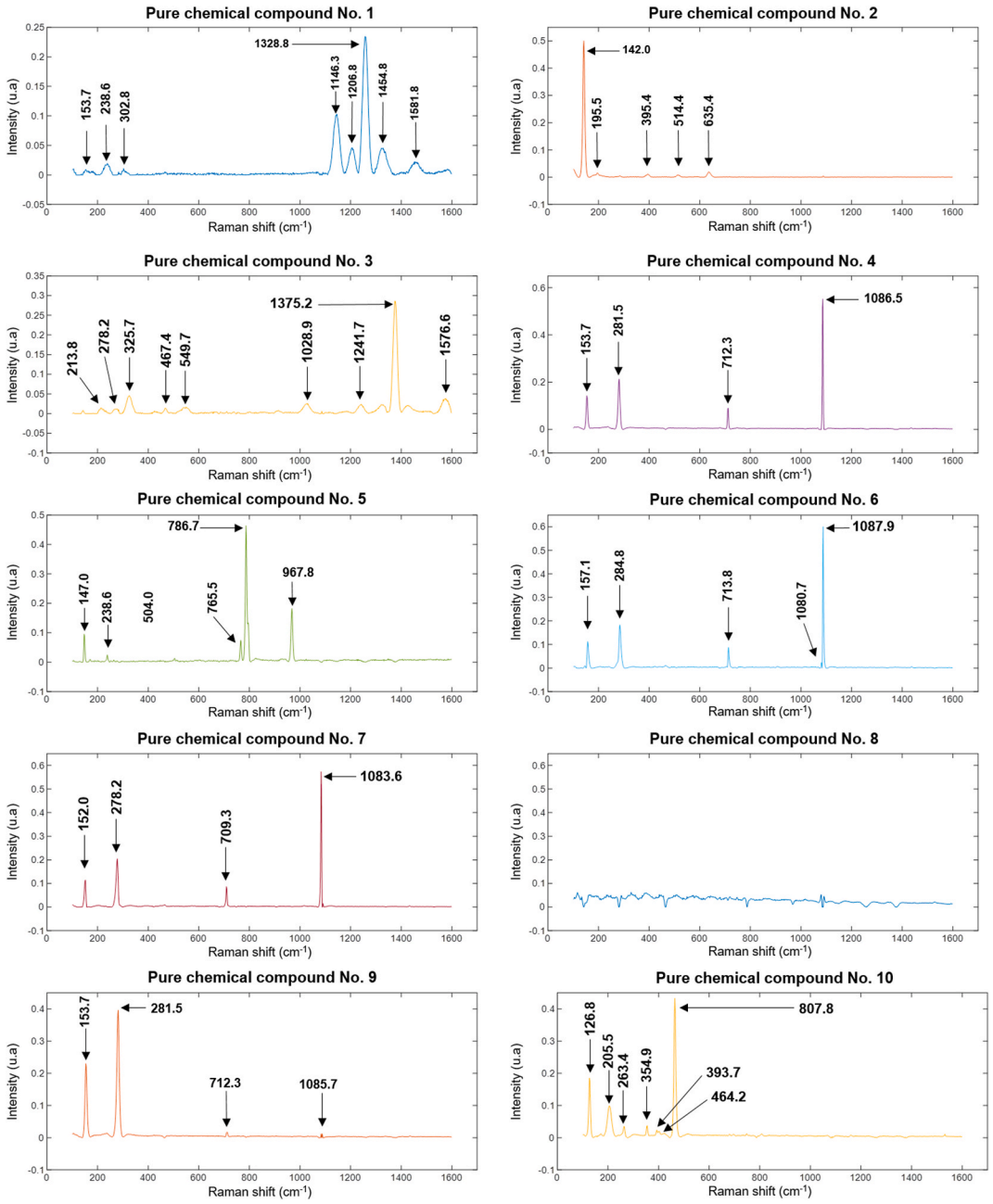


Fig. 5. The Multivariate Curve Resolution and Alternating Least-square (MCR-ALS) analyses of the archaeological sample. Here is the spectral matrix $\mathbf{S}_{\mathbf{opt}}^{\mathbf{T}}$ which contained 10 pure chemical compounds.

and spectra solutions with the addition of a normalization on the spectra and a convergence criterion of 0.01. Although the quality of the regressions is similar according to the method, the number of iterations to converge on the optimum is better with our approach, i.e. the spectral initialisation matrix deduced from segmentation and multiple OPA. This implies that our initial guesses are closer to ALS solutions than those proposed by SIMPLISMA or OPA alone. There is, therefore, a lower risk of converging towards global minimums. In terms of identifying chemical compounds, the results are interesting in order to understand the origin of the mortar.

4.2. *The archaeologist's point of view*

The most outstanding archaeological result of this study is the detection of biogenic materials (magnesium calcites or amino-acids) of animal origin, in particular fish or shellfish. This result is also confirmed in Fig. S8 by other techniques, in particular Scanning Electronic Microscopy (SEM). For example, if the smallest MCR-ALS map of the contribution No. 10 (i.e. quartz) is compared to the biggest observed area on the SEM map (F), they are quite similar. This is because quartz is a common mineral species in the silicate group, a subgroup of the tectosilicates, composed of silicon dioxide, or silica, with the chemical formula SiO_2. The same remarks apply to other common compounds, such as iron or magnesium. This also confirms the interest of the data analysis pipeline presented here. In fact, despite the complexity of the archaeological artefact, the chemical compounds identified by multivariate curve analysis in Raman imaging are in agreement with the elements found by SEM.

However, the local origin of the materials seems difficult to establish, but could nonetheless be the result of a supply of various materials by

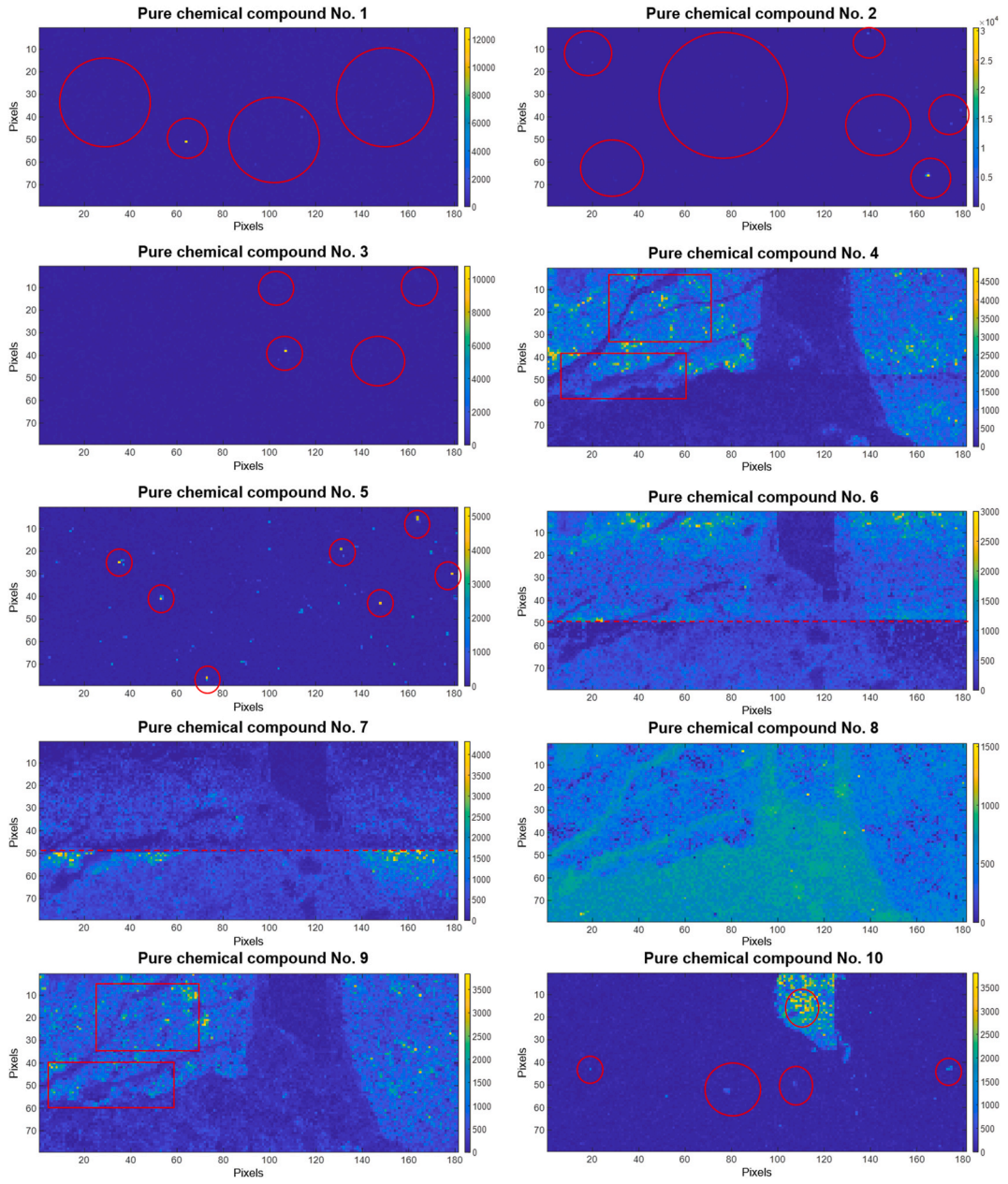


Fig. 6. The Multivariate Curve Resolution and Alternating Least-square (MCR-ALS) analyses of the archaeological sample. Here is the concentration matrix $\mathbf{C}_{opt}$ which contained the 10 pure chemical compounds.

Table 3
ALS optimisation with three different spectral initialisation matrices.

The spectral initialisation matrix	ALS-Optimisation		
	Number of iterations	Lack of fit (%)	The variance explained (%)
Segmentation and multiple OPA	184	24.10	94.17
SIMPLISMA	327	24.14	94.17
OPA	212	24.14	94.25

trade covering the whole of the Roman Empire. The historical dating and the exact role of the glue between the mosaic and the stone is still debated from an archaeological point of view.

Code availability

The code (OPA and threshold-based clustering) employed for the study is available from the corresponding author upon reasonable request. The MCR-ALS algorithm can be found on http://www.mcrals.info/. The pre-processed made in this study are performed by the Eigenvector toolbox available on https://eigenvector.com/.

CRediT authorship contribution statement

Marc Offroy: Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Mario Marchetti:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis. **Thomas H. Kauffmann:** Writing – review & editing, Methodology, Investigation. **Patrice Bourson:** Writing – review & editing, Methodology, Investigation. **Ludovic Duponchel:** Writing – review & editing, Validation, Methodology,

Investigation. **Laurent Savarese:** Writing – review & editing, Investigation. **Jean-Michel Mechling:** Writing – review & editing, Validation, Investigation, Formal analysis.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

The authors greatly acknowledge the spectroscopy platform of the LMOPS, University of Lorraine, as well as the electronic microscopy and micro-analyses platform of GeoRessources Laboratory, University of Lorraine. The authors thank Michelle Adrian co-head of ENSIC languages department.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2024.125955.

References

[1] B.T. Nigra, K.F. Faull, H. Barnard, Analytical chemistry in Archeological research, Anal. Chem. 87 (2015) 3–8, https://doi.org/10.1021/ac5029616.
[2] D.R. Brothwell, A.M. Pollard, Handbook of Archaeological Sciences, Wiley Blackwell, New York, 2008.
[3] R.E. Taylor, Radiocarbon dating in archaeology, in: C. Smith (Ed.), Encyclopedia of Global Archaeology, Springer, New York, 2014.
[4] G. Di Maida, M.A. Mannino, B. Krause-Kyora, T.Z.T. Jensen, S. Talamo, Radiocarbon dating and isotope analysis on the purported Aurignacian skeletal remains from Fontana Nuova (Ragusa, Italy), PLoS One 14 (3) (2019) e0213173, https://doi.org/10.1371/journal.pone.0213173.
[5] B. Genc Oztoprak, M.A. Sinmaz, F. Tülek, Composition analysis of medieval ceramics by laser-induced breakdown spectroscopy (LIBS), Appl. Phys. A 122 (557) (2016) 3–11, https://doi.org/10.1007/s00339-016-0085-9.
[6] A. Nevin, G. Spoto, D. Anglos, Laser spectroscopies for elemental and molecular analysis in art and archaeology, Appl. Phys. A 106 (2012) 339–361, https://doi.org/10.1007/s00339-011-6699-z.
[7] A. Hauptmann, S. Schmitt-Strecker, F. Begemann, A. Palmieri, Chemical composition and lead isotopy of metal objects from the Royal tomb and other related finds at Arslantepe, Eastern Anatolia, Paleorient 28 (2002) 43–69, https://doi.org/10.3406/paleo.2002.4745.
[8] J. Ling, S. Stos-Gale, L. Grandin, K. Billström, E. Hjärthner-Holdar, P.-O. Persson, Moving metals II: provenancing Scandinavian Bronze Age artefacts by lead isotope and elemental analyses, J. Archaeol. Sci. 41 (2014) 106–132, https://doi.org/10.1016/j.jas.2013.07.018.
[9] A. Pecci, Chromatography and archaeological materials analysis, in: S.L. López Varela (Ed.), The Encyclopedia of Archaeological Sciences, 2018.
[10] J.P. Ogalde, B.T. Arriaza, E.C. Soto, Identification of psychoactive alkaloids in ancient Andean human hair by gas chromatography/mass spectrometry, J. Archaeol. Sci. 36 (2) (2009) 467–472, https://doi.org/10.1016/j.jas.2008.09.036.
[11] R. Salzer, H.W. Siesler, Infrared and Raman Spectroscopic Imaging, Wiley-VCH, Weinheim, 2009.
[12] M. Offroy, M. Moreau, S. Sobanska, P. Milanfar, L. Duponchel, Pushing back the limits of Raman imaging by coupling super-resolution and chemometrics for aerosols characterization, Sci. Rep. 5 (2015) 12303, https://doi.org/10.1038/srep12303.
[13] D. Neff, L. Bellot-Gurlet, P. Dillmann, S. Reguer, L. Legrand, Raman imaging of ancient rust scales on archaeological iron artefacts for long-term atmospheric corrosion mechanisms study, J. Raman Spectrosc. 37 (2006) 1228–1237, https://doi.org/10.1002/jrs.1581.
[14] S. Piqueras, L. Duponchel, M. Offroy, F. Jamme, R. Tauler, A. de Juan, Chemometric strategies to unmix information and increase the spatial description of hyperspectral images: a single-cell case study, Anal. Chem. 85 (13) (2013) 6303–6311, https://doi.org/10.1021/ac4005265.
[15] A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault, M. Maeder, Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis, Trends Anal. Chem. 23 (2004) 70–79, https://doi.org/10.1016/S0165-9936(04)00101-3.
[16] L. Bellot-Gurlet, S. Pagès-Camagna, C. Coupry, Raman spectroscopy in art and archaeology, J. Raman Spectrosc. 37 (2006) 962–965, https://doi.org/10.1002/jrs.1615.
[17] A. Dominguez-Vidal, M.J. de la Torre-López, M.J. Campos-Suñol, R. Rubio-Domene, M.J. Ayora-Cañada, Decorated plasterwork in the Alhambra investigated by Raman spectroscopy: comparative field and laboratory study, J. Raman Spectrosc. 45 (2014) 1006–1012, https://doi.org/10.1002/jrs.4439.
[18] M.L. Roldán, S.A. Centeno, A. Rizzo, An improved methodology for the characterization and identification of sepia in works of art by normal Raman and SERS, complemented by FTIR, Py-GC/MS, and XRF, J. Raman Spectrosc. 45 (2014) 1160–1171, https://doi.org/10.1002/jrs.4620.
[19] M. Gutman, M. Lesar-Kikelj, A. Mladenović, V. Čobal-Sedmak, A. Križnar, S. Kramar, Raman microspectroscopic analysis of pigments of the Gothic wall painting from the Dominican Monastery in Ptuj (Slovenia), J. Raman Spectrosc. 45 (2014) 1103–1109, https://doi.org/10.1002/jrs.4628.
[20] P.T.C. Freire, J.H. Silva, F.E. Sousa-Filho, B.T.O. Abagaro, B.C. Viana, G.D. Saraiva, T.A. Batista, O.A. Barros, A.A.F. Saraiva, Vibrational spectroscopy and X-ray diffraction applied to the study of Cretaceous fish fossils from Araripe Basin, Northeast of Brazil, J. Raman Spectrosc. 45 (2014) 1225–1229, https://doi.org/10.1002/jrs.447.
[21] L. Medeghini, P.P. Lottici, C. De Vito, S. Mignardi, D. Bersani, Micro-Raman spectroscopy and ancient ceramics: applications and problems, J. Raman Spectrosc. 45 (2014) 1244–1250, https://doi.org/10.1002/jrs.4583.
[22] D. Lambert, C. Muehlethaler, L. Gueissaz, G. Massonnet, Raman analysis of multilayer automotive paints in forensic science: measurement variability and depth profile, J. Raman Spectrosc. 45 (2014) 1285–1292, https://doi.org/10.1002/jrs.4490.
[23] E.M.J. Schotsmans, A.S. Wilson, R. Brettell, T. Munshi, H.G.M. Edwards, Raman spectroscopy as a non-destructive screening technique for studying white substances from archaeological and forensic burial contexts, J. Raman Spectrosc. 45 (2014) 1301–1308, https://doi.org/10.1002/jrs.4526.
[24] V. Otero, D. Sanches, C. Montagner, M. Vilarigues, L. Carlyle, J.A. Lopes, M. J. Melo, Characterisation of metal carboxylates by Raman and infrared spectroscopy in works of art, J. Raman Spectrosc. 45 (2014) 1197–1206, https://doi.org/10.1002/jrs.4520.
[25] M.J. Baxter, Archaeological data analysis and fuzzy clustering, Archaeometry 51 (2009) 1035–1054, https://doi.org/10.1111/j.1475-4754.2008.00449.x.
[26] S. Cardinal, Sets, graphs, and things we can see: a formal combinatorial ontology for empirical intra-site analysis, J. Comp. App. in Arch. 2 (1) (2019) 56–78, https://doi.org/10.5334/jcaa.16.
[27] J. Monnier, L. Bellot-Gurlet, D. Baron, D. Neff, I. Guillot, P. Dillmann, A methodology for Raman structural quantification imaging and its application to iron indoor atmospheric corrosion products, J. Raman Spectrosc. 42 (2011) 773–781, https://doi.org/10.1002/jrs.2765.
[28] P. Colomban, A. Tournié, On-site Raman identification and dating of ancient/modern stained glasses at the Saint-Chapelle, Paris, J. Cult. Herit. 8 (2007) 242–256, https://doi.org/10.1016/j.culher.2007.04.002.
[29] S. Blomberg, T. Garland, A. Ives, Testing for phylogenetic signal in comparative data: behavioral traits are more labile, Evolution 57 (2003) 717–745, https://doi.org/10.1111/j.0014-3820.2003.tb00285.x.
[30] M. Offroy, L. Duponchel, Topological data analysis: the next big data exploration tool in biology, analytical chemistry and physical chemistry, Anal. Chim. Acta 910 (2016) 1–11, https://doi.org/10.1016/j.aca.2015.12.037.
[31] S. Gourvénec, C. Lamotte, P. Pextiaux, D.L. Massart, Use of the orthogonal projection approach (OPA) to monitor batch processes, Appl. Spectrosc. 57 (2003) 80–87.
[32] L. Valderrama, R.P. Gonçalves, P.H. Março, D.N. Rutledge, P. Valderrama, Independent components analysis as a means to have initial estimates for multivariate curve resolution-alternating least squares, J. Adv. Res. 7 (5) (2016) 795–802, https://doi.org/10.1016/j.jare.2015.12.001.
[33] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, J. Chemometr. 15 (2001) 627–646, https://doi.org/10.1002/cem.654.
[34] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (14) (1994) 1425–1432, https://doi.org/10.1016/j.chemolab.2004.06.009.
[35] F. Cuesta Sánchez, J. Toft, B. van den Bogaert, D.L. Massart, Orthogonal projection approach applied to peak purity assessment, Anal. Chem. 68 (1) (1996) 79–85, https://doi.org/10.1021/ac950496g.
[36] E.C. Muñoz, F. Gosetti, D. Ballabio, S. Andò, O. Gómez-Laserna, J.M. Amigo, E. Garzanti, Characterization of pyrite weathering products by Raman hyperspectral imaging and chemometrics techniques, Microchem. J. 190 (2023) 108655, https://doi.org/10.1016/j.microc.2023.108655.
[37] R. Vitale, S. Hugelier, D. Cevoli, C. Ruckebusch, A spatial constraint to model and extract texture components in Multivariate Curve Resolution of near-infrared hyperspectral images, Anal. Chim. Acta 1095 (2020) 30–37, https://doi.org/10.1016/j.aca.2019.10.028.
[38] A. de Juan, M. Maeder, T. Hancewicz, R. Tauler, Use of local rank-based spatial information for resolution of spectroscopic images, J. Chemometr. 22 (2008) 291–298, https://doi.org/10.1002/cem.1099.
[39] P. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636, https://doi.org/10.1021/ac034173t.
[40] H. Martens, S. Jensen, P. Geladi, Multivariate linearity transformation for near-infrared reflectance spectrometry, in: Proc. Nordic Symposium. Of Applied Statistics, 1983, pp. 205–234 (Norway: Stokkland Forlag).

[41] A. de Juan, M. Maeder, T. Hancewicz, L. Duponchel, R. Tauler, in: R. Salzer, H. W. Siesler (Eds.), Chemometric Tools for Image Analysis in Infrared and Raman Spectroscopic Imaging, Wiley-VCH, Weinheim, Germany, 2009, pp. 65–106. Ch. 2.
[42] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics (Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.
[43] D. Demirović, An implementation of the mean shift algorithm, Image Process. Line 9 (2019) 251–268, https://doi.org/10.5201/ipol.2019.255.
[44] K. Arai, A.R. Barakbah, Hierarchical K-means: an algorithm initialization for K-means, Rep. Fac. Sci. Engrg. 36 (1) (2007) 25–31. Saga University.
[45] M. Mittal, R.K. Sharma, V.P. Singh, Validation of k-means and threshold based clustering method, Int. J. Adv. Technol. 5 (2014) 153–160.
[46] W. Windig, D.A. Stephenson, Self-modeling mixture analysis of second-derivative near-infrared spectral data using the simplisma approach, Anal. Chem. 64 (1992) 2735–2742.
[47] S. Gourvénec, D.L. Massart, D.N. Rutledge, Determination of the number of components during mixture analysis using the Durbin-Watson criterion in the orthogonal Projection Approach and in the SIMPLe-to-use Interactive Self-modelling Mixture Analysis approach, Chemometr. Intell. Lab. Syst. 61 (2002) 51–61, https://doi.org/10.1016/S0169-7439(01)00172-1.
[48] K. De Braekeleer, D.L. Massart, Evaluation of the orthogonal projection approach (OPA) and the SIMPLISMA approach on the Windig standard spectral data sets, Chemometr. Intell. Lab. Syst. 39 (1997) 127–141, https://doi.org/10.1016/S0169-7439(97)00060-9.
[49] G. Chikvaidze, N. Mironova-Ulmane, A. Plaude, O. Sergeev, Investigation of silicon carbide polytypes by Raman spectroscopy, Latv. J. Phys. Tech. Sci. 51 (3) (2014) 51–57, https://doi.org/10.2478/lpts-2014-0019.
[50] W.D. Bischoff, S.K. Sharma, F.T. Mackenzie, Carbonate ion disorder in synthetic and biogenic magnesium calciltes: a Raman spectral study, Am. Mineral. 70 (1985) 581–589.
[51] J. Urmos, S.K. Sharma, F.T. Mackenzie, Characterization of some biogenic carbonates with Raman spectroscopy, Am. Mineral. 76 (1991) 641–646.
[52] R.G.C. Bathurst, Carbonate Sediments and Their Diagenesis, second ed., Elsevier, Amsterdam, 1975.
[53] K.E. Chave, Aspects of the biogeochemistry of magnesium 1, Calcareous marine organisms, J. Geol. 62 (1954) 266–283.
[54] K.E. Chave, Aspects of the biogeochemistry of magnesium 2, Calcareous sediments and rocks, J. Geol. 62 (1954) 587–599.
[55] Y. Suffren, F.-G. Rollet, O. Levasseur-Grenon, C. Reber, Ligand-centered vibrational modes as a probe of molecular and electronicstructure: Raman spectroscopy of cis-Fe(1,10-phenanthroline)2(NCS)2 and trans-Fe(pyridine)4(NCS)2 at variable temperature and pressure, Polyhedron 52 (2013) 1081–1089, https://doi.org/10.1016/j.poly.2012.06.070.
[56] R.L. Frost, J. Bouzaid, I.S. Butler, Raman spectroscopic study of the molybdate mineral szenicsite and comparison with other parageneticaly related molybdate minerals, Spectrosc. Lett. 40 (4) (2007) 603–614, https://doi.org/10.1080/00387010701301220.
[57] G. Zeng, K.-K. Li, H.-G. Yang, Y.-H. Zhang, Micro-Raman mapping on an anatase TiO2 single crystal with a large percentage of reactive (0 0 1) facets, Vib. Spectrosc. 68 (2013) 279–284, https://doi.org/10.1016/j.vibspec.2013.08.012.
[58] S. Kos, M. Dolenec, J. Lux, S. Dolenec, Raman microspectroscopy of garnets from S-fibulae from the archaeological site lajh (Slovenia), Minerals 10 (4) (2020) 325–336, https://doi.org/10.3390/min10040325.
[59] R.C. Lord, G.J. Thomas, Raman spectral studies of nucleic acids and related molecules—I Ribonucleic acid derivatives, Spectrochim. Acta Mol. Spectros 23 (9) (1967) 2551–2591, https://doi.org/10.1016/0584-8539(67)80149-1.
[60] A. Raj, K. Raju, H.T. Varghese, C.M. Granadeiro, H.I.S. Nogueira, C.Y. Panicker, IR, Raman and SERS spectra of 2-(methoxycarbonylmethylsulfanyl)-3,5-dinitrobenzene carboxylic acid, J. Braz. Chem. Soc. 20 (3) (2009) 549–559, https://doi.org/10.1590/S0103-50532009000300021.