

An overview of Principal Components Analysis approaches in Raman studies of cultural heritage materials

Alessia Coccato¹ | Maria Cristina Caggiani² 

¹Faculty of Classics, Ioannou Centre for Classical and Byzantine Studies, University of Oxford, Oxford, UK

²Department of Biological, Geological and Environmental Sciences, University of Catania, Catania, Italy

Correspondence

Alessia Coccato, Faculty of Classics, Ioannou Centre for Classical and Byzantine Studies, University of Oxford, 66 St Giles', Oxford OX1 3LY, UK.
Email: alessia.coccato@classics.ox.ac.uk

Maria Cristina Caggiani, Department of Biological, Geological and Environmental Sciences, University of Catania, Catania, Italy.
Email: mariacristina.caggiani@unict.it

Abstract

The present overview answers the need of assessing the current state of the art concerning the application of principal components analysis (PCA) to Raman spectroscopy investigations of cultural heritage and related materials. An increment of the employment of this multivariate statistic technique to Raman results in the mentioned field began between 15 and 10 years ago, after a very slow start at the turn of the millennium. A delay of about a decade was observed with respect to PCA applied to elemental quantitative data of archaeometric analyses, likely a consequence of the required spectral pre-treatment and to results of complex interpretation. Therefore, it is by now the time to summarize this evolution in a comprehensive, yet very specific way. In this overview, painting constituents were considered, both colouring materials and binders, in addition to natural and synthetic glasses, and biogenic and mineral gemmological materials. A marked unbalance between the studies pertaining to the different sections has been noticed, revealing a concentration of the work mainly on painting materials, including the study of ageing and alteration. The different aims of PCA application to Raman spectra, the various approaches and the achievable results, with the possible arising problems, were underlined, too. Special attention was given to the pre-treatment of the spectra, which was observed to be essential to overcome the influence of several issues concerning bands intensity, spectral noise, background, fluorescence and so on.

KEYWORDS

gemmology, glass, painting materials, PCA, statistical analysis

1 | BACKGROUND

Multivariate statistical methods are nowadays extensively used in the data treatment of archaeometric and conservation-connected analyses. At the end of the 20th

century, when conservation of cultural heritage was still considered an emerging area within scientific and interdisciplinary studies, few works including the use of statistical methods could be found,¹ mainly concerning pottery and glass, with the first examples of applications on mortars in wall paintings and archaeological textiles, but the number of papers using statistical methods was gradually

No data were generated for the present paper.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Journal of Raman Spectroscopy* published by John Wiley & Sons Ltd.

increasing.² Two reviews of 2006^{3,4} already registered a wide employment of statistical methods in archaeometric studies and included principal components analysis (PCA) among the most commonly used,³ though Madariaga⁵ highlighted that few studies concerned the use of chemometric methods in the characterization of building materials and/or outdoor exposed artworks degradation.

PCA is included within the so-called unsupervised learning, or pattern recognition approaches, for which the aim is mainly that of identifying a hidden structure—such as a pattern or a grouping—in the dataset, without a preliminary knowledge of the pattern itself.³ It works as a multidimensional scaling (MDS) method, as it transforms a high dimensional data matrix into a lower dimensional one, reducing a large set of variables into a smaller set of orthogonal ones, so called principal components (PCs), oriented according to the maximal variation, thus reducing the number of dimensions to be considered but keeping the maximum of the useful information.⁶

The unsupervised approach is employed when a supervising guidance (labelling) is not available: apart PCA, various cluster analysis algorithms (e.g., K-means and hierarchical cluster analysis [HCA]) belong to this group of methods. Supervised methods, instead, label the classes to be discriminated and work with two subsequent phases: the training phase and the prediction (testing) one. The former uses a training—labelled—dataset to find the pattern; in the latter, the data excluded from the training set are validated, using for example discriminant analysis (DA), multiple linear regression (MLR), principal components regression (PCR), partial least squares (PLS) and support vector machines (SVMs). On one hand, supervised methods have the disadvantage of depending on labelling data, which could become inconvenient when dealing with a high number of observations; on the other hand, unsupervised methods, like PCA, are less effective in discrimination.⁶ For this reason, PCA, though also working as a stand-alone approach, is often used as a starting point for further supervised, as well as unsupervised data treatments.^{6–8}

The basic assumption of PCA is that variation in the dataset contains information and, at the same time, that some variables might be superfluous or redundant in describing it. PCs are in fact calculated as linear combinations of the original variables, and they are ranked according to the largest variation in the data, that is, most information. Once PC1 has been identified, PC2 explains the largest residual variation and is orthogonal to the first and so on. Usually, a few PCs are sufficient to capture the most information from the dataset. The so-obtained reduced number of features simplifies interpretation of the dataset. The calculated scores can be plotted on the new orthogonal space defined by the PCs. A clearer

understanding of the impact of the different variables can be obtained by plotting the loadings in the PCs space. Variables are strongly correlated when they are in the same direction from the origin, and the smaller the angle with an axis, the stronger its impact on the represented PC. Finally, it is important to work with homogeneous datasets, to ensure the maximum information is not influenced by outliers. The model built by PCA is strongly dependent on the starting data. PCA itself can be used to identify and remove outliers from the dataset in an iterative process and to separate signals from noise.^{3,7} It clearly emerges that PCA is useful for data treatment when a large number of variables are available, for example, in the analytical results. Spectroscopic data can be considered a typical example; here, in fact, all the wavenumbers of a given spectral region (hundreds or thousands) constitute the variables, and all the spectra acquired are the different observations. Therefore, multivariate analysis can be extremely useful to keep the maximum quantity of the spectroscopic information, while simultaneously handling a large and complex dataset, where the whole spectral variance, constituted of thousands of spectral channels, can be explained by the first few dominant PCs.⁶ The benefits of applying PCA, apparently the most frequently used of the multivariate statistical techniques for groups identification, among spectral data⁹ such as Raman or infrared ones, are clear, allowing for example to group/separate spectra that are very similar to each other's and may exhibit minimal differences difficult to be spotted by visual observation. Nevertheless, the use of PCA for the interpretation of Raman spectroscopy results, in particular in the field of cultural heritage investigation, has been for long time limited,¹⁰ especially if compared with its application to elemental analyses. The reason can be mainly found in the necessary pre-processing of the raw data.⁶ A first necessary preprocessing requires the alignment of all the spectra on a common *x*-axis (interpolation) and potentially data reduction. Next to this, every spectrum should be also appropriately corrected for spike removal, broad fluorescence background and noise reduction. Often these latter are a combination of Raman cross section of the materials, mixtures of pigments and the presence of binders, focus of the laser, analytical set-up, orientational effects and so on.¹¹ Moreover, in Raman spectroscopy, as the intensity of the signal is affected by numerous parameters, it is important to assess which normalization technique is best suited for the problem at hand. Sometimes the intensity of the spectrum is standardized (max intensity = 1) to limit the effect of laser power, measurement time and focus.¹¹ Additionally, the pre-treated spectra have to be normalized,¹¹ or auto-scaled, so that the variables are centred on the origin, and their variance is

equal to 1. At this point, PCs can be calculated. Furthermore, the visual approach to loadings is complicated in spectroscopy, as each signal corresponds to numerous wavenumbers. Most commonly, the loading values are plotted against the variables, resulting in a spectrum-like plot. In the case of identification of materials based on a database of spectra, the quality and completeness of the reference collection are also crucial. In the latter case, scores for new data can be easily calculated for projecting them into the reduced space for predictive purposes.^{3,7}

Multivariate analysis of Raman spectra has become common more recently, also in relation with increasing computational power of computers: a review of 2016⁸ reports a fair number of works employing chemometric techniques for the study of Raman spectra of cultural heritage materials, indicating PCA together with HCA and PLS-DA as by far the most used ones in the field. Centeno in 2016¹² reviewed seven papers dealing with PCA for the purpose of artistic materials identification with Raman spectroscopy. On the other hand, among the papers concerning cultural heritage in a special issue of Environmental Science and Pollution Research devoted to Multivariate Analysis and Chemometrics in Cultural Heritage and Environment Fields, none involved Raman spectroscopy.¹³ In a review published in 2020, Chiriu et al.¹⁴ included two papers specifically dealing with PCA on Raman spectroscopic data. According to Edwards et al.,¹⁰ the perspectives for the near future are the exploitation of multivariate analysis for Raman imaging and the combination of data coming from different, or hyphenated instruments at the same time, also known as data fusion, while there are already some examples of combination of Raman results with those of other elemental (X-ray fluorescence [XRF], laser-induced breakdown spectroscopy [LIBS]) and molecular (fiber optics reflectance spectroscopy [FORS], infrared) techniques.

The present work is aimed at providing an overview on the use of PCA for the interpretation of Raman spectroscopy results concerning cultural heritage materials. Table 1, summarizing the papers by studied materials according to this overview's structure, indicates relevant pretreatments used in each work, as clearly stated by the individual authors. Proportionally to the papers published up to the present day, a large part will concern painting materials, mainly colouring substances, both inorganic and organic, and binders and paints; smaller sections will regard other geomaterials like glass, both natural (obsidian) and synthetic, as well as gemmological materials, including the biogenic ones such as amber, ivory and pearls (Table 1). The different approaches will be described, comprising the pre-processing, the type of application (chosen spectral parameters used as variables or the spectra themselves) and the combination with

other techniques (separately or through data fusion). Only for the colouring materials, however, a separation based on these criteria will be followed strictly due to the much higher number of reported papers. The benefits achieved in the interpretation will be discussed.

2 | COLOURING MATERIALS

The wide variety of pigments used in cultural heritage since prehistory, their application to various supports with different binders, the onset of alteration processes depending on the used materials and their exposure to the atmosphere create numerous analytical challenges for the archaeometrist. Identification of pigments, alone or as mixtures with other pigments and/or binders, which might have undergone alteration processes, in spot analyses or in micro- or macroscopic imaging approaches are all aspects where PCA has been successfully applied to Raman spectroscopy. Namely, PCA has been applied to pre-treated spectra (or to a limited range of wavenumbers), sometimes after calculation of first and second order derivatives, to spectral parameters (band positions, widths, areas), and in combination with datasets coming from other complementary techniques (data fusion). These three types of datasets will be discussed separately in the following.

2.1 | PCA on Raman spectra

Synthetic and natural organic colourants were the first class of materials where PCA was successfully applied, both for identification purposes and as data reduction for further chemometrical approaches. The first application of PCA to pigmenting agents was proposed by Vandena-beele et al.,¹¹ who specifically developed a PCA-based search algorithm for the Raman spectra of synthetic organic pigments (SOPs). This appeared as an ideal field of application of chemometrical approaches to Raman spectra, as the range of SOPs is wide, each showing numerous Raman bands that make the task of automated identification very desirable. In this case, micro-Raman spectra were acquired both on reference samples, on test materials (commercially available crayons) and on micro-samples from a painting by Delvaux (20th century). A 780-nm laser with a 1200 grooves/mm grating was used, allowing to obtain a spectral resolution of 1 cm⁻¹. The algorithm computed the PCs of the reference materials and of the unknown spectrum, both appropriately pre-treated (6th grade polynomial baseline correction, linear interpolation to reduce data points by a factor 10, 300–1800 cm⁻¹ range extraction), and then the Euclidean distance between the PCs of the two sets was used for identification purposes, as it measures similarity. The authors

TABLE 1 Summary of the papers cited in the manuscript, according to the type of materials studied. Type of samples and pre-processing of spectra for PCA are reported according to the explicit indication of the individual authors. 'No' indicates the unsuccessful application of the selected pre-treatment; 'yes/no' refers to a comparison of results with and without the selected pre-treatment.

	Samples		Pre-processing							First derivative
	Reference/ natural samples	Real cultural heritage samples	Baseline	Interpolation	Data reduction	Smoothing	Selection of range	Spectral decomposition		
Colouring agents	Spectra	Vandenabeele et al. (2001)	X	X	X		X			
Colouring agents	Spectra	Vandenabeele et al. (2003)	X				X		X	
Colouring agents	Spectra	Daly et al. (2018)	X	X			X		X	
Colouring agents	Spectra	Daly et al. (2019)	X	X			X		X	
Colouring agents	Spectra	Piantanida et al. (2013)	X				X			
Colouring agents	Spectra	Capone et al. (2021)	X					X		
Colouring agents	Spectra	Castanys et al. (2006)	X							
Colouring agents	Spectra	Castanys et al. (2011)	X		X					
Colouring agents	Spectra	Gonzalez-Vidal et al. (2016)	X	X						
Colouring agents	Spectra	Gonzalez-Vidal et al. (2015)	X	X						
Colouring agents	Spectra	Navas et al. (2010)	X				X		No	
Colouring agents	Spectra	Romero Pastor et al. (2011)	X				X		No	
Colouring agents	Spectra	Otero et al. (2014)	X	X			X			
Colouring agents	Spectra	Lofrumento et al. (2012)	X							
Colouring agents	Spectra	D'Elboux Bernardino et al. (2014)	X		X				X	
Colouring agents	Spectra	Festa et al. (2022)	X	X		X				
Colouring agents	Spectra	Martins et al. (2021)	X	X					X	
Colouring agents	Spectra	Pozzi et al. (2013)	X	X		X	X			
Colouring agents	Spectra	Doherty et al. (2014)	X				X			
Colouring agents	Spectra	Marengo et al. (2004)	X		X	X				
Colouring agents	Spectra	Marengo et al. (2005)	X			X				
Colouring agents	Spectra	Vandenabeele and Rousaki (2021)	X	X	X					
Colouring agents	Spectra	Colomban et al. (2023)	X	X						
Colouring agents	Parameters	Tomasini et al. (2012)	X							
Colouring agents	Parameters	Botticelli et al. (2020)	X			X				
Colouring agents	Parameters	Colomban et al. (2023)	X	X				X		
Colouring agents	Data fusion	Ramos et al. (2008)	X			X				
Colouring agents	Data fusion	Deneckere et al. (2011)	X	X						
Colouring agents	Data fusion	Pallipurath et al. (2013)	X			X	X			
Colouring agents	Data fusion	Pallipurath et al. (2014)	X							
Colouring agents	Data fusion	Carlesi et al. (2016)	X	X					X	
Colouring agents	Data fusion	Carlesi et al. (2018)	X	X			X		X	

TABLE 1 (Continued)

			Samples		Pre-processing						First derivative
			Reference/ natural samples	Real cultural heritage samples	Baseline	Interpolation	Data reduction	Smoothing	Selection of range	Spectral decomposition	
Binding media		Nevin et al. (2007)	X			X	X		X		
Binding media		Nevin et al. (2008)	X			X			X		
Binding media		Romero Pastor et al. (2013)	X								
Binding media		Manzano et al. (2012)	X		X						
Binding media		Otero et al. (2014)	X	X	X				X		
Binding media		Daher et al. (2013)	X						X	X	
Glassy materials	Obsidian	Carter et al. (2009)	X	X				X			
Glassy materials	Obsidian	Carter et al. (2012)	X	X							Yes/no
Glassy materials	Obsidian	Kelloway et al. (2010)	X	X							X
Glassy materials	Obsidian	McMillan et al. (2019)	X	X							
Glassy materials	Glass and glazes	Colomban et al. (2006)	X							X	
Glassy materials	Glass and glazes	Van Pervenage et al. (2020)		X						X	
Glassy materials	Glass and glazes	Ricci et al. (2007)	X	X							
Gemmological materials	Biogenic	Teodor et al. (2010)	X	X							
Gemmological materials	Biogenic	Badea et al. (2015)	X		X				X		
Gemmological materials	Biogenic	Barone et al. (2016)	X								
Gemmological materials	Biogenic	Peris Diaz et al. (2018)	X	X			X				
Gemmological materials	Biogenic	Daher et al. (2013)	X	X	X				X	X	
Gemmological materials	Biogenic	Shimoyama et al. (1997)	X						X		
Gemmological materials	Biogenic	Brody et al. (2001)	X		X				X		
Gemmological materials	Biogenic	Park et al. (2009)	X		X				No		
Gemmological materials	Biogenic	Stenger et al. (2021)	X		X			X			
Gemmological materials	Mineral	Bi et al. (2015)	X				X				
Gemmological materials	Mineral	Chang et al. (2016)	X						X		
Gemmological materials	Mineral	Liu et al. (2022)	X		X						
Gemmological materials	Mineral	Dumańska-Słowik et al. (2020)	X		X			X			
Gemmological materials	Mineral	Caggiani et al. (2023)		X	X			X	X		X
Miscellaneous		Zoppi et al. (2008)	X	X	X					X	
Miscellaneous		Bianchi et al. (2016)	X					X		X	
Miscellaneous		Quintero Balbas et al. (2021)	X	X	X				X		X
Miscellaneous		Simonetti et al. (2016)		X							

TABLE 1 (Continued)

[illegible]

TABLE 1 (Continued)

	Pre-processing										Total
	Second derivative	Standard normal variate	Normalization	Least square weighing	Mean centring	Weight standardization	Standardization	Centring	Scaling	Autoscaling	
Binding media					X			X	X		
Binding media			X								
Binding media		X		X	X						
Binding media											
Glassy materials			X								6
Glassy materials											
Glassy materials											
Glassy materials											4
Glassy materials											
Glassy materials			X								
Glassy materials									X		3
Glassy materials											7
Gemmological materials											
Gemmological materials			X								
Gemmological materials											
Gemmological materials											
Gemmological materials	X			X							
Gemmological materials											
Gemmological materials			X								
Gemmological materials			X								
Gemmological materials			X								9
Gemmological materials			X								
Gemmological materials											
Gemmological materials			X								
Gemmological materials			X								
Gemmological materials			X								
Gemmological materials	X		X								5
Gemmological materials											14
Miscellaneous											
Miscellaneous					X						
Miscellaneous	X	X						X			
Miscellaneous		X									4

stressed how a successful identification clearly depends on the quality of the spectra and on the completeness of the database, highlighting as well some issues on the discrimination of different phthalocyanine blues (PB15). In a later study by the same research group,¹⁵ PCA was used as a data reduction method before applying supervised (linear discriminant analysis [LDA]) and unsupervised (HCA) pattern recognition methods, in order to be able to discriminate between synthetic and natural indigo samples based on their micro-Raman spectra (785 nm laser, 1200 grooves/mm grating, spectral resolution of 1 cm^{-1}), which is not evident by visual observation of the spectra. First it was used on autoscaled Raman spectra to monitor the dye obtained in different conditions. It was observed that the fluorescence background strongly affected the spectra, leading to misclassifications. Hence, first and second derivatives were used, together with autoscaling before PCA. PCA on the Raman spectra, however, only showed partially satisfactory results, but it proved an essential step in spectral processing and for further unsupervised and supervised multivariate analyses, as HCA and LDA, respectively. HCA proved successful in identifying synthetic indigo samples using the first 3 PCs. Classification improved when LDA was applied on PCs 2 to 4 from the raw spectra, and on the first 4 PCs from the second derivative spectra. The authors note that interpretation of the loadings of the latter is complicated by the shape of peaks' second derivatives.

Carbon black pigments are another interesting field of application of PCA, as the pigments show broad bands ascribable to different carbon structures, and the nomenclature does not correspond specifically to geological materials.¹⁶ Moreover, as these materials have been continuously used since prehistory, they are relevant trackers for technological developments and artistic choices through time, by comparison with relevant references. Both the spectra and spectral parameters (see Section 2.2) have been subjected to PCA. Carbonaceous black drawing materials on paper have been subjected to investigations by Daly et al.^{17,18} Subtle differences in the Raman spectra acquired directly on such materials are hard to detect, especially in the presence of fluorescence due to supports, binders and fixatives. Both reference materials from historical manufacturers (14 in total) and nine drawings from the J. Paul Getty Museum (7 by 19th century French artists published in 2018¹⁷ and 2 more by Odilon Redon in 2019¹⁸) have been studied with 785-nm excitation, as tests with 514-nm laser were dominated by fluorescence. PCA has been applied on the range $800\text{--}2000\text{ cm}^{-1}$. The spectra were also smoothed and corrected by a multiplicative scatter correction using the mean of the whole dataset, normalized by the D band intensity, and the first derivative was calculated, which

proved effective in reducing the background contribution without affecting the G and D bands. A further least square weighing and mean centring have been applied to the dataset to further reduce spectral noise. Outliers were identified and removed by checking their residuals, and the model revised iteratively. More than 150 spectra of the reference materials were used to build the model. It appeared that positive PC1 values are associated to materials with carbon bands that are either closer to one another, or broader, as in the case of charcoals, with chalks having the opposite behaviour (well separated bands). PC2 was linked with the G band profile: positive values are observed for waxy and oily pigments; an additional contribution is linked to higher background intensities. It seemed that PC2 was therefore related to commercial drawing materials, where the presence of a binder modifies the background of the Raman spectrum, and its positioning in the PCA space. The authors stressed the importance of a reference dataset consistent with the expected artistic materials, reflecting the materials available in their historical context.¹⁷ The interpretation of the drawings has also included historical considerations, such as the preferential use of commercial materials by artists and the changes in production techniques and recipes. A variety of commercial drawing materials (up to six in Redon's 'Apparition'), both with high and low contents of binding agents, have been observed, corresponding to different areas of the works of art, according to the desired colour and texture, and to the different stages of the creative process.¹⁸ It has been moreover observed that, in Redon's pastels, the use of a fixative to protect the underlying layers from smudging is recurring, which affects the Raman spectra by increasing fluorescence and producing skewed PCA coordinates. Although these works were successful in identifying a variety of materials coherent with both general trends in pigments use and manufacture, and in the artists' practice, they suggest care in using the proposed PCA model to works of art of different epochs/geographic origin. Moreover, they suggest complementary techniques for clarifying the role of organic binders in the Raman spectra of drawing materials.

Besides carbon-based pigments, inks are also a successful field of application of PCA. Both identification and authenticity questions were tackled with this approach. Piantanida et al.¹⁹ applied PCA to Raman spectra of iron-based inks, in order to obtain clearer signals for further multivariate analyses, supporting the feasibility of non-destructive analyses for organic compounds in works of art. A 785-nm laser was used to collect spectra. A micro-destructive characterization based on chromatographic techniques was also performed. Visual inspection of multiple spectra acquired on each sample allowed to identify a variety of compounds (CaSO_4 , CuSO_4 , carbon

black) in addition to iron gall inks and sometimes tannins. Even though the spectra recorded were noisy, affected by fluorescence and far from ideal, the authors strongly advocated against smoothing procedures, as they expected the noise to be systematic and related to degraded organic materials. Only linear baseline correction was applied. Moreover, each of the 5 acquisitions was treated as individual samples. Nevertheless, PCA was performed on the region 900–1700 cm^{-1} , and the loading plot allowed to identify only the first 5 PCs as distinguishable from noise. These 5 PCs were further subjected to multivariate analyses (LDA). A combination of Raman and PCA has also been applied in the authenticity quest of the *Artemidoros papyrus*.²⁰ As carbon blacks were often used as writing materials, and their Raman spectra are fairly similar, PCA has been used for a two-step comparison with a spectral library assembled for the scope of the paper, which included both natural and synthetic modern pigments; 170 micro-Raman spectra (785-nm laser, spectral resolution of 3 cm^{-1}) were acquired on all three sections of the papyrus, on both sides, and considering text, geometric and figurative elements. Spectral decomposition was also performed, to estimate each carbon configuration's contribution to the spectrum. Correlation matrices between the reference materials spectra were created and diagonalized to obtain the PCs. Single linkage clustering was later applied to the ink spectra projected on the PC space. A first PCA was carried out as a screening and allowed to discard some reference materials identified as irrelevant for the purpose of the study and to focus on only eight of them. A second PCA was run on this subset, with PC1 accounting for 40% of the variance. The contribution of PC1 was removed from the ink spectra, and the resulting signal decomposed into their individual structural carbon components. A distinction could be made among bituminous materials and artificially produced carbons and another group showing sp^3 diamond hexagonal sites present, for example, in lonsdaleite, a mineral, or in modern manufactured carbon. The inks showed a wide variety of carbon structures, and their distribution does not seem to correspond to specific areas of the papyrus, confuting the hypothesis of its writing during three separate periods of time (1st century BC—1st century AD).

Castanys and co-workers²¹ used PCA in the pre-processing of spectra, after baseline correction and normalization of intensities, to reduce the dimensions of Raman spectra of pigments for identification purposes. Anatase, azurite, chrome yellow, Naples yellow, rutile, ultramarine and vermillion were used as reference spectra for identification of pigments and binary mixtures. The same authors later devised a machine learning methodology to support the analyst,²² also based on PCA of

Raman spectra: here, 20 PCs were obtained from 1599 data points for 21 yellow SOPs. The authors stated that system should be flexible enough to adapt to different spectral libraries.

Binary pigment mixtures were explored by González-Vidal et al.²³ They used PCA to reduce the size of the dataset (baseline corrected, intensity normalized, interpolated spectra). The reference dataset was composed of 20 inorganic and 31 SOPs. The authors stressed the need for an adapted set of reference materials depending on the type, chronology and context of the polychrome object under study. Artificial mixtures of rutile with ultramarine, of PY1 with PR3, of PY1 and PB60, as well as a commercial pigment reported to contain PY1 and PR4, were studied. Fictitious spectra of mixtures were directly built in the reduced space obtained by PCA of the individual pigments spectra and based on similarity criteria of the individual pigments as expressed by the reliability factor (which is a function of the Euclidean distance between the unknown spectrum and those present in the reference database). The authors pointed out that band distortion and fluorescence could negatively affect the results of identification of mixtures and that the results of identification and the reliability factor are intended to guide the analyst in spectral interpretation. González-Vidal et al. also proposed a more advanced version of automatic identification of pigments in mixtures²⁴ that did not rely on the operator's input and that provided a confidence factor to guide the decision-making process. The pre-processing required baseline correction, calibration and interpolation of the x-axis, and a min-max normalization, the latter to improve reproducibility across instruments while maintaining peak intensity ratios. PCA was applied on these spectra as a data reduction tool. Spectra of 53 reference materials were used to build the PC space, and the unknown spectrum projected onto it, with Euclidean distance and squared cosine being used for identification of the best candidates, and to assess the reliability of the identification. For mixtures, an additional step was proposed using independent components analyses (ICAs), which will not be discussed here. A combination of PCA and multiple discriminant analysis (MDA) was also proposed²³ to avoid the repetitive, subjective and time-consuming process of manual classification. A classification space could be built by the user, assigning spectra to reference classes. PCA was used to extract features and MDA to assign the PCA results to classes. Euclidean and Mahalanobis distances (MDs) were used for this purpose. A few tests were performed, showing good results: discrimination between natural and synthetic ultramarine and its identification in mock-ups and real paintings; discrimination of the copper-phthalocyanine blues (α , β , ϵ) by considering both

spectra acquired by the authors and by different researchers, hence comparing different acquisition systems and measurement conditions; a dataset of paint samples containing PB15:1, PB15:3 and PB15:6 was tested to account for the binder.

PCA is also a powerful tool to study paints, both in order to identify the pigments and their interactions with different binders, including the formation of metal carboxylates. This has been applied both to mock-ups at different stages of curing/ageing and to works of art. Navas et al.²⁵ tested nine pigments and their relative tempera paints (egg binder), highlighting how PCA was not successful on the raw micro-Raman spectra (514-nm laser, spectral resolution of 1 cm^{-1} , $200\text{--}3800\text{ cm}^{-1}$ range; acquisition parameters given in the reference), while it showed interesting results on their first derivative. Centring and scaling of the data were both tested, in order to improve the obtainable results on a case-by-case approach. However, the pre-processing complicates the identification of the loadings, as both maxima and minima in the spectrum equal to zero in their first derivative. Also, it is interesting to note that the authors applied PCA on each colour-group separately (white, red, blue). For each colour, a limited spectral range was selected, where indicative bands appear, sometimes discarding the more intense ones as the 401 cm^{-1} in azurite or the 462 cm^{-1} in smalt. Differences between blue pigments and their paint could also be seen, likely as a result of pigment–binder interactions. The laser-induced degradation of red lead could also be tracked in the PCA. Further tests were carried out on pigment–binder interactions by the same research group, using the same micro-Raman spectrometer and the same experimental setup.²⁶ Binary mixtures of egg yolk with cinnabar, raw Sienna, lead white, gypsum, calcite, azurite, lapis lazuli and smalt were prepared and tested separately by colour, focussing on the $2800\text{--}3100\text{ cm}^{-1}$ region. In this case, autoscaling of the spectra before PCA was beneficial as the studied range was narrow. From PCA, it was possible to infer that raw Sienna, cinnabar, lead white, gypsum, azurite and smalt interact with the aminoacidic moieties of the binder, with calcite interacting at a lower degree, and lapis not at all. From a detailed observation of the spectra and of the loadings, the authors were able to propose preferential interactions between metallic cations and amino acids. A detailed study on the organic binders has also been carried out by the same research group²⁷ and is discussed in Section 3.

Micro-Raman spectra of synthetic metal carboxylates and of microsamples of 19th century paintings were successfully identified by PCA.²⁸ The 22 reference materials were synthesized with palmitic, stearic, oleic and azelaic acid, with calcium chloride dihydrate, manganese(II)

chloride tetrahydrate, cadmium chloride hydrate, zinc chloride, copper acetate monohydrate and lead nitrate, in order to simulate the variety of carboxylates one is expected to find in works of art. Micro-Raman analyses were carried out with both 633- and 532-nm lasers. The authors pointed out that Raman bands between 1000 and 1150 cm^{-1} (C–C stretching vibrations) are sensitive to the carbon chain length both for the acids and their metal carboxylates. Also, oleates are easily identified by the C=C band at $1640\text{--}1675\text{ cm}^{-1}$. PCA on palmitates, stearates and azelates was performed on the C–C stretching region at $1040\text{--}1120\text{ cm}^{-1}$ and has allowed to discriminate saturated and dicarboxylate anions. The first PC is related to the azelates, whereas the third allows to group palmitates versus stearates. Raman spectroscopy was successful in discriminating among different chain lengths but less so in the differentiation of different cations coordinated to the same length chain, as this vibration is not always detected. However, Raman microscopy did not allow to discriminate the coordinated ions, which is, in turn, possible by Fourier transform infrared spectroscopy (FTIR). The PCA model was then used for interpreting the painting materials. The authors highlighted however high values for the model's residual statistic, which decreased the confidence level of the assignment, and has been ascribed to low signal to noise ratios of the experimental spectra, compared with the reference ones. Nevertheless, the identification of specific carboxylates, as supported by PCA on Raman spectra, allowed to suggest reaction paths involving the original painting materials.

In the framework of prehistoric rock-art, the use of Raman spectroscopy is beneficial as it is a powerful tool in identifying the red and black pigments commonly used. Nevertheless, unravelling the production technique of such decorations and supporting chronological interpretation of these early demonstration of artistic behaviours is extremely interesting. Haematite-containing pigments from Ethiopian rock paintings,²⁹ separated in two groups on stylistic basis, were studied with micro-Raman spectroscopy (514- and 785-nm lasers, respective spectral resolution of 6 and 2 cm^{-1}). Area-normalized spectra in the range $260\text{--}750\text{ cm}^{-1}$ allowed to support such grouping: it appeared that the older paintings were made using impure haematite, containing some hydrated iron oxides as indicated by the shoulder at about 390 cm^{-1} visible in PC1. PC2 on the other hand appeared linked to a band at ca. 660 cm^{-1} , attributed to disordered haematite (i.e., naturally occurring or artificially produced by heating goethite at $T < 1000^\circ\text{C}$). D'Elboux Bernardino and co-workers studied common iron- and manganese-containing pigments used in rock art and more specifically their effect of the used binders (vegetable and animal fats).³⁰ The catalytic action of Mn^{4+} and Fe^{2+} on organic

molecules appears to be well known, but not fully understood, as inhibition of oxidative reactions and radical formation occur in some cases. The effect of MnO_2 and $\alpha\text{-Fe}_2\text{O}_3$ (containing Fe^{3+} ions) on the degradation of methyl linoleate and fats, including the possible role of light, was studied by accelerated ageing of binders and paints by gas chromatography–mass spectrometry (GC–MS), FTIR and micro-Raman spectroscopy (785 nm). Spectra were baseline corrected and mean normalized for the vegetable fats, whereas for the methyl linoleate reference material, the first derivative was computed, confirming the viability of PCA when combined with a case-by-case approach. The authors also reported excessive fluorescence when using a 633-nm laser. PCA was used to identify the bands most sensitive to degradation, by comparing baseline corrected spectra on intact and aged methyl linoleate in the region $2800\text{--}3050\text{ cm}^{-1}$. In this case, only 2 PCs were considered: PC1 indicated the non-aged material, and PC2 reflected the ageing-induced modifications of the spectrum. The latter showed negative values for decreasing intensity of the original bands (970 , 1263 , 1656 , 3013 cm^{-1}) and positive ones for bands appearing or increasing in intensity (858 , 1640 , 1670 , 1695 , 1720 cm^{-1}). Therefore, bands not appearing in PC2 are likely related to stable chemical groups, as for example the terminal CH_3 at 2904 cm^{-1} . A detailed assignment of Raman (and IR) bands is also provided, which is in good agreement with the expected reactions and molecular rearrangements taking place during ageing, also showing the role of light in the process. MnO_2 appeared to strongly enhance the linoleate degradation, with no synergistic effect of light. PCA was then conducted on the first derivative of baseline corrected spectra ($700\text{--}1800\text{ cm}^{-1}$) of the methyl linoleate set of samples, confirming that spectra of paints are indistinguishable according to their ageing in the darkness or not. It appeared that the metal oxides are the most important factor controlling degradation processes of both animal and vegetable fat.

Polychrome works of art of more recent epochs also benefitted from PCA combined with Raman spectroscopy. In addition to a portable Raman with sequentially shifted excitation (785 and 853 nm), an innovative commercial instrument combining XRF and Raman (785 nm) spectroscopies has been used to characterize a wall painting in northern Sicily, as well as on a set of 48 reference pigments.³¹ FTIR was also applied to the samples; 220 spectra were retained of the 316 acquired, excluding spectra with too much noise and/or no signals (such as mainly green, blue, black and organic pigments and lakes). Linear interpolation, normalization and weight standardization of the Raman spectra have been carried out prior to PCA. All the vibrational spectra have been treated together but separately for inorganic and organic

pigments, over the range $2\text{--}3999\text{ cm}^{-1}$, and this is reflected in the scores plots. SVD was applied to the organic matrices spectra and nonlinear iterative partial least squares (NIPALS) to the inorganic ones. Not all the pigments gave usable spectra; hence, the Raman dataset had to be reduced. However, the contribution of the Raman bands to identification of pigments is not straightforward. The scores plot of PC1 versus PC2 shows a grouping according to the used instrument. The considerations of the authors on the observed groups appearing in the different quadrants are likely affected by a strong variability in the Raman and infrared spectral profile, which affects the calculated PCs. No loading plot is shown about the Raman spectra. The authors of the paper claim effectiveness of the method over non-homogeneous datasets, but in the writers' opinion, such issues would benefit from a different approach, such as data fusion or iterative PCA on consistent datasets from the same analytical technique.

PCA has also been applied to an illustrated book by Henri Matisse,³² within the framework of a non-destructive multi-analytical characterization of paints. Both direct analyses with a portable instrument (sequentially shifted source, 785 and 853 nm), micro-Raman (785 nm) and surface-enhanced Raman spectroscopy (SERS) (532 nm) on samples were carried out. First derivative Raman spectra were calculated and PCA used as an exploratory method on pigments with similar Raman, IR and XRF signatures. Unfortunately, only the raw Raman spectra are shown in the paper and in the supplementary material, and there is no clear description of the PCA results on the Raman spectra. The approach nevertheless allowed to identify a total of 39 gouaches.

In 2013, the first combination of multivariate approaches and SERS of dyes was published. Pozzi et al.³³ stressed how improvements in the technique and its increasing application in a variety of fields highlighted the need for searchable databases. A wide range of archaeological objects, historic commercial catalogues, applied arts specimens, sculptures, musical instruments, paintings and watercolours, as well as a mock-up of a late Ottoman decoration, were subjected to the test. Both FT (1064 nm) and dispersive (488 , 633 and 785 nm coupled with 1800 and 1200 grooves/mm gratings, spectral resolution of $3\text{--}5\text{ cm}^{-1}$) Raman analyses were carried out, on the sample as is or prepared for SERS with Ag nanoparticles. All the spectra were pre-treated (baseline subtraction, normalization, moving-average smoothing) over the range $420\text{--}1675\text{ cm}^{-1}$. PCA on 99 spectra was used to compute the covariance matrix and the 100th spectrum projected in the PC space and assigned to a group based on the Euclidean distance. The number of PCs for identifying unknown spectra was based on the success rate of

identification and assessed on a case-by-case approach. The success rate for PCA significantly increased (ca. 35–76%) when the second derivative of baseline corrected, range extracted, smoothed and normalized spectra was used. However, spectral variability observed in the reference spectra was exploited to compare PCA with a library search method (correlation coefficient algorithm). PCA proved less efficient in terms of classification power, but it was still able to pinpoint the chemical class of all the studied dyes, notwithstanding different sample preparation procedures and acquisition conditions.

Recently after, Doherty et al.³⁴ also used a variety of instruments and sample preparations for vibrational analyses, in combination with chemometrics, for the investigation of 10 synthetic triarylmethane dye powders (green, blue and violet) and respective solutions on paper. The authors highlighted the need of exploiting portable Raman devices in the analysis of cultural heritage and started to build a reference database for the portable instrumentation, with further data processing by means of PCA. Both the laboratory and portable Raman spectrometers were equipped with a 532-nm laser, the latter also having the option of a 785-nm source. The portable one was used for SERS, obtained by silver colloids deposited on the powder or on the dyed paper. Even though this approach should be considered micro-invasive as it stains the paper support, its effectiveness proved unparalleled. PCA was performed on the baseline corrected and normalized spectra in the range 300–1800 cm^{-1} . The authors report a noticeable drop in the quality of the spectra of dyes acquired with the portable Raman spectrometer, even though the 785-nm source was expected to provide lower fluorescence, due to the interference of external illumination, and the lack of a confocal system. SERS allowed to reduce fluorescence and amplify weak signals, which remained fairly broad due to the lower spectral resolution of the portable device (8 vs. 2 cm^{-1}). PCA on the conventional micro-Raman spectra allowed to separate the dyes into three groups: di-amino, di-phenyl-naphthalene and tri-amino dyes based on the first 2 PCs. PCA on SERS spectra further separated the acidic and basic dyes, as a consequence of their different interaction with the silver colloids. When applied on paper, conventional Raman measurements only yielded fluorescence, and SERS allowed to obtain excellent spectra, both on concentrated and diluted dyes. PCA was not applied on the Raman spectra of the paper samples.

Besides identification of pigments and paints, PCA was also included as part of the design of control charts for monitoring the early onset of alteration processes on pigments, based on accelerated ageing experiments. Marengo et al.^{35,36} exploited a multivariate approach that took into account Fourier transform Raman (FT-Raman)

spectra (1064-nm laser, 4 cm^{-1} spectral resolution, 50–4000 cm^{-1}) to reduce the large number of correlated variables (the intensity for each registered wavenumber) by means of PCA. Spectra were acquired before and after exposure to acidic environment and ultraviolet (UV) light of lead chromate PbCrO_4 ³⁵ as well as ultramarine and red ochre³⁶ in linseed oil on canvas at set time intervals. Baseline correction was applied, and a smoothing procedure was used to reduce the number of variables by a factor of 20. In each case, PCA was applied on the whole dataset (unexposed and exposed), allowing to highlight some spectral changes. These affected both the binder, the support and the pigments. Furthermore, PCs were also used to develop control charts that clearly show the first appearance of alterations.

Additionally, PCA appears to be a widely used method when dealing with Raman imaging; however, no examples are given on the specific topics of archaeometry in a recent review.³⁷ Vandenabeele and Rousaki³⁸ have exploited PCA to process almost 20 000 baseline corrected and normalized spectra from a 785-nm macro-Raman mapping of a 70 × 68.5 mm area with a step of 500 μm , whose acquisition lasted approximately 15 h. PCA has been applied as a data reduction step. The loadings of the first 4 PCs showed resemblance with three identifiable pigments, and one of them had no evident interpretation. The visualization of the Raman signals has been achieved by mapping the scores of each PC: each pixel contained information on the PCs and allowed to map the used pigments (copper phthalocyanine PB15 and lithol rubine PR57:1) in a printed cartoon, revealing the presence of mixtures. Signals somehow related to the paper and to the support (calcite) were also visible. The yellow pigment could not be identified, as each acquisition lasted 1 s only, which was probably not enough for detecting its Raman signals.

Finally, a variety of yellow inorganic synthetic materials used for colouring glazes in European and Chinese artefacts (17th–18th centuries) was the subject of a dedicated PCA study of the Raman spectra,³⁹ both as such and as parameters (see Section 2.2), and including considerations on the glassy matrix (see Section 4.2). The spectra with the lowest contribution from the glassy matrix were selected and corrected with singular normal variate transformation, but in the end, this spectral approach seemed not the best-suited to distinguish the contribution of the pigments.

2.2 | PCA on spectral parameters

A limited range of pigments has been studied by applying PCA on spectral parameters. As already mentioned, carbon

blacks, with their broad bands and multiple possible origins, are well suited for that, and the precious cinnabar showed promising results. Yellow pigments in glazes represent a very recent application of such an approach.

Tomasini et al.⁴⁰ have used the band positions and full width at half maximum (FWHM) of the G and D bands of carbonaceous commercial pigments, as well as the intensity ratio of the two bands to discriminate among 8 reference materials (10 points per sample, 14-nm laser, 0.1 to 2 W/mm², 1800 gr/mm, spectral resolution of 1.5 cm⁻¹), and to identify two samples issued from early 18th century colonial paintings on different supports. Both PCA and HCA were used, and the results were confirmed by scanning electron microscopy—Energy Dispersive X-ray Spectroscopy (SEM-EDS) morphological and chemical data.

Raman spectral parameters of a variety of cinnabar samples have been subjected to PCA by Botticelli et al.⁴¹ A 633-nm source was used in combination with a 1200 grooves/mm grating, obtaining a spectral resolution of 2.7 cm⁻¹. Spectral pre-treatment included baseline correction, smoothing and intensity normalization. Second derivative was applied to reveal the expected doublets of the E degenerate modes of HgS, with a Gaussian profile used for band decomposition. The retained values were peak area by integrating data, FWHM, peak maximum height, peak gravity centre and peak area by integrating data (%). Peak height was not considered as it is affected by instrumental parameters. FWHM and the peak area were considered separately as they are dependent variables. Two approaches were tested on multiple replicates on each sample, as they were either averaged or treated individually as different specimens. The correlation matrix method was preferred for PCA. In the first case, differences could be highlighted between Chinese and European cinnabar, as well as among the European samples, and between synthetic and natural ones. Nevertheless, the dataset was strongly reduced as a consequence of the non-resolvable doublet in numerous cases. PCA was then carried out on individual spectral features of European cinnabar only (relative area and position of the bands at ~252, ~342 and ~350 cm⁻¹). Samples from Carnia, Idrija and Almadén could be discriminated based on the position of the band at 342 cm⁻¹ and the relative area of that at 350 cm⁻¹. These two latter sources have shown the same sulfur isotopic signature and could not be discriminated in former studies. Moreover, Carnia and Idrija show detectable differences even though they are geographically close. Both synthetic and Chinese cinnabar appeared easily identifiable. However, an effect of morphology (powder vs. bulk) has been observed on the Raman spectral parameters, which should be taken into account when dealing with pigments and paints.

Also, significant differences between artificial mercuric sulphide from historical productions and contemporary ones are evident, so that the identification of the former might be challenging.

The already mentioned yellow pigments used in glazes³⁹ could not be successfully discriminated based on their spectra, due to the unavoidable signal from the silicatic glaze. Raman band positions, widths and areas of phases belonging to 'Naples yellows' of known composition/mineralogy at around 120–140, 330, 450 and 510 cm⁻¹ were considered from literature. Data were centred but not scaled for this PCA. Such classification proved much more effective in discriminating lead stannate phases (Pb₂SnO₄–PbSnO₃) of pyrochlore structure and speciation of Sn (Pb₂Sn_xO_{6-δ}); a second group includes Pb₂Sb₂O₇ with Sn, Si and Zn substituting Sb; a third group is composed of lead antimonates with either Fe, Zn or Sn. The first two PCs allow to discriminate both chemistry and oxygen stoichiometry. Raman parameters from spectra acquired by the authors were then included in the PCs space built from the references. A group of green Chinese enamels clearly differs from the others and has no correspondence with the references. HCA with Euclidean distance and Ward linkage confirms these results. Both geographical provenance and manufacture technology support these findings.

2.3 | PCA in data fusion

An early proposal of micro-Raman and XRF data fusion for pigments analysis was published by Ramos et al. in 2008,⁴² specifically dealing with ochres and testing a new hyphenated instrument performing the two analyses on the same spot (633- and 785-nm lasers for the Raman analyses, 150–2000 cm⁻¹). Natural ochres show a wide range of colours, both depending on their mineralogical composition (iron oxides and oxi-hydroxides, other minerals) as well as their particle size. The spectra were subjected to smoothing, de-noising, background correction and normalization. As the data came from two different sensors, both a low- and a mid-level fusion were possible, that is, respectively, the fusion of the raw data, or of the most relevant features. PCA was used as an intermediate step, before performing PLS-DA. Three PCs were necessary to grouping the samples in their respective classes based on Raman spectroscopy and XRF data considered individually. The identification appeared more robust when fused data are used.

Similarly, Deneckere et al.⁴³ applied PCA on a fused dataset comprising Raman spectra acquired on 24 natural and synthetic inorganic pigments, and on a polychrome object, a porcelain card and the corresponding elemental

information (areas obtained from micro-XRF spectra). Raman spectra were acquired with both a 785- and a 532-nm laser, the former being generally able to provide good quality spectra for identification purposes. The green-excited Raman spectra were only kept where they provided additional information. After baseline correction and scaling to 300-s measuring time, 22–24 Raman bands were selected for PCA. The loading plot of the fused dataset showed that the Raman bands and elements of each pigment are not strictly correlated, because pigments often share the same key element, and also some Raman bands are common. Nevertheless, the processing of fused data showed a greater success in the grouping of pigments on a polychrome object. However, the use of the reference dataset for the purpose of identification of pigments on the porcelain cards posed some issues. In fact, the ubiquitous presence of a lead white layer affected all the measurements, hampering a correct identification even when processing the fused dataset. Signals related to Pb and Pb compounds in both datasets were removed, and the PCA of fused data allowed to increase the success rate of identification to 70%. The unsuccessful cases were, probably, due to trace impurities and to the simultaneous presence of signals from different layers in the XRF spectra.

Another possible combination of techniques for application of PCA on fused data regards Raman and FORS to non-invasively identify binding media in lead-based paints. This combination was applied to simulated mediaeval paints by Pallipurath and co-workers,⁴⁴ who mixed lead white, red lead and lead-tin yellow with gum arabic, egg (white, yolk, whole) and siccativ oils (linseed, poppy, walnut). The PCA was performed on interpolated, vector-normalized and smoothed (see reference for details on smoothing) FT-Raman spectra. Here, the sum of variance of the first two PCs was not considered indicative, with the distance of points within a cluster to the centroid being preferred. Pigment–binders interactions appeared in the range 1000–3200 cm⁻¹, and different spectral regions were subjected to PCAs separately (see the reference for details). The signals of pigments were not considered, as their bands would cause PCA to give more weight to such intense signals compared with the binders. The whole range and the C-H region proved better in discriminating lipidic and proteinaceous binders mixed with the same pigment, with different groupings observed for the same paints obtained with different pigment-binder ratios. Tight groups were nevertheless observed for the oils, and the other binders appeared more spread out and overlapping. The discrimination of gum arabic from egg white was confirmed to be challenging (see Section 3). The same binder was tested across the three pigments and at different pigment-binder ratios:

PCA showed separate groups for the pigments and their mixing ratio, independently on the spectral region. It was observed that vector normalization equalizes band intensities, while band broadening of the C-H due to bound pigments was retained. Fluorescence appeared as a separate contribution in the loading plot and, in this case, helped in differentiating lead tin yellow from the other paints. The built PCA space was then used to classify paint samples (lead-tin yellow in oil) with a 60% success rate. Nevertheless, issues with fluorescence enhancement in the pre-processing (vector normalization) and Raman spectra with worse signal to noise ratios were raised. The combination of separately vector normalized Raman and FORS spectra proved successful in improving the clustering, especially for egg white and gum arabic. The loading plots of the fused dataset showed that PC1 included signals from both Raman and FORS and PC2 mainly from Raman. A later study by the same authors⁴⁵ was aimed at assessing the relative proportion of pigment and binder in three paint systems (lead white + egg yolk, lead-tin yellow light or dark + poppy oil) from PCA on FT-Raman and FORS spectra. The workable mixture range constrained the obtainable paint mix for each pigment (full details in the reference). The same spectral processing described in Pallipurath et al.⁴⁴ was applied here, but standardization was not carried out at first, as band intensities were considered the key source of spectral variations. The spectra were instead normalized to the 2926 cm⁻¹ band (C-H stretching). The PCA on lead white paint spectra allowed to assess reasonable estimates of proportions just by considering the first PC (>90% of total variance explained), which proved sensitive to lead white phonon and carbonate vibrations. In a later step, standardized data were considered: by considering the first 2 PCs (>90% of total variance explained), it was still possible to estimate the paint composition. The two lead tin yellows were treated together to assess if PCA could work simultaneously for qualitative and quantitative purposes. In this case, PC1 separated the two pigments. PC2 (21.3% of the variance) follows the mixture composition for both pigments. Several unknown paint samples (described as rich and lean in binder) were investigated using the established PCA method, but for lead white, it appeared that its composition (anhydrous vs. basic) negatively affected the estimation of the binder content. On the other hand, lead-tin yellow paints showed a more predictable behaviour, except in the lean samples that could be either made of light or dark lead tin yellow. A separate study of the two pigments allowed to better assess the paint composition. The authors pointed out that PCA is not a quantitative approach in itself, but the potential for this type of investigations based on a set of reference materials surely enforces further applications, in parallel

to the assessment of paint ageing and paint-binder interactions, which is necessary in the field of cultural heritage analysis.

Another set of oil paints using poppy and linseed oil, with lead white, azurite, natural ultramarine, phthalocyanine blue, zinc oxide and synthetic ultramarine was characterized with Raman and FTIR and subsequently with PCA, by Carlesi et al.⁴⁶ Furthermore, the 48 prepared paint samples, and the two binders, were studied after 9 months of natural ageing (incomplete polymerization of the oil without the occurrence of hydrolysis or oxidation reactions); 320 Raman spectra (785 nm, spectral resolution of 4 cm^{-1}) were acquired in the range $200\text{--}2000\text{ cm}^{-1}$ and showed good reproducibility; they were then baseline corrected, average reduced by a factor of 2 and their first derivative calculated. Vector normalization was applied on both sets of vibrational spectra, which were fused. Then PCA was performed by applying a mean centring onto data column-wise and the NIPALS algorithm. PCA on first derivative micro-Raman spectra did not allow to separate the spectra of pigments from those of the paints, except for zinc white, whose spectra grouped with the drying oils. On the other hand, PCA on first derivative FTIR spectra allowed to identify the binders based on the first two PCs. PCA on the combined spectra of all the materials allowed to select the first 5 PCs explaining 81% of the total variance: from the score plots, it was possible to separate the oils, from the mixtures, from the pure pigments, to discriminate the type of pigment, including differences between natural and artificial ultramarine, the peculiar behaviour of zinc white paints. PCA was also performed on pigments, with a focus on lead white and azurite, as they affect the binders (whose modifications are more evident in the infrared spectra and are not discussed here). The same research group further studied the effect of lead white and zinc white on the ageing of linseed oil paints.⁴⁷ The mock-ups were followed up for 24 months, and also compared with 10-year-old samples of comparable composition. Micro-Raman spectroscopy with a 785-nm laser was combined with Fourier transform-near infrared spectroscopy (FT-NIR), in order to include a variety of vibrations and structural modifications that affect the binders during curing (that also produce polar functional groups that are better detected by infrared). Only the $1000\text{--}1800\text{ cm}^{-1}$ region of the Raman spectra was considered, the spectra baseline corrected, average reduced by a factor 2, smoothed, vector normalized and their first derivative calculated. The authors avoided the second derivative as it would have introduced artefacts and more complexity to the spectra. The combined Raman and FT-NIR spectra had therefore 1450 data points, 400 of which from the Raman spectra. PCA was applied on the spectra in

the same way as in Carlesi et al.⁴⁶ Each model was validated by Leave One Out (LOO) cross-validation. The spectra of the mock-ups were pre-processed and projected into the specific PCA models. PC1 indicated two different ageing trends for the two pigments. It appeared that, for both paints, a prolonged ageing does not induce additional reactions compared with what already described for the 24-month samples.

3 | BINDING MEDIA

When studying paintings, colouring and binding materials are often treated together, because not only they are bound to interact with each other, but also their analysis is influenced by their co-presence. Therefore, many works involving binders are reported in Section 2.

Nevin and co-authors⁴⁸ selected eight proteinaceous binders among those most used throughout history: ox glue, rabbit skin glue, parchment glue, fish glue, egg white, egg yolk, milk and extracted casein, characterized by the presence of different proteins. The Raman spectra (785 nm) were subjected to multivariate analysis in the range between $2700\text{ and }3200\text{ cm}^{-1}$ with the aim of deciding if the spectra acquired on the different binding media could be considered statistically different from each other. In order to overcome the problem of instrumental calibration, that could cause shifts in the bands position from day to day, the data were interpolated obtaining 251 points in the selected range with a spacing of 2 cm^{-1} ; furthermore, the spectra were vector normalized to avoid variations influencing the overall recorded intensity; finally, the average intensity was subtracted from every position in the spectrum and then divided by the standard deviation of the peak intensities. The more significative spectral region was found to be the $2800\text{--}3100\text{ cm}^{-1}$ one. The PC plots clearly show that the protein could easily be grouped: in detail, not all the samples could be distinguished within the collagen-based materials cluster (e.g., rabbit skin glue and ox bone glue) while within the dairy proteins cluster, all differences were significant. Besides, egg yolk and egg white could be differentiated and the former well separated from the rest of the materials too; milk as well was clearly distinguished from all the other samples, thanks to the co-presence of other proteins apart from casein and sugars within it. Furthermore, a sample of yellow ochre mixed with egg white was studied. The abovementioned spectral region, not affected by the pigment's signals, was extracted and interpolated, and scores calculated in order to compare it with the reference database of binders. The proteinaceous binder was identified correctly. In continuation with the previous work, the same group developed a study

concerning the artificial ageing of protein-based binders.⁴⁹ Actually, ageing is the main subject of several works reported in this section. In the latter study, the already mentioned eight binders⁴⁸ were analysed, but half of the samples were subjected to natural ageing, while the remaining half were exposed to artificial ageing corresponding to 100 years of indoor light exposure. In this case too, micro-Raman spectra (785 nm) were interpolated and normalized before performing multivariate statistical analysis. PCA was performed on the 2700–3200 cm^{-1} region, the CH stretching one, considered as the best choice for absence of possible inorganic pigments signals, lower luminescence and less bands of the material itself but still informative with respect to the fingerprint region. The PCA results showed a good differentiation of most of the binding materials, even though, as an example, casein and egg white separation was complicated by the highest similarity of their spectra. Furthermore, PCA was useful to provide some hints about spectral modifications due to ageing: studying the loadings diagram, it could be deduced, for example, that the decrease of the band at 3050 cm^{-1} (aromatic C-H stretching) and the increase in the intensity of lower wavenumber bands (oxidation of aromatic rings) were the cause of the shift towards less negative PC1 for egg yolk and egg white. Another work⁵⁰ analysed through micro-Raman spectroscopy (785 nm) different mixtures of quartz and albumin in variable percentages, as such and after thermal treatments at different temperatures. The PCs were obtained using both the covariance data matrices (scaling by mean-centred data) and the correlation data matrices (scaling by unit variance), with better results for the latter, which were the actually used data. The PCA results were less and less significant as the percentage of quartz increased, because quartz peaks prevailed, masking the signals of albumin, the most affected by the heating. The study by Manzano and co-authors²⁷ focussed instead on the ageing of three drying oils, namely, linseed oil, poppy oil and walnut oil, and of egg yolk, with which 50:50% (v/v %) mixtures were prepared and left to naturally age for 6 years. The work was carried out by means of FT-Raman (1064 nm). After baseline correction, normalization and mean centring, the spectra were subjected to PCA, mainly with the aim of explaining the ageing observations; PLS-DA was also performed. As a result of the combined approach, at the same time, lipidic binding media could be differentiated according to their composition, and information about ageing was obtained, too. In detail, 2893 cm^{-1} (CH stretching) was identified as the most significant band for discrimination, while the peak at 2939 cm^{-1} (CH_2 asymmetrical stretching) showed its importance in linseed oil alteration. Fatty acids were studied by means of micro-Raman

by Otero et al.²⁸ and were employed to synthesize various metal carboxylates, then subjected to a comprehensive Raman + PCA study including also two case studies (see Section 2.1).

As already seen in previous work,⁴⁹ Daher and co-workers⁵¹ highlighted how the morphology of the C-H stretching region of the Raman spectra could be extremely informative for terpenoids, proteins, polysaccharides and triglycerides, especially if coupled with PCA. This work was mainly focussed on terpenoids, and its approach is described more in detail in Section 5.1. It can be here summarized that the PCA carried out on the pre-treated Raman spectra resulted in a weak clustering of the scores, due to the spectral noise in regions without characteristic and significative Raman peaks. On the other hand, the approach exploiting chosen variables of the spectra leads to an easy recognition of the scores of reference materials: for example, oils were found grouped in the upper right quadrant of the PC1 versus PC2 score plot.

4 | GLASSY MATERIALS

Applications of multivariate statistical analysis to Raman spectra of glassy materials are surprisingly few. A more frequent exploitation of PCA for the interpretation of Raman data would be expected for glassy phases, due to the characteristic features of their Raman spectra. In fact, where a clear distinction of sharp peaks as in crystalline materials is not possible, the differentiation of spectra might be more challenging, and the aid of PCA could be requested. The case of combined XRF or SEM-EDS and Raman studies of pottery glazes is more frequent, where PCA is applied to the elemental data and not to the molecular ones, see for example.⁵²

4.1 | Obsidian

Raman spectroscopy has been widely employed in the literature for the analysis of obsidian focussing on the detection of their spectra, the structural changes due to compression, the influence of water content and provenance discrimination of western Mediterranean provenances, but the first to apply PCA to a dataset of obsidian Raman spectra was a group of Australian researchers. They aimed at the distinction of the Pacific Region sources on an unbiased basis going beyond the mere visual examination and at the tentative provenance attribution of some archaeological artefacts.^{53–55} In the first work,⁵⁵ a laboratory instrumentation was used, equipped with a 785-nm excitation wavelength to analyse

43 obsidian samples. PCA carried out on the averaged (3 per sample), smoothed and normalized Raman spectra, distinguished three main clusters corresponding to the three major geographical regions (West New Britain, Manus and Vanuatu). Furthermore, two additional spectra acquired on an obsidian stemmed tool from Pitt Rivers Museum (University of Oxford, UK) were included in the dataset, and thanks to PCA, the provenance of this artefact was tentatively ascribed to the New Britain group. A step forward consisted in assessing the possibility of employing a portable Raman spectrometer coupled with PCA to identify the geological source of obsidian, comparing the obtained results with those of a laboratory instrumentation,⁵⁴ where both are equipped with a 785-nm laser. First derivative was performed prior to PCA to reduce or eliminate baseline and background effects at the same time emphasizing bandwidth. The number of samples was increased to 65 (West New Britain, Manus and a small sample set from the Banks Islands, Vanuatu) with respect to the previous work. The spectral resolution for the two instruments was different (6 vs. 1–2 cm^{-1}); therefore, ascertaining that similar discriminations could be obtained in the two cases was vital. Actually, no influence of this parameter was observed. Neither the broad background nor the focussing issues caused significant differences in the data collected from most of the samples with the portable instrument. MD calculations were employed too, to compare the clusters obtained with PCA on the two datasets: a similar discrimination could be obtained, with only minor differences. On the other hand, if the obtained clusters are comparable, the loadings revealed that the contribution of the variables to the formation of the clusters themselves was different in the two cases, due to the varying intensities of specific bands (e.g., those of magnetite) if acquired with portable or laboratory instrumentations.⁵⁴ To correctly interpret the loadings, it was necessary to compare the pre-processed data after first derivative with the corresponding original spectra. A further work adopted the same approach⁵³ aiming at classifying four archaeological obsidian artefacts kept at Museum of Victoria, which was achieved by adding their Raman spectra, acquired once again with portable instrumentation, to the previous dataset of reference spectra of different provenance.

A different approach was instead adopted in a more recent work,⁵⁶ where PCA was carried out on the raw, unprocessed Raman spectra of obsidian from British Columbia. The aim was that of investigating its structure and composition, including trace elements and Pb isotopes, starting with a non-destructive survey involving portable XRF and laboratory Raman spectroscopy (532 nm) and proceeding with a minimally invasive phase with (laser ablation)–inductively coupled plasma–mass spectrometry ([LA]-ICP-MS). Both geological source

materials and 14 archaeological artefacts were analysed. The multi-variable methodology proved useful for increasing the confidence in the obsidian sourcing.

4.2 | Glass and glazes

The first PCA approach to Raman spectra of synthetic glass was performed by Colomban and co-authors in 2006,⁵⁷ when a differentiation guide of glass types based on Raman spectra was proposed. A selection of 30 Raman spectra of previously analysed glass, porcelain, stoneware, *terracotta* and faience glaze was used to establish relationships between the Raman parameters and the raw materials/technological procedures employed, resulting in a given chemical composition. The chemical composition of these samples was also determined, or it was already known from previous works. The parameters that can be extracted from a Raman spectrum of an amorphous phase are various; therefore, PCA, together with HCA, was used not really for the discrimination of groups of samples in this case but rather to evaluate which parameters can be considered most significant for establishing a reproducible and simple procedure of differentiation: for this reason, the authors chose to extract specific variables from the Raman spectra rather than processing them as such. The 10 considered variables were polymerization index (I_p), Si–O stretching peak maximum, wavenumbers and peak areas of Q_0 , Q_1 , Q_2 and Q_3 components. The loadings indicated the more independent parameters, as, for example, I_p , Si–O stretching peak maximum wavenumber and some areas of the Q_n components. A similar aim, testing the possibility of distinguishing different tiles manufactures, lead other authors⁵⁸ to the same type of choice: considering Raman spectral parameters rather than spectra themselves, as in Colomban et al.,⁵⁷ but in combination with elemental data extracted from XRF analyses. The Raman instrument used was fixed and equipped with a green laser (532 nm). The data were vector-normalized prior to the PCA, which was carried out on 20 variables, including the Raman ones (I_p , wavenumber[Q_0] to wavenumber[Q_4] and area[Q_1] to area[Q_4]) and the elemental ones (normalized weight percentages [wt%] of Al_2O_3 , K_2O , CaO , TiO_2 , MnO , Fe_2O_3 , CuO , ZnO , PbO and SrO). In this work, too, the most discriminative parameters were identified, and they resulted to be MnO and Fe_2O_3 as markers for the brown glazes, whereas I_p , area(Q_1), K_2O , CaO , PbO and Al_2O_3 for the white glazes.

A different type of application concerned glassy ceramics coatings.⁵⁹ It consisted in an experimental work where tin-opacified lead glazes were prepared following Renaissance recipes with firing at 300°C, 450°C, 600°C, 750°C, 850°C, 920°C and 990°C; 31 micro-Raman spectra

(532 nm) were acquired on the products fired at the different temperatures, and PCA was carried out directly on the spectra in the 700–1200 cm^{-1} region, with the aim of discriminating the spectral features of the different glassy phases and of creating a pattern for real cases examples. Spectra were pre-processed by using auto-scaling to remove systematic variation due to baseline effect. The experimental glass spectra formed three clusters, and the loadings were useful to determine which spectral region was most significative to identify each given firing temperature. Real case Deruta majolica glazes were analogously analysed and treated: they fell in the field of 920–990°C firing group.

A very recent study by Colombari et al.³⁹ aimed at discriminating yellow glazes and enamels of both European and Chinese provenance using a high number of previously acquired Raman spectra. The paper was mainly aimed at the pigmenting phases study, but the contribution of the glassy matrix resulted important in the approach involving the statistical treatment of the spectra themselves (comparison is reported in paragraphs 2.1 and 2.2). A first test concerned the 100–1200 cm^{-1} range of spectra acquired with a portable Raman system with a 532-nm laser. This allowed to identify glazes containing cassiterite SnO_2 and As-bearing phases, at 630 and 775–820 cm^{-1} , respectively. However, this was apparent from a visual observation of the spectra, and a narrower range was selected to avoid such strong features (100–700 and 100–590 cm^{-1}). In the latter (which excludes cassiterite's main band), it appears that PC2 and PC3 contain the most relevant information and allow to discriminate glazes according to their geographical origin, which corresponds, from the spectral point of view, to pyrochlore phases (substitutions in Naples yellow) in the European manufactured objects. Actually, the information about the lead-based pigmenting phases was not easily separated from that related to the glassy matrix, the latter being lead-rich in Europe and lead-alkali in China. The authors underline how the variability of spectra acquired in situ, the simultaneous presence (with different intensities) of both the pigment and the glassy matrix signals and the complex solid solutions across the studied materials hamper a successful application of PCA, while a visual discrimination of the materials appears more reliable.

5 | GEMMOLOGICAL MATERIALS

5.1 | Biogenic

Among the gemmological materials of biologic origin, amber is that more frequently studied by means of Raman spectroscopy coupled with PCA, even though the group of published papers on the subject is relatively small.

Analogously to obsidian, the general aim is that of distinguishing ambers of various provenances, differentiating the most studied type of amber, succinite, of Baltic origin, from others such as romanite, from Romania,^{60,61} simetite, from Sicily⁶² and valchovite, from Czech Republic.⁶³ Within this scheme, attention is often paid to extract those spectral parameters that are more discriminative. FT-Raman is in most cases considered more suitable for amber characterization, and groups of geological samples from the chosen geographical regions are compared, sometimes adding archaeological items with the aim of attempting a provenance attribution of the latter.¹⁴

A Romanian research group carried out Raman investigations, in parallel to infrared ones, aimed at fully characterizing romanite, whose distinction from Baltic amber was made possible through PCA on the baseline-subtracted and normalized FT-Raman spectra in the 250–1800 cm^{-1} region. Functional groups specific of the two types of amber were also identified thanks to the attribution of the single bands and the study of the loadings. Forty-three samples collected from fragmentary beads found at Cioclovina and kept at the National History Museum in Bucharest were studied too, and their assignment to a specific geological origin was more accurately performed by selecting PC1 and PC2 scores as variables in DA. The great part of these amber beads was identified as Romanite though with a high internal variability of each sample.⁶⁰ A further step focussed on the artificial weathering of Baltic and Romanian geological amber, where PCA on FT-Raman spectra was used to understand which environments (air, salt, acid, alkaline) initiated or catalysed their thermal alteration starting from an unaltered condition. In order to avoid the influence of low signal-to-noise ratio on the PCA results, the authors chose to pre-treat the spectra selecting the region with more intense signals (3100–3500 cm^{-1}); then they subtracted the baseline and normalized the spectra. Furthermore, thanks to the experience gained in the previous work,⁶⁰ they managed to reduce the high internal variability of amber, removing one of the three replicate spectra on the same sample, the eccentric one. In this way, it was possible to notice that Baltic amber is less affected by alteration—with the exception of that in saline environment—with respect to romanite, for which the most altered seemed the sample in alkaline environment. In order to obtain significant results on both the type of weathering and the provenance distinction, the spectrum had to be cut in the 1570–1700 cm^{-1} region; the results were in accordance with those obtained with infrared spectroscopy.⁶¹ The study concerning simetite, from Sicily,⁶² involved also Baltic and Dominican samples and was carried out with a micro-Raman instrument equipped with a 785-nm laser. With the support of solid-state nuclear magnetic resonance (NMR) and of PCA, it was successful in

distinguishing the three geographical origins. The approach employed in the work by Peris-Díaz and co-authors⁶³ was totally different because it coupled Raman spectroscopy to PLS and employed PCA only as a preliminary step of distinction between FT-Raman spectra of amber samples from the Baltic and Moravian regions (Czech Republic, valchovite) and objects dating back to the Upper Cretaceous and Cenozoic ages.

Within a broader work aimed at developing a methodology to discriminate natural organic compounds and including different families of organic materials (terpenoids, proteins, polysaccharides and triglycerides [see Section 3]), Daher and co-workers⁵¹ studied 26 reference samples of terpenoid resins: diterpenic and triterpenic resins and shellacs. Furthermore, 10 samples of archaeological copal from Yemen were added to the corpus to target different aims: evaluating the representativity of the references, the potentiality of the method for ancient samples and the effect of surface alteration on the materials distinction. The samples were analysed both by FT-Raman and by attenuated total reflectance–Fourier transform infrared (ATR-FTIR) spectroscopy. The pre-treatment regarded the choice of the spectral region (2730–3185 cm⁻¹), including the most intense signal at ca. 2900 cm⁻¹, a linear baseline subtraction, spectra normalization and their decomposition; the latter was carried out on the basis of the second derivative of the spectra. The spectra normalization was an important step in the pre-processing to avoid the effect of scaling on the PCA results because the analysed materials were of different size or thickness, and the resulting spectra were characterized by variable intensities. PCA was performed both on the treated spectra before decomposition and on the fitting parameters (band positions, FWHM, area and profile) of the 10 single bands. Better results were obtained with the second type of approach. The archaeological copals were grouped with the African/Madagascar reference ones; furthermore, their conservation state did not seem to affect their identification and distinction. Here, as in Badea et al.,⁶¹ PCA was also used to study the material alteration. In detail, the internal part of a copal sample was used as reference of unaltered material, while the FWHM of two specific bands (associated to C-H₂ groups and to C-H vinyl groups) were correlated variables affecting PC1, which separated less altered samples (yellow copals) from more altered ones (red copals).

Another gemmological material that has been subjected to Raman analyses with the subsequent assistance of PCA is ivory. FT-Raman has been used to discriminate among hard ivories, soft ivories and mammoth tusk, processing with PCA the spectra in the 400–1800 cm⁻¹ region.⁶⁴ Study of the loadings was useful to understand that the differentiation was based on the

ratio between collagen and hydroxyapatite. PLSR was also associated to make a calibration model for specific gravity prediction of the ivories. Another work,⁶⁵ aimed at discriminating dentine from six mammalian species, used a different approach for PCA: the collected spectra of dentine, cement and bone were divided into 11 regions; each of them was baseline-subtracted, and the area value of one of the regions was fixed to be 1; the areas of the other 10 regions were normalized to the fixed one, and these values were processed by PCA and DA. A successful differentiation was achieved, notwithstanding some overlaps that did not allow 100% classification of each species. In the framework of a study concerning painting materials ageing⁵⁰ (see Section 3), PCA was applied to Raman spectra (785 nm) of natural and thermally treated (150°C, 200°C, and 250°C) powdered bone in the 650–1730 cm⁻¹ region. The scores diagram showed that untreated bone spectra separate from the differently treated ones; furthermore, the approach was useful to confirm that the principal hydroxyapatite signal at about 958 cm⁻¹ is not affected by heating at the selected temperature range.

Finally, an application to pearls study was published.⁶⁶ PCA was employed with a different kind of objective: not really for the discrimination of fresh water, Akoya and South seawater pearls, for which PLS was used, but rather to validate the reproducibility of a method. The Wide Area Illumination (WAI) scheme was in fact used: a 785-nm laser was magnified up to a 6-mm illumination area, thus covering a large sample area; the scattered radiation had then to be collected by an array of 50 optical fibres. Once PLS had proved successful in pearls discrimination, 20 spectra were collected on one pearl only, each time repositioning the sample in a random orientation. PCA was performed on all the 100–1800 cm⁻¹ spectra of the 14 analysed pearls with the addition of the 20 spectra collected on the single pearl. The comparison of the scores distribution for the different pearls and for the repeated spectra of the same pearl showed that these latter are positioned very close to each other's, pointing to an acceptable reproducibility of the WAI method. A more classic type of application concerned the pigment in the shell of *Pinctada margaritifera* oysters used for producing gem-quality pearls, for which PCA on Raman spectra allowed to separate the albino, yellow and red varieties from the green and black ones.⁶⁷

5.2 | Mineral

A combined Raman and PCA approach to minerals of gemmological interest is extremely rare, and examples of published studies are sparse and on disjointed subjects.

Bi and co-authors⁶⁸ took into account six minerals (gypsum, spodumene, barite, haematite, moonstone and labradorite) and analysed them by means of Raman spectroscopy (532 nm) and LIBS. They then fused the obtained results into a new single matrix, after normalization to the maximum intensity, and performed a series of multivariate analyses on it: PCA, PLS-DA, artificial neural network (ANN) and SVM. The total number of variables in the matrix was too high; therefore, PCA was used to reduce them. The first 7 PCs explained 97.05% of the total variance for the fused data; therefore, they were subjected to PLS-DA, ANN and SVM. PCA results for the first 3 PCs resulted in a misclassification of gypsum and barite, if only Raman data were considered (in our opinion due to the spectral resolution of 6 cm^{-1}), while the fusion with LIBS data gave the best separation for all the minerals. Probably, analogous problems of spectral resolution (4 cm^{-1}) were encountered by Chang et al.⁶⁹ when comparing Raman spectra (785 nm) acquired with the WAI scheme on 20 natural and 21 Be-diffused sapphires. The positions of the three major bands found (577 , 645 and 750 cm^{-1}) were shifted of 1 to 3 cm^{-1} —therefore less than the spectral resolution—towards lower wavenumber in Be-diffused gemstones with respect to the natural ones due to the altered lattice structure. Authors performed PCA on the Raman spectra in the $460\text{--}780\text{ cm}^{-1}$ range: the scores plot of PC1 and PC2 fairly discerned the two groups, but some scores resulted ambiguous. Another Raman system (785 nm) with a low spectral resolution (7 cm^{-1}) was used to distinguish Shoushan (SS) (China), from Changhua (CH) (China) and Laos (LA) Tianhuang stones, considered precious stones with great cultural and economic value.⁷⁰ After Raman spectra pre-processing, including background removal and normalization, PCA was applied to the latent variables of seven bands (268 , 336 , 434 , 460 , 748 , 795 , and 915 cm^{-1}). PCA was associated to latent Dirichlet allocation, allowing to establish the classification accuracy, which was higher for SS/CH distinction (94.8%) and lower for SS/LA (75%) and CH/LA (67.1%) ones. Random forest (RF) analysis was associated too. No hints to possible problems deriving from the spectral resolution were given by the authors.

Another work⁷¹ aimed at the characterization and distinction of two types of phosphate (turquoise - planerite) from Carico Lake Valley in Nevada, slightly differing in colour, exploiting Raman spectroscopy with the support of PCA to investigate these inhomogeneous and nonstoichiometric species. Twenty-seven blue and 30 green crystals were analysed, and the Raman spectra acquired with the 442-nm laser were processed with PCA. Besides the distinction of the samples, the method allowed, through the observation of the loadings,

to identify the bands characteristic for green and blue phosphate.

The only study coupling Raman to PCA for the investigation of minerals of gemmological interest directly constituting cultural heritage items seems at present to be that by Caggiani and co-authors.⁷² Here, PCA was applied to the Raman spectra of a group of chalcedony glyptics. The aim was double: (1) testing this approach to distinguish chalcedony spectra with different intensities of the ca. 503 cm^{-1} band, avoiding a more complex and long band decomposition and (2) identifying groups of samples with possible different chalcedony composition. Both aims were achieved but the authors had to be careful with the pre-processing of the spectra, that, acquired with a portable instrumentation, were affected by a high background and an invasive noise due to the multi-layer filter. Spectra were considered only in the region of interest for the two main peaks ($435\text{--}525\text{ cm}^{-1}$), smoothed (Savitsky–Golay function), baseline-subtracted (multi-point baseline) and normalized to the maximum. Furthermore, it was observed that a better correlation between the results of PCA and those of the band decomposition could be achieved after performing first or second derivative of the spectra: this procedure allowed to abate the contribution of the filter noise and guaranteed a higher reliability of the PCA results.

6 | MISCELLANEOUS

As concerns one of the most studied materials in the field of archaeometry, that is, ceramic, a detailed Raman study on Al-bearing haematites⁷³ as found in Sigillata wares from Gaul also included PCA on the micro-Raman spectra subjected to baseline subtraction, though no pre-treatment preparatory to PCA was highlighted by the authors. This study allowed to highlight the dependence of band positions on the Al-for-Fe substitution, especially of the E_g Raman band at ca. 300 cm^{-1} . Both band intensity ratios and PCA on synthetic Al-doped haematites allowed to monitor chemical variations as they affected the Raman spectrum. Otherwise, for pottery ceramic bodies, the most frequent application of PCA to molecular data involves infrared spectra (e.g., Medeghini et al.⁷⁴), which can be explained by the difficulty of obtaining significant Raman results on ceramic phases.

In the field of fibres study, Bianchi and co-authors⁷⁵ used a micro-Raman with both 532- and 780-nm lasers to analyse industrial cotton, polyester and polyamide textiles, subsequently stained with different dyes, before and after artificial ageing through washing and simulated sun exposition. Raman spectra were pre-processed (smoothed, deconvoluted, mean centred) prior to PCA and LDA

analysis aimed at spotting trends and clusters within the samples. Though the work concerned modern textiles, the results could be useful also in view of cultural heritage analysis, including contemporary art. Quintero Balbas et al.,⁷⁶ instead, studied wool and silk fibres dyed with turmeric and saffron both with micro-Raman (785 nm) and FT-Raman (1064 nm). The authors aimed at avoiding SERS as the samples can be affected by the pre-treatment, and by-products of nanoparticle synthesis appear in the Raman spectra. A real sample from a 19th century Persian carpet was compared with mock ups subjected to natural (36 years) and artificial ageing. The range 751–1800 cm^{-1} of micro-Raman spectra was considered and pre-treated as follows: second derivative baseline subtraction, standard normal variate (SNV) to correct both baseline and intensity variations and column centring. PCA allowed to group samples according to the combination of fibre and dye and to highlight differences in the spectra of naturally and artificially aged samples, especially visible in saffron-dyed silk. PC1 is more sensitive to the dye signals, whereas PC2 to the amide I band, to sericin protein and to amino acids (alanine, tyrosine). Finally, the historical fibre was plotted in the PCA model and was grouped together with turmeric-dyed wool, proving the effectiveness of this fully non-destructive approach in the simultaneous characterization of fibre and dye.

A peculiar case study⁷⁷ concerned the characterization of gum applied on the back side of ancient stamps; 108 samples were studied by means of FT-NIR reflectance and FT-Raman spectroscopy. The authors have addressed the problem of minimizing baseline shifts and overall intensity variations, performing two types of spectral pre-processing: the SNV transform and the first derivative, both alone and in sequence, obtaining the most consistent results with the latter solution. The two datasets were subjected to PCA both separately and in data fusion mode. The latter included low-level and mid-level approaches, explained in detail in the paper. The PCA distinguished stamps gummed with animal glue, gum arabic and polyvinyl acetate. The most characteristic FT-Raman spectra bands of the latter were identified at 631, 1085, 1437, 1737 and 2937 cm^{-1} ; besides, the PC1 versus PC2 scores plot showed that the multivariate analyses are able to discriminate the stamps purposely regummed with polyvinyl acetate from the original ones, as the areas around 600, 1450, 1750 and 2900 cm^{-1} of the FT-Raman spectra showed some differences in the two cases.

7 | CONCLUSIVE REMARKS

One of the first impressions deriving from this overview on PCA application to Raman investigations of cultural

heritage concerns the unbalance of published contributions about different materials. On one hand, Raman spectroscopy dedication to pigments and paintings materials study is well known,¹⁰ and as a consequence, a proportional larger number of studies involving PCA could be expected. On the other hand, the scarce exploitation of this combined approach for the study of broad Raman bands of some amorphous materials (e.g., glass) or for distinction of mineral series/solid solutions is surprising.

The aims for which PCA has been coupled to Raman spectroscopy in the reviewed fields are manifold.

1. Computer-assisted and semi-automatic identification of materials, alone or in mixtures/paint layers for colouring materials, especially when the spectrum is complex (e.g., organic materials, SOPs and carbon black pigments), supporting both specialists and less-experienced Raman users in the task;
2. Distinction of groups of reference materials, with or without the addition of one or more specific samples to be attributed to a given group, as often happens for provenance issues of archaeological findings, such as obsidian or amber;
3. Highlighting the band position dependence on the mineral structure;
4. Data reduction in view of further multivariate analyses;
5. Overcoming of the contribution of fluorescence, which is a well-known issue in Raman spectroscopy;
6. Loadings study for the identification of the most significant Raman parameters for the discrimination of given compositions/provenance, as well as for evaluating of degradation state and/or of its causes;
7. Comparison of datasets acquired by different scientists with different devices and measurement conditions;
8. Validation of the reproducibility of a method.

It however appears that the instrumental spectral resolution, which can be insufficient to achieve a significative distinction, could make the discrimination of samples into different groups less relevant. From the methodological point of view, the numerous research groups applying PCA to Raman studies do so with varying degrees of confidence and within different research questions: in some cases, PCA is a solid foundation of the whole work; in others, it is a complementary tool to target specific research aspects. Moreover, a large variety of pre-processing methods is available to PCA users, who have to make a selection. Sometimes, this is done according to the spectral features to be considered or avoided (bands intensity ratio, spectral noise and background reduction, fluorescence abatement, etc.), and it is specified by the authors with an explicit aim within the experimental design.

The overview of the chronological spread of PCA in archaeometry allowed to highlight a delay in its application to Raman spectroscopy (as for example compared to quantitative results) likely a consequence of limited processing power of computers for both pre-processing and processing of high numbers of variables. Moreover, it was possible to notice, after a peak in publications about a decade ago, a slight decrease in interest of this approach, resulting in a few papers per year. Therefore, this overview can be considered a useful milestone to assess past successes and pitfalls and to promote a renewed interest in this exploratory method as a standalone or coupled to other chemometric techniques.

In conclusion, it appears that the coupling of Raman spectroscopy with PCA is a versatile approach for numerous tasks in the field of archaeometry and conservation science, from exploration of an unknown dataset to understanding very specific structural modifications.

CONFLICT OF INTEREST STATEMENT

The authors declare no known conflicts of interest.

ORCID

Maria Cristina Caggiani  <https://orcid.org/0000-0001-8475-1175>

REFERENCES

- [1] T. J. Reedy, C. L. Reedy, *Archaeometry* **1994**, 36, 1.
- [2] G. Musumarra, M. Fichera, *Chemom. Intell. Lab. Syst.* **1998**, 44, 363.
- [3] M. J. Baxter, *Archaeometry* **2006**, 48, 671.
- [4] M. J. Baxter, I. C. Freestone, *Archaeometry* **2006**, 48, 511.
- [5] J. M. Madariaga, *Anal. Methods* **2015**, 7, 4848.
- [6] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, *EPJ Tech. Instrum.* **2015**, 2, 1.
- [7] R. Wehrens, *Chemometrics with R - multivariate data analyses in the natural and life sciences*, Springer, Berlin, Heidelberg **2010**, 45.
- [8] F. Casadio, C. Daher, L. Bellot-Gurlet, *Top. Curr. Chem.* **2016**, 374, 62.
- [9] R. G. Brereton, *Applied Chemometrics for scientists*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore **2007**.
- [10] H. G. M. Edwards, P. Vandenabeele, P. Colomban, *Raman spectroscopy in cultural heritage preservation*, Springer, Cham, Switzerland **2023** 53.
- [11] P. Vandenabeele, A. Hardy, H. G. M. Edwards, L. Moens, *Appl. Spectrosc.* **2001**, 55, 525.
- [12] S. A. Centeno, *J. Raman Spectrosc.* **2016**, 47, 9.
- [13] G. Visco, P. Avino, *Environ. Sci. Pollut. Res.* **2017**, 24, 13863.
- [14] D. Chiriu, F. A. Pisu, P. C. Ricci, C. M. Carbonaro, *Materials (Basel)*. **2020**, 13, 13.
- [15] P. Vandenabeele, L. Moens, *Analyst* **2003**, 128, 187.
- [16] A. Coccato, J. Jehlicka, L. Moens, P. Vandenabeele, *J. Raman Spectrosc.* **2015**, 46, 1003.
- [17] N. S. Daly, M. Sullivan, L. Lee, K. Trentelman, *J. Raman Spectrosc.* **2018**, 49, 1497.
- [18] N. S. Daly, M. Sullivan, L. Lee, J. K. Delaney, K. Trentelman, *Herit. Sci.* **2019**, 7, 1.
- [19] G. Piantanida, E. Menart, M. Bicchieri, M. Strlič, *J. Raman Spectrosc.* **2013**, 44, 1299.
- [20] B. Capone, P. Biocca, P. Corsi, C. Meneghini, M. Bicchieri, *J. Cult. Herit.* **2021**, 48, 1.
- [21] M. Castanys, M. J. Soneira, R. Perez-Pueyo, *Laser Chem.* **2006**, 2006, 1.
- [22] M. Castanys, R. Perez-Pueyo, M. J. Soneira, E. Golobardes, A. Fornells, *J. Raman Spectrosc.* **2011**, 42, 1553.
- [23] J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, *J. Raman Spectrosc.* **2016**, 47, 1408.
- [24] J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Appl. Spectrosc.* **2015**, 69, 314.
- [25] N. Navas, J. Romero-Pastor, E. Manzano, C. Cardell, *J. Raman Spectrosc.* **2010**, 41, 1196.
- [26] J. Romero-Pastor, C. Cardell, E. Manzano, Á. Yebra-Rodríguez, N. Navas, *J. Raman Spectrosc.* **2011**, 42, 2137.
- [27] E. Manzano, J. García-Atero, A. Dominguez-Vidal, M. J. Ayora-Cañada, L. F. Capitán-Vallvey, N. Navas, *J. Raman Spectrosc.* **2012**, 43, 781.
- [28] V. Otero, D. Sanches, C. Montagner, M. Vilarigues, L. Carlyle, A. Lopes, M. J. Melo, J. A. Lopes, M. J. Melo, *J. Raman Spectrosc.* **2014**, 45, 1197.
- [29] C. Lofrumento, M. Ricci, L. Bachechi, D. De Feo, E. M. Castellucci, *J. Raman Spectrosc.* **2012**, 43, 809.
- [30] N. D'Elboux Bernardino, T. Sevilhano Puglieri, D. L. A. De Faria, *Vib. Spectrosc.* **2014**, 70, 70.
- [31] G. Festa, C. Scatigno, F. Armetta, M. L. Saladino, V. Ciaramitaro, V. M. Nardo, R. C. Ponterio, *Molecules* **2022**, 27, 1.
- [32] A. Martins, A. Catherine, M. Duranton, A. Haddad, C. Daher, A. G. Bail, *T. Tang* **2021**, 4, 4205.
- [33] F. Pozzi, S. Porcinai, J. R. Lombardi, M. Leona, *Anal. Methods* **2013**, 5, 4205.
- [34] B. Doherty, M. Vagnini, K. Dufourmantelle, A. Sgamellotti, B. Brunetti, C. Miliani, *Spectrochim. Acta - Part a Mol. Biomol. Spectrosc.* **2014**, 121, 292.
- [35] E. Marengo, E. Robotti, M. C. Liparota, M. C. Gennaro, *Talanta* **2004**, 63, 987.
- [36] E. Marengo, M. C. Liparota, E. Robotti, M. Bobba, M. C. Gennaro, *Anal. Bioanal. Chem.* **2005**, 381, 884.
- [37] H. Mitsutake, R. J. Poppi, M. Breitreitz, *J. Braz. Chem. Soc.* **2019**, 30, 2243.
- [38] P. Vandenabeele, A. Rousaki, *Anal. Chem.* **2021**, 93, 15390.
- [39] P. Colomban, J. Burlot, D. Vangu, L. Bellot-Gurlet, *J. Raman Spectrosc.* **2023**. <https://doi.org/10.1002/jrs.6600>
- [40] E. P. Tomasini, E. B. Halac, M. Reinoso, E. J. Di Liscia, M. S. Maier, *J. Raman Spectrosc.* **2012**, 43, 1671.
- [41] M. Botticelli, A. Maras, A. Candeias, *J. Raman Spectrosc.* **2020**, 51, 1470.
- [42] P. M. Ramos, I. Ruisánchez, K. S. Andrikopoulos, *Talanta* **2008**, 75, 926.
- [43] A. Deneckere, L. de Vries, B. Vekemans, L. Van De Voorde, F. Ariese, L. Vincze, L. Moens, P. Vandenabeele, *Appl. Spectrosc.* **2011**, 65, 1281.
- [44] A. Pallipurath, J. Skelton, P. Ricciardi, S. Bucklow, S. Elliott, *J. Raman Spectrosc.* **2013**, 44, 866.

- [45] A. Pallipurath, R. V. Vofély, J. Skelton, P. Ricciardi, S. Bucklow, S. Elliott, *J. Raman Spectrosc.* **2014**, 45, 1272.
- [46] S. Carlesi, M. Ricci, C. Cucci, C. Lofrumento, M. Picollo, M. Becucci, *Microchem. J.* **2016**, 124, 703.
- [47] S. Carlesi, M. Picollo, M. Ricci, M. Becucci, *Vib. Spectrosc.* **2018**, 99, 86.
- [48] A. Nevin, I. Osticioli, D. Anglos, A. Burnstock, S. Cather, E. Castellucci, *Anal. Chem.* **2007**, 79, 6143.
- [49] A. Nevin, I. Osticioli, D. Anglos, A. Burnstock, S. Cather, E. Castellucci, *J. Raman Spectrosc.* **2008**, 39, 993.
- [50] J. Romero-Pastor, C. Cardell, Á. Yebra-Rodríguez, A. B. Rodríguez-Navarro, *J. Cult. Herit.* **2013**, 14, 509.
- [51] C. Daher, L. Bellot-Gurlet, A. S. Le Hô, C. Paris, M. Regert, *Talanta* **2013**, 115, 540.
- [52] A. Alonso-Olazabal, L. A. Ortega, M. C. Zuluaga, C. Alonso-Fernández, J. Jimenez-Echevarría, A. Sarmiento, *J. Raman Spectrosc.* **2022**, 53, 1204.
- [53] E. A. Carter, S. J. Kelloway, N. Kononenko, R. Torrence, in *Analytical Archaeometry*, (Eds: H. G. M. Edwards, P. Vandenabeele), The Royal Society of Chemistry, Cambridge, UK **2012**.
- [54] S. J. Kelloway, N. Kononenko, R. Torrence, E. A. Carter, *Vib. Spectrosc.* **2010**, 53, 88.
- [55] E. A. Carter, M. D. Hargreaves, N. Kononenko, I. Graham, H. G. M. Edwards, B. Swarbrick, R. Torrence, *Vib. Spectrosc.* **2009**, 50, 116.
- [56] R. McMillan, M. Amini, D. Weis, *J. Archaeol. Sci. Reports* **2019**, 28, 102040.
- [57] P. Colomban, A. Tournie, L. Bellot-Gurlet, *J. Raman Spectrosc.* **2006**, 37, 841.
- [58] J. Van Pevenage, M. Baeck, E. Verhaeven, L. Vincze, L. Moens, P. Vandenabeele, *J. Cult. Herit.* **2020**, 41, 27.
- [59] C. Ricci, C. Miliani, F. Rosi, B. G. Brunetti, A. Sgamellotti, *J. Non-Cryst. Solids* **2007**, 353, 1054.
- [60] E. S. Teodor, E. D. Teodor, M. Virgolici, M. M. Manea, G. Truică, S. C. Lijescu, *J. Archaeol. Sci.* **2010**, 37, 2386.
- [61] G. I. Badea, M. C. Caggiani, P. Colomban, A. Mangone, E. D. Teodor, E. S. Teodor, G. L. Radu, *Appl. Spectrosc.* **2015**, 69, 1457.
- [62] G. Barone, D. Capitani, P. Mazzoleni, N. Proietti, S. Raneri, U. Longobardo, V. Di Tullio, *Appl. Spectrosc.* **2016**, 70, 1346.
- [63] M. D. Peris-Díaz, B. Łydźba-Kopczyńska, E. Sentandreu, *J. Raman Spectrosc.* **2018**, 49, 842.
- [64] M. Shimoyama, H. Maeda, H. Sato, T. Ninomiya, Y. Ozaki, *Appl. Spectrosc.* **1997**, 51, 1154.
- [65] R. H. Brody, H. G. M. Edwards, A. M. Pollard, *Anal. Chim. Acta* **2001**, 427, 223.
- [66] S. C. Park, M. Kim, J. Park, H. Chung, H. Y. Kim, *J. Raman Spectrosc.* **2009**, 40, 2187.
- [67] P. Stenger, C. Ky, C. Reisser, J. Duboisset, H. Dicko, P. Durand, L. Quintric, S. Planes, J. Vidal-Dupiol, *Genes (Basel)*. **2021**, 12, 1.
- [68] Y. Bi, Y. Zhang, J. Yan, Z. Wu, Y. Li, *Plasma Sci. Technol.* **2015**, 17, 923.
- [69] K. Chang, S. Lee, J. Park, H. Chung, *Talanta* **2016**, 149, 335.
- [70] T. Liu, L. Kong, L. Lin, H. Xu, Z. Zhou, M. Huang, *Laser Phys.* **2022**, 32, 1.
- [71] M. Dumańska-Słowik, A. Wesełucha-Birczyńska, L. Natkaniec-Nowak, A. Gawel, A. Włodek, K. Kulmaczewska, *J. Raman Spectrosc.* **2020**, 51, 346.
- [72] M. C. Caggiani, M. Cavarra, G. Barone, A. Coccato, A. M. Manenti, P. Mazzoleni, *J. Raman Spectrosc.* **2023**, 1. <https://doi.org/10.1002/jrs.6588>
- [73] A. Zoppi, C. Lofrumento, E. M. Castellucci, P. Sciau, *J. Raman Spectrosc.* **2008**, 39, 40.
- [74] L. Medeghini, S. Mignardi, C. De Vito, A. M. Conte, *Microchem. J.* **2016**, 125, 224.
- [75] F. Bianchi, N. Riboni, V. Trolla, G. Furlan, G. Avantiaggiato, G. Iacobellis, M. Careri, *Talanta* **2016**, 154, 467.
- [76] D. Quintero Balbas, G. Lanterna, C. Cirrincione, M. Ricci, M. Becucci, R. Fontana, J. Striova, *J. Raman Spectrosc.* **2021**, 53, 593.
- [77] R. Simonetti, P. Oliveri, A. Henry, L. Duponchel, S. Lanteri, *Talanta* **2016**, 149, 250.

How to cite this article: A. Coccato, M. C. Caggiani, *J Raman Spectrosc* **2024**, 55(2), 125. <https://doi.org/10.1002/jrs.6621>