OpenPCA and Raman mapping to decipher complex spectral datasets from multi-component samples: application to cannabis trichomes

Janani Balasubramanian[1], Elisa Crocioni[2], Mattia Frattini[2], Scott Hill[1], Darryen Sands[1], Chiara Zanchi[2], Matteo Tommasini[2], Nisha Rani Agarwal[1*]

[1]Nano-imaging and Spectroscopy Laboratory, Faculty of Science, University of Ontario Institute of Technology, 2000 Simcoe Street North, Oshawa ON L1G 0C5, Canada

[2]Department of Chemistry, Chemical and Materials Engineering 'Giulio Natta', Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milan 20133, Italy

*Corresponding author: nisha.agarwal@ontariotechu.ca

Abstract

The development of analytical techniques that decode chemical information in complex biochemical samples to discriminate different structural components may open the way for several new findings. In this study, principal component analysis (PCA) is carried out using an *ad hoc* Matlab coding that provides a transparent access to multivariate analysis of Raman mapping datasets. Here, we illustrated the efficacy of this method to extract meaningful results from Raman images of *Cannabis sativa* trichomes. A large dataset of *Cannabis* trichome comprising of 441 Raman spectra was examined for the first time using our OpenPCA. By mapping the chemical distribution in the trichome, we could locate the secretary vesicles in the PC score maps generated from the mapped Raman spectra. Black-box PCA solutions available in commercial software can be limited by rigid input interfaces which may prevent obtaining information by tuning the PCA analysis on selected wavenumber ranges. Hence, the OpenPCA scripts facilitate the task of obtaining key information from widely distributed range of wavenumbers that are characteristic to a specific cannabinoid, namely Δ^9-THC and CBD. Overall, the PCA-coding algorithm shows advantages in decoding Raman spectra that could be extended to handle all kinds of datasets with simultaneous spatial and chemical details.

Keywords: OpenPCA; Raman mapping; Cannabinoids; Trichomes; chemical analysis of spectra

1. Introduction

Cannabis sativa from family *Cannabaceae* is predominantly a dioecious annual herb that have significant importance in the industrial and medicinal field [1]. It is widely utilized and consumed for manifold purposes including personal-care products, natural fungicides, food additives, essential oils, and medical formulations. The cannabinoids are the prime bioactive substance in *C. sativa* and are a promising therapeutic candidate for cancer treatment, neurological diseases, appetite disorders, and inflammation [2]. The cannabinoids are secreted from trichome structures which are the specialized hairs covering female inflorescences. Three kinds of trichomes were observed with different morphologies, namely capitate-stalked, capitate-sessile, and bulbous [3]. The abundantly present capitate-stalk of trichomes that contains highest cannabinoid levels features a basal cell, secretory cells, several stalk cells, and a large sub-cuticular storage cavity [2]. Phytocannabinoids are naturally synthesized in the trichomes, and constitute a unique group of terpenophenolics compounds with three leading members, namely, tetrahydrocannabinol (THC), cannabinol (CBN), and cannabidiol (CBD). Among them, Δ^9-THC is considered as an illicit drug because it possesses psychotropic effects and hence it is restricted in most countries [4]. However, CBD is a substance with non-psychoactive nature, and acts as an antagonist to THC effects. Further, CBD displays neuroprotective, anti-rheumatoid arthritis, anti-nausea, anxiolytic, anti-spasmodic, and anticonvulsant properties [5].

Despite the medicinal and economic significance of trichomes, cannabinoid levels, chemical profile, and its distribution in the trichomes remain uninvestigated and poorly understood [6]. To analyse the trichomes and to extract their chemical composition, a powerful non-destructive Raman spectroscopy technique was adopted, but no systematic information was reported about the spatial distribution of the chemical species in the plant [7, 8]. However, the high spatial resolution of confocal micro-Raman spectroscopy (proved for instance in ref.

[9]) can be applied to the context of Cannabis characterization. One could directly probe Cannabis samples such as the trichomes by micro-Raman mapping techniques and gather information about the chemical distribution of the cannabinoid substances at the micrometric scale. To assign the chemo-markers to the experimental trichome spectra, the Raman spectra of pure cannabinoids that are present in high levels and free of other chemicals are now gaining momentum and being reported in the literature (see e.g. [8, 10]). Hence Raman mapping is a powerful tool that would eventually allow one to detect the presence of specific cannabinoids, their spatial distribution, and their concentration in a sample.

The Raman spectroscopic analysis of the trichome could be useful in several ways. For instance, the determination of cannabinoid levels can be employed to design a sensor by which the point of maximum maturation of the plant and its best harvesting time can be identified. The development of analytical methods for testing cannabinoid substances could display a potential application in law enforcement and forensic applications. In addition, according to the legal framework established by governments and regulatory bodies, the farmers would be able to distinguish between two varieties of hemp (cultivated for fibre production) or marijuana (cultivated for drug and medical purposes) based on the chemical threshold levels, *e.g.*, Δ^9-THC in hemp is ≤ 0.2 % and in marijuana it is > 0.2 % [5].

Despite its benefits, the decoding of large dataset of Raman spectra is an arduous task [11]. Normally, tens to thousands of spectra are collected from pixel scans on a sample to create a Raman mapping. As a result, it is often a tedious, time consuming process to decode this large data matrix made of a multitude of signal intensities at different wavenumbers [12]. In this work, a sample area of 21 µm × 21 µm with a pixel size of 1 µm × 1 µm were scanned to acquire 441 spectra as a matrix. Commonly, a wavenumber that is specific to a cannabinoid of interest is selected and a Raman image is mapped based on the varying intensity at that point. If the chosen wavenumber is a unique characteristic peak of the specific compound of interest, with no background interference, the image generated with relation to the scanned area carries direct information about the spatial distribution of the chemical species of interest. The drawback behind carrying out this method in regular software includes the limited information associated with the produced image that corresponds only to the selected wavenumber, and may display low signal-to-noise ratio and/or loss of signal. Usually, key information contained within the dataset matrix of the Raman spectra is widely distributed throughout the dataset [13], and a single-peak data analysis approach may lose significant details. Principal component analysis (PCA) is an effective statistical method capable to handle a complex large data matrix by reducing the dimensionality while preserving the most critical features [14]. However, despite PCA being a well-established analysis tool, it often shows up as a black box in the software driving spectrometers, and its true potential may be hard to be fully appreciated. In this work, a novel coding approach is presented to introduce PCA in Matlab where the background process is observable, open and it is a white box approach. The software coding is based on a fully algebraic approach that focuses on the variance-covariance matrix of the dataset and its spectral decomposition. This allows the easier control over the multivariate dataset, and facilitates the analysis and tuning of the right parameters. In this study, we carry out the analysis of a novel dataset of Raman spectra of *C. sativa* trichomes by the implementation of such white-box PCA approach. This work demonstrates the potential of a label-free and non-destructive method based on principal component analysis of the micro-Raman mapping of trichome to understand different structures and chemo-types along with their spatial distribution. Nevertheless, this technique is not limited only to *Cannabis*, but it could be amply extended for handling and investigating all kinds of natural or technological processes that deal with simultaneous spatial and chemical details.

2. Materials and methods

2.1 Experimental

The Raman spectra of pure THC and CBD cannabinoids were analysed. For micro Raman analysis, 5 µL of the Δ^9-THC solution (1 mg/mL) prepared with methanol solvent was dropped on a glass slide. In the case of CBD, a sample from pure CBD powders was directly used (with no solvent preparation) to carry out micro-Raman measurements. The Cannabis seeds were obtained from a Cannabis licensed distributor in Oshawa (ON, Canada). The trichomes were procured from the grown plant during the flowering phase. The trichome sample was used as received to collect the Raman spectra. The scanned area of the trichome was over a grid of 21 x 21 points, with 1 µm spacing. For the two mappings reported here, each single point Raman collection had a duration of 10 sec with either 1 accumulation (sample 1) or 10 accumulations (sample 2).

2.2 Instrumentation

The Raman Spectra were obtained using a Renishaw Raman instrument equipped with a 532 nm laser. The spectra were acquired at a laser power of 1%, with a 50x objective, the exposure time was 10 s, 1 - 10 accumulations, and ranged from 100 cm^{-1} to 4000 cm^{-1}. Raman spectrum of methanol was acquired as well for control measurements.

The fluorescence background associated with the obtained Raman spectra, especially with shorter wavelengths makes it harder to read. To resolve this issue, a completely automated software from Renisha, Windows®-based Raman Environment (WiRE), included with the Raman spectrometer was applied for the acquisition of the mappings and the removal of background signal. In addition, the WiRE has control over both Raman data acquisition and data processing options. Thus, the fluorescence background subtraction allows for a clearer visualization of the Raman data. However, numerical artifacts could also be introduced in this process and should be carefully noted to avoid misleading conclusions.

2.3 PCA – as implemented in OpenPCA

The PCA was introduced in 1933 by Harold Hotelling in the context of psicometric data analysis [15]. PCA has been widely applied to many fields where multivariate datasets have to be dealt with. However, PCA remain as a black box that is poorly understood. A novel coding approach is required to introduce PCA in Matlab that allows the background process to be observable, and modifiable. The easier approach to introduce PCA, by also taking into consideration its numerical implementation in Matlab, is through a fully algebraic approach that focuses on the variance-covariance matrix of the dataset and its spectral decomposition. Let us introduce first the multivariate dataset matrix $\mathbf{X}_{ov}$, which along each row stores the results of one multivariate observation along a given number of variables (N_v). The adopted notation for the dataset matrix highlights the different role of row *vs*. column indexes. The different observations are identified in the $\mathbf{X}_{ov}$ matrix by the row index (o), whereas the different variables of each multivariate measurement (observation) are identified by the column index (v). In the context of spectroscopy, each row represents one spectrum, and the different variables are the wavenumbers at which the instrument has recorded a given spectral intensity (e.g., Raman intensity, or absorbance). Hence, because of the adopted notation, we have the following identities:

$$\boldsymbol{X} = \boldsymbol{X}_{ov} \qquad (1a)$$

$$\boldsymbol{X}_{vo} = (\boldsymbol{X}_{ov})^t = \boldsymbol{X}^t \quad (1b)$$

where ^{t} indicates matrix transposition. As described later, the variance-covariance matrix among the variables of the dataset can be straightforwardly introduced through the matrix of the centered dataset, $\boldsymbol{\chi}_{ov}$:

$$\boldsymbol{\chi}_{ov} = \boldsymbol{X}_{ov} - \langle \boldsymbol{X}_v \rangle \quad (2)$$

Where $\langle \boldsymbol{X}_v \rangle$ represents the row vector of the average values of the variables over the number of N_o observations, and its v-th element is given by:

$$\langle X_v \rangle = \frac{1}{N_o}\sum_{o=1}^{N_o} X_{ov} \quad (3)$$

we adopt in Eq. (2) the same abuse of notation used in Matlab: by subtracting a row vector to a matrix actually one subtracts the given row vector to each row of the matrix. Hence Eq. (2) is implemented in Matlab as simply as chi = X - mean(X), because the Matlab function mean(X) gives the row vector corresponding to the average of all the rows of the **X** matrix – which effectively corresponds to averaging out with respect to the available observations (see above). By using the cantered dataset matrix, the variance-covariance matrix among the variables of the dataset ($\mathbf{\Sigma}_{\mathrm{vv}}$) can be introduced as follows:

$$\mathbf{\Sigma}_{\mathrm{vv}} = \frac{1}{N_o - 1}\,\boldsymbol{\chi}_{vo}\boldsymbol{\chi}_{ov} \quad (4)$$

Clearly, by definition, **Σ** is a symmetric matrix, and it is positive definite. Therefore it admits spectral decomposition by the orthogonal matrix of its eigenvectors, and the eigenvalues are positive quantities [16]. The matrix eigenvalue problem of the variance-covariance matrix is written as:

$$\boldsymbol{\Sigma}_{vv}\boldsymbol{L}_{vs} = \boldsymbol{L}_{vs}\boldsymbol{\sigma}_{ss} \qquad (5)$$

In Eq. (5) $\boldsymbol{\sigma}_{\mathrm{ss}}$ is the diagonal matrix of the eigenvalues of $\mathbf{\Sigma}_{\mathrm{vv}}$ and $\mathbf{L}_{\mathrm{vs}}$ is the orthogonal matrix of the eigenvectors of $\mathbf{\Sigma}_{\mathrm{vv}}$. The orthogonality of $\mathbf{L}_{\mathrm{vs}}$ implies:

$$\boldsymbol{L}_{vs}\boldsymbol{L}_{sv} = \mathbf{1}_{vv} \qquad (6)$$

$$\boldsymbol{L}_{sv}\boldsymbol{L}_{vs} = \mathbf{1}_{ss} \qquad (7)$$

Therefore, by left-multiplying Eq. (5) by $\mathbf{L}_{\mathrm{sv}}$, and by considering its orthonormality, one obtains the spectral decomposition of the variance-covariance matrix:

$$\boldsymbol{L}_{sv}\boldsymbol{\Sigma}_{vv}\boldsymbol{L}_{vs} = \boldsymbol{\sigma}_{ss} \qquad (8)$$

By substituting in the right-hand side of Eq. (8) the definition of $\mathbf{\Sigma}_{\mathrm{vv}} = \chi_{vo}\chi_{ov}$ / ($\mathrm{N_o}$ - 1) (cfr. Eq. 4), one obtains:

$$\boldsymbol{\sigma}_{ss} = \frac{1}{N_o - 1} \boldsymbol{L}_{sv}\boldsymbol{\chi}_{vo}\boldsymbol{\chi}_{ov}\boldsymbol{L}_{vs} \qquad (9)$$

Similarly to the definition of a variance-covariance matrix (Eq. (4)), it is then possible to identify in the right-hand side of Eq. (9) a structure given by the product of a matrix (defined **S**) by its transpose:

$$\boldsymbol{\sigma}_{ss} = \left[\frac{1}{\sqrt{N_O - 1}}\boldsymbol{L}_{sv}\boldsymbol{\chi}_{vo}\right]\left[\boldsymbol{\chi}_{ov}\boldsymbol{L}_{vs}\frac{1}{\sqrt{N_o - 1}}\right] = \boldsymbol{S}_{so}\boldsymbol{S}_{os} = \boldsymbol{S}^t\boldsymbol{S} \qquad (10)$$

The rows of such a matrix ($\mathbf{S}_{os}$) - named the scores matrix - define the observations (o label) through the so-called principal components (s label):

$$\boldsymbol{S}_{os} = \left[\frac{1}{\sqrt{N_O - 1}}\boldsymbol{\chi}_{ov}\boldsymbol{L}_{vs}\right] \qquad (11)$$

The matrix of the eigenvectors of the variance-covariance matrix ($\mathbf{L}_{vs}$), which is named the loadings matrix, defines the linear relationship existing between each principal component and the set of variables. The PCA scatterplot e.g. of the first two principal components (PC1, PC2) can be obtained by plotting on the Cartesian xy plane the first column of the **S** matrix (x coordinates) vs. the second column of the **S** matrix (y coordinates). This plot immediately allows judging data clustering or the presence of outliers. Such scatterplots can be of course extended to other principal components (*i.e.*, to other columns of the **S** matrix). Sometimes, the scores matrix is normalized in such a way to produce an associated variance-covariance matrix (over the s variables) that is a unit matrix. This normalization is simply done as follows:

$$\boldsymbol{S}'_{os} = \boldsymbol{S}_{os}\boldsymbol{\sigma}_{ss}^{-\frac{1}{2}} \qquad (12)$$

It is then straightforward to show that the variance-covariance matrix associated to $\boldsymbol{S}'_{os}$ is a unit matrix:

$$(\boldsymbol{S}')^t\boldsymbol{S}' = \left(\boldsymbol{\sigma}_{ss}^{-\frac{1}{2}}\boldsymbol{S}_{so}\right)\left(\boldsymbol{S}_{os}\boldsymbol{\sigma}_{ss}^{-\frac{1}{2}}\right) = \boldsymbol{\sigma}_{ss}^{-\frac{1}{2}}\boldsymbol{\sigma}_{ss}\boldsymbol{\sigma}_{ss}^{-\frac{1}{2}} = \mathbf{1} \qquad (13)$$

3. Results and discussion

3.1 Spectroscopic characterization of Δ^9-THC and CBD

The volatile nature of Δ^9-THC makes it hard to control the formation of solid samples that tend to sublimate (which is why Δ^9-THC is supplied as methanol solution). To our fortune some microcrystalline aggregate was stable for just enough time to perform a Raman mapping. Unfortunately, the replication of the same conditions to obtain the microcrystals was not succeeded anymore and only this single Raman data set was obtained on Δ^9-THC. The reason behind this finding could be hypothesized to the presence of a nucleation site created from a piece of dirt on the glass slide that prevented evaporation. This is the reason behind the lack of data with respect to pure cannabinoid samples. Most of the Raman mapped area did not belong to Δ^9-THC, therefore the obtained Raman data set was processed by PCA to get 2D Raman mapping image. Cluster analysis was performed to evaluate the variation in the Raman signal with respect to the position of the laser spot in the measurement. Matlab was used to perform multivariate analysis and to plot the spectra. The Raman spectrum of the ephemeral sample of Δ^9-THC and CBD is shown is Figure 1 along with its chemical structure that displays very sharp and defined peaks, most likely due to the ordered crystalline state of the sample. It is remarkable to note that X-ray diffraction data of Δ^9-THC was not available in the literature owing to its volatile nature. The peaks corresponding to CH stretching modes observed at 2847, 2913, 2990 and 3056 cm^{-1} are not highly structure-specific markers, but the expected peaks for both aliphatic (2847, 2913, 2990 cm^{-1}) and aromatic (3056 cm^{-1}) structures were observed. The OH stretching modes that should be observable around 3500 cm^{-1} were not detected. The fingerprint region of the spectrum is observed in the range of 200 to 1800 cm^{-1} that contains the characteristics collective modes of the molecule. All the Raman spectra of Δ^9-THC found in literature show only the fingerprint region, so the comparison will be limited to this range of frequencies [8, 17, 18]. The Raman spectrum of Δ^9-THC recorded in this work with a 532 nm laser and the reference Raman spectra retrieved from 633 nm laser spectrum in the study of

Islam et al. (2020) [18] show similar strong peaks close to 1002, 1087 and 1605 cm^{-1}, however the relative Raman intensities in our Δ^9-THC Raman spectrum significantly differ from those reported, for instance, in ref. [8]. We believe that this is caused by the joint effect of the laser polarization and uncontrolled crystalline orientation of the ephemeral Δ^9-THC sample. The intense peak at 1002 is assigned to the breathing of the aryl group. The bands observed in the 1600 – 1670 cm^{-1} range are assigned to the ring stretching of the aryl group and to the stretching of the C=C bond in one of the rings of THC [18]. The shifting of the bands compared to the previous reports can be caused by many different factors: the wavelength of the light used for the analysis, the different aggregation states of the samples, and the effect of the solvents from which the sample was obtained [19, 20]. The other peaks in the fingerprint region is assigned as follows: 715 cm^{-1} CH deformation, 1032 cm^{-1} C-C stretching, 1437 cm^{-1} CH_3 twist and bend [21].

The CBD sample is a pure crystalline powder and it provides a very neat FT-Raman spectrum with sharp and well-defined peaks. Our spectrum compares well with literature data [8] but it also reports the signal in the CH stretching region. Out of the whole spectrum of CBD, the peaks at 1433 and 1662 cm^{-1} in the fingerprint region and the peak at 2927 cm^{-1} in the high frequency region are predominant and could be clearly recognized. The peak at 1433 cm^{-1} is ascribed to the vibrations of the hydroxyl (OH) group, hexene ring stretch, and CH bend of the benzene ring [18]. The peak at 1662 cm^{-1} corresponds to the C=C stretch in cyclohexane [22]. In both Δ^9-THC and CBD, the high frequency region between 2500 and 3600 cm^{-1} is ascribed to the CH (around 3000 cm^{-1}) and OH (around 3500 cm^{-1}) stretching vibrations. Δ^9-THC and CBD have similar chemical structures. Compared to Δ^9-THC, CBD has a strong peak at 1433 cm^{-1} which can be used to distinguish CBD from Δ^9-THC.

3.2 Spectroscopic characterization of trichomes

Glandular trichomes are the structures that are observed covering the surface of each floral inflorescence of *C. Sativa* and are the site of production of metabolites (cannabinoids). We examine here by micro-Raman spectroscopy how information may be collected about the chemical composition and microscopic spatial distribution of cannabinoids in trichomes. Here, we aim to differentiate the structure of secretary vesicles in the whole trichome using Raman spectroscopy based on the expected variation in the cannabinoids levels. According to Livingston et al. [6] some regions in the trichomes can be identified as the secretory vesicles that are characterized by the presence of higher level of cannabinoids than others. The bright field image of trichome samples 1 and 2 is depicted in Figure 2a and 2d, respectively. The average spectrum computed over the two mapped areas of the trichomes is reported in Figure 2c and 2f for sample 1 and 2, respectively. An intense Raman peak at 1295 cm^{-1} is observed in the average spectra of both samples. Based on literature [8], such Raman signal can be assigned to a few cannabinoids (THC, THCA and CBGA) that exhibit a strong Raman peak in this position of the spectrum. The Raman map generated with such peak (1295 cm^{-1}) of the Raman spectra of both trichome samples is reported in Figure 2b and 2e and demonstrate the information which could be obtained from direct inspection of the data at a particular wavenumber, with no advanced dataset processing.

3.3 PCA of the micro-Raman mapping of a Cannabis trichome

The PCA of the Raman spectra of the map can identify the most important variations of the spectra across the dataset, which can be used to produce unsupervised grouping of image pixels on the basis of their Raman signature, which reflects the chemical composition in that location. Once the PCA scripts are run in the Matlab environment, a screeplot in the logarithmic scale is obtained which allows to identify the most relevant principal components (PCs) as shown in Figure 3a (screeplot of sample 1). The screeplot is a representation of the principal variances in the multivariate dataset, where the principal components are sorted by decreasing principal variance (*i.e.*, by decreasing eigenvalues of the covariance matrix). In Figure 3a it is noticed how quickly the principal variances decrease along the screeplot. For this reason, just the loadings along the first four PCs were analyzed. The components starting from PC5 have been neglected since their variance is very low compared to the previous PCs.

At first, the Raman analysis of sample 1 is presented. The PC loadings were investigated to obtain the chemical information behind different PCs. The scoremaps and loadings of PC1 to PC4 are reported in Fig. 3f. The loadings of PC1 to PC3 convey chemical information, whereas PC4 mostly displays an undulatory behaviour over a noisy signal. The PC1 loadings clearly display an intense fluorescence background, which is less evident in the PC2 loadings. In PC1 and PC2 it is not possible to fully decouple the fluorescence and Raman contributions, as PC1 and PC2 are both characterized by a fluorescence and a Raman component. However, while in PC1 it is the fluorescence component to be more intense, in PC2 it is the Raman contribution to be dominant. Hence, with a little degree of approximation, when considering the associated scoremaps we may assume that PC1 describes the areas of the trichome that are more fluorescent, whereas PC2 indicates the areas with stronger Raman signal. Since the fluorescence signal is less strongly related to the chemical structure of the compounds than it is the Raman signal, one may expect to get chemical information out of PC2. The PC1 to PC4 scoremaps were obtained with the Matlab PCA scripts and are also reported in Figure 3. In the scoremap, each single measurement point in the Raman mapping experiment is represented as separate pixel whose colour shade identifies the score value. The scoremaps are a representation of the spatial variation of the Raman spectra along a specific principal component – for instance high PC1 score values indicate a strong fluorescence, whereas a high PC2 score indicates an overall strong Raman signal. By inspecting a given scoremap associated to a given Raman peak one may infer the local changes of the cannabinoid levels in the trichome sample. In the first scoremap (Fig. 3f), we could observe three dark spots. Since the PC1 loading displays the same shape as the average spectrum the PC1 can be regarded as the overall strength of the signal (which includes a strong fluorescence background). To be more precise, this indicates the overall variation that arises from the different point to point focusing on the curved bulb of the trichome surface. In the second scoremap, we can identify a central dark region. However, since it is a complex data set with the combination of several chemical species, it is difficult to interpret this structure more precisely. Therefore, using the PCA scripts in Matlab, the range of wavenumber is selected from 1580 cm^{-1} to 1700 cm^{-1}, which contains a characteristic peak of cannabinoids observed in both THC and CBD. The Fig. 3g represents the PC1 and PC2 scoremap and loadings within the selected 1580-1700 cm^{-1} wavenumber range. The PC1 scoremap does not include very specific chemical information as the loading looks similar to the average spectra (Fig. 3d), hence this scoremap can be associated with the florescence background over the Raman spectra. In the PC2 scoremap three dark circular spots are observed that indicate high concentration of cannabinoids, while the bright regions corresponds to low concentration (the negative scale of the PC2 score is due to the negative sign of the peaks in the PC2 loadings). The dimensions of the dark spots are about 8-12 µm. This compares well with the dimension of the vesicles in the *C. Sativa* trichome. In general, THC is accumulated in this specific region of the trichomes called vesicles. Based on literature [6] the dimension of the vesicles in the trichomes can be of the order of 10 µm, which is compatible with the size of the dark region in scoremap 2. We may conclude that the dark regions, which are characterized by a higher accumulation of cannabinoids, represent the vesicles of the trichomes. In addition, the region of the most intense peak in the trichome spectra 1280-1310 cm^{-1} containing the sharp band at 1295 cm^{-1} was analyzed. Also in this case we ignore PC1 (see above). The PC2 loadings show that the spectral change with respect to the average spectrum follows a pattern where the peak get more narrow and intense. The corresponding scoremap displays three bright circular regions that confirms the accumulation of cannabinoid substances in the trichome, which can be attributed to the presence of the secretary vesicles.

In Fig 4a we report the screeplot of the trichome sample 2 obtained from the PCA analysis of the Raman spectra over the full wavenumber range. The associated PC1-4 loadings and scoremap are reported in Fig. 4f. As for sample 1 the first scoremap (Fig. 4f) displays little chemical information since the corresponding PC1 loading is mainly interpreted as signal background. In the second scoremap, we can identify a few circular dark spots that could be associated to the overall presence of chemical species. To extract more details about the specific chemical species we selected the wavenumber range that contains characteristic peaks of cannabinoids (1600-1700 cm^{-1}). The Fig. 4b and d represents the screeplot and average spectrum corresponding to this range. Fig. 4g represents scoremap and loadings of PC1 and PC2. As mentioned earlier, the scoremap of PC1 does not include any specific chemical information and may be associated with the overall background. We notice a large dark spot in the PC2 scoremap. Since the dimensions of the observed dark area is 20 µm, this compares with the size of a trichome structure [6]where the presence of chemical species make it Raman active. The PC3 loading does not show background (fluorescence) contributions. In the PC3 scoremap we can observe three dark spots with a size of about 8-10 µm that is compatible with the size of vesicles [6]. The same pattern was observed for the PCA analysis in the restricted wavenumber region comprinsing the high intensity Raman peak (1270-1290 cm^{-1}), see Fig. 4c, e, and h. The reproducible pattern observed in the scoremaps computed with different spectral ranges confirms the accumulation of cannabinoids in those regions.

4. Conclusions

Raman mapping of chemically complex samples can provide access to chemical compositional information though the analysis of the spatial variation of the Raman signal. This is very tedious to do manually and it is greatly simplified by applying principal component analysis to the dataset. The OpenPCA framework offers a way to carry out routine PCA analysis of Raman mappings, customising the spectral range and the selection of the principal components of interest. By plotting the scores of selected PCs on the map, one can easily spot regions of the samples where chemical variations occur, as they are witnessed by the changes is the Raman markers of given species. This method was implemented to spot the vesicles structures in the cannabis trichome head based on the rich accumulation of cannabinoids. This could open the doors to post-process various datasets that deals with chemical heterogeneity and its spatial distribution.

Acknowledgements

The authors would like to thank the researchers Dr. Cinzia Citti and Dr. Giuseppe Cannazza (CNRNANOTEC- Institute of Nanotechnology, Lecce) for the interesting discussions on cannabinoids and for providing the crystalline Cannabidiol sample used in this work.

References

1. Hesami, M., et al., *Recent advances in cannabis biotechnology.* Industrial Crops and Products, 2020. **158**: p. 113026.
2. Rodziewicz, P., et al., *Cannabinoid synthases and osmoprotective metabolites accumulate in the exudates of Cannabis sativa L. glandular trichomes.* Plant Science, 2019. **284**: p. 108-116.
3. Carretero, P.L., et al., *Glandular trichomes affect mobility and predatory behavior of two aphid predators on medicinal cannabis.* Biological Control, 2022. **170**: p. 104932.
4. Liu, Y., et al., *Cannabis sativa bioactive compounds and their extraction, separation, purification, and identification technologies: An updated review.* TrAC Trends in Analytical Chemistry, 2022: p. 116554.
5. Micalizzi, G., et al., *Cannabis Sativa L.: A comprehensive review on the analytical methodologies for cannabinoids and terpenes characterization.* Journal of Chromatography A, 2021. **1637**: p. 461864.

6. Livingston, S.J., et al., *Cannabis glandular trichomes alter morphology and metabolite content during flower maturation.* The Plant Journal, 2020. **101**(1): p. 37-56.

7. Ramos-Guerrero, L., et al., *Classification of Various Marijuana Varieties by Raman Microscopy and Chemometrics.* Toxics, 2022. **10**(3): p. 115.

8. Sanchez, L., D. Baltensperger, and D. Kurouski, *Raman-based differentiation of hemp, Cannabidiol-rich hemp, and Cannabis.* Analytical chemistry, 2020. **92**(11): p. 7733-7737.

9. Badou, A., et al., *New insight on spatial localization and microstructures of calcite-aragonite interfaces in adult shell of Haliotis tuberculata: Investigations of wild and farmed abalones by FTIR and Raman mapping.* Journal of Structural Biology, 2022. **214**(2): p. 107854.

10. Tay, L.-L., J. Hulse, and R. Paroli, *FTIR and Raman Spectroscopic Characterization of Cannabinoids.* Canadian Journal of Chemistry, 2022(ja).

11. von der Esch, E., et al., *TUM-ParticleTyper: A detection and quantification tool for automated analysis of (Microplastic) particles and fibers.* Plos one, 2020. **15**(6): p. e0234766.

12. Sobhani, Z., et al., *Identification and visualisation of microplastics/nanoplastics by Raman imaging (i): Down to 100 nm.* Water research, 2020. **174**: p. 115658.

13. Fang, C., et al., *Identification and visualisation of microplastics via PCA to decode Raman spectrum matrix towards imaging.* Chemosphere, 2022. **286**: p. 131736.

14. Halstead, J.E., et al., *Assessment tools for microplastics and natural fibres ingested by fish in an urbanised estuary.* Environmental Pollution, 2018. **234**: p. 552-561.

15. Hotelling, H., *Analysis of a complex of statistical variables into principal components.* Journal of educational psychology, 1933. **24**(6): p. 417.

16. Schott, J.R., *Matrix analysis for statistics*. 2016: John Wiley & Sons.

17. Fedchak, S., *Presumptive field testing using portable raman spectroscopy.* Las Vegas Metropolitan Police Department, USA, 2014: p. 1-59.

18. Islam, S.K., et al., *An analysis of tetrahydrocannabinol (THC) and its analogs using surface enhanced Raman Scattering (SERS).* Chemical Physics, 2020. **536**: p. 110812.

19. Leonard, J., et al., *SERS, Raman, and DFT analyses of fentanyl and carfentanil: Toward detection of trace samples.* Journal of Raman Spectroscopy, 2017. **48**(10): p. 1323-1329.

20. Wahadoszamen, M., et al., *Laser Raman spectroscopy with different excitation sources and extension to surface enhanced Raman spectroscopy.* Journal of Spectroscopy, 2015. **2015**.

21. Yüksel, S., et al., *Trace detection of tetrahydrocannabinol (THC) with a SERS-based capillary platform prepared by the in situ microwave synthesis of AgNPs.* Analytica Chimica Acta, 2016. **939**: p. 93-100.

22. Sigworth, K., *Raman spectroscopy study of delta-9-tetrahydrocannabinol and cannabidiol and their hydrogen-bonding activities.* 2020.

Figures

Figure 1

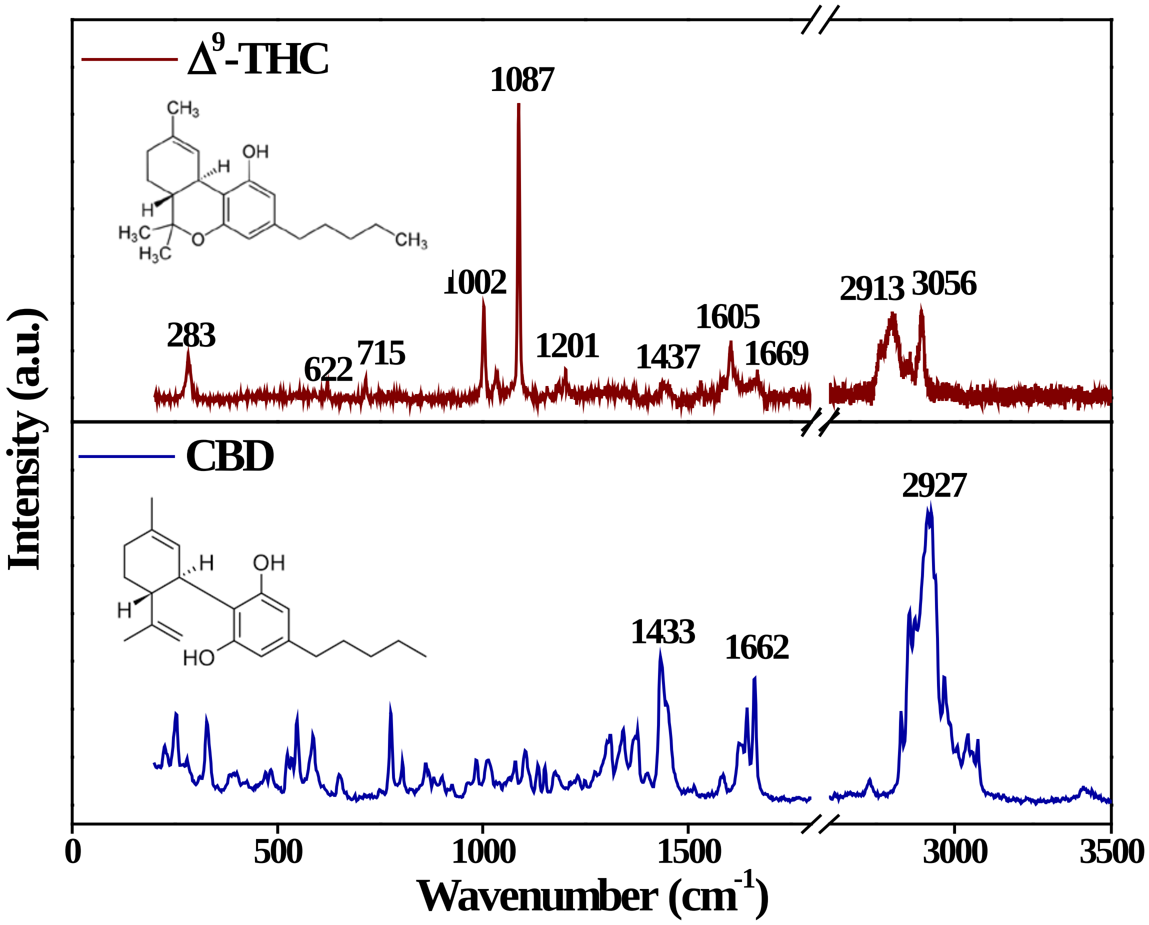

Figure 1. Pure cannabinoid spectra: the Raman spectrum of the ephemeral crystal of Δ9-tetrahydrocannabinol and the Raman spectrum of the pure cannabidiol; (inset - chemical structure of Δ9-THC and CBD).

Figure 2

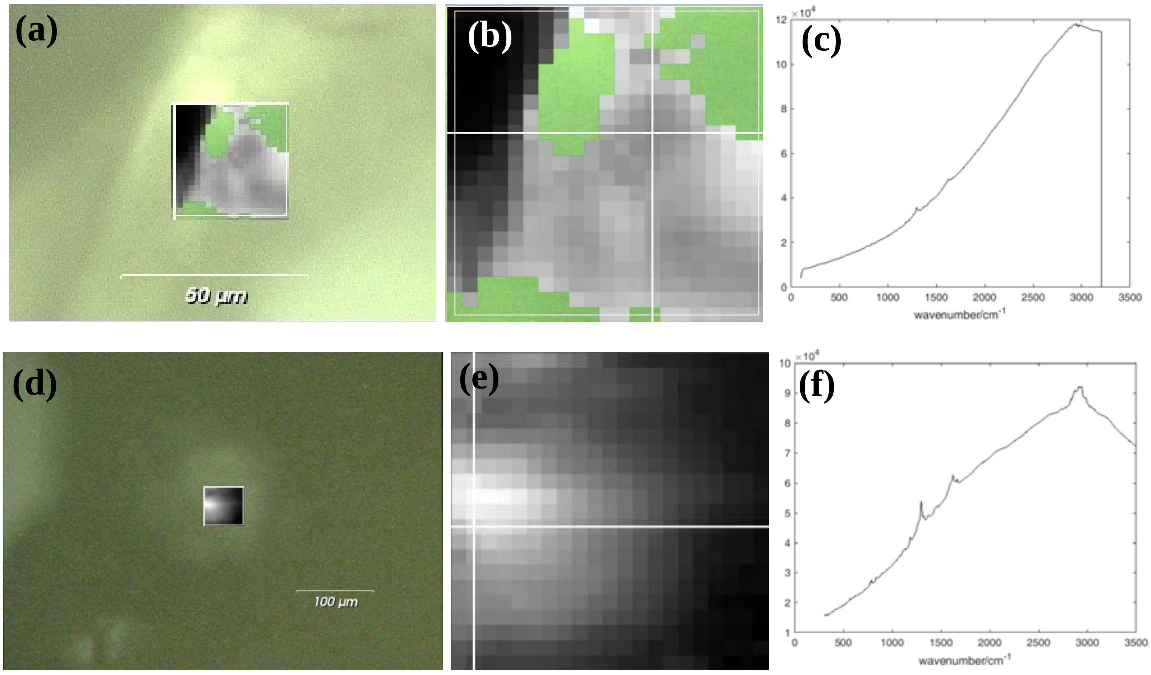

Figure 2. Trichome sample 1: (a) Bright field image of the sample, (b) Map of the Raman intensity of the peak at 1295 cm^{-1} and (c) average Raman spectrum over the mapped area of the trichome sample. Trichome sample 2: (d) Bright field image of the sample, (e) Map of the Raman intensity at 1295 cm^{-1} and (f) average Raman spectrum over the mapped area of the trichome sample.

Figure 3

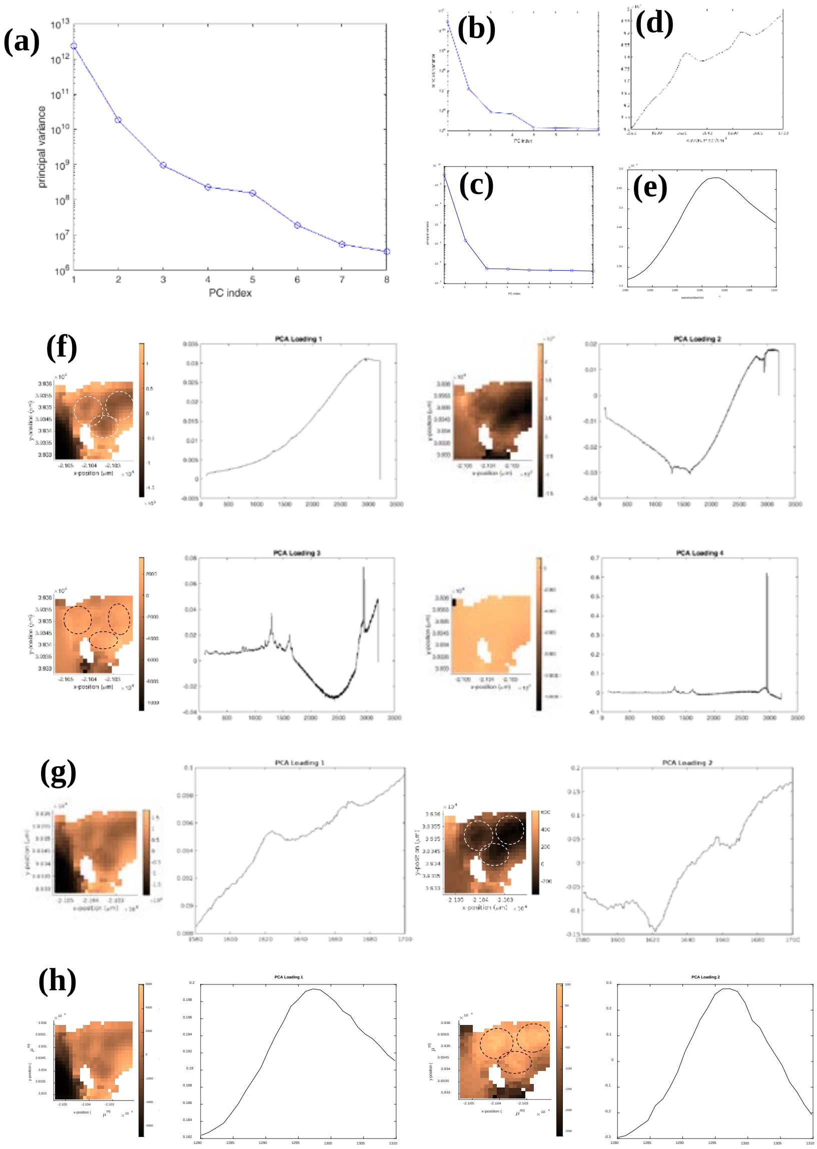

Figure 3. PCA analysis of trichome sample 1: (a) Screeplot of the Principal Components in the logarithmic scale on the y axis: variance of the dataset as a function of the PC index(s), (b) Screeplot of the filtered dataset in the spectral range 1580-1700 cm^{-1}, (c) Screeplot of the filtered dataset in the spectral range 1280-1310 cm^{-1}; (d) average spectrum in the region between 1580-1700 cm^{-1}, (e) average spectrum in the region between 1280-1310 cm^{-1}; (f) Scoremaps and loadings of PC1, PC2, PC3 and PC4, (g) Scoremaps and loadings of PC1 and PC2 of the filtered dataset in the spectral range 1580-1700 cm^{-1} and (h) Scoremaps and loadings of PC1 and PC2 of the filtered dataset in the spectral range 1280-1310 cm^{-1}.

Figure 4

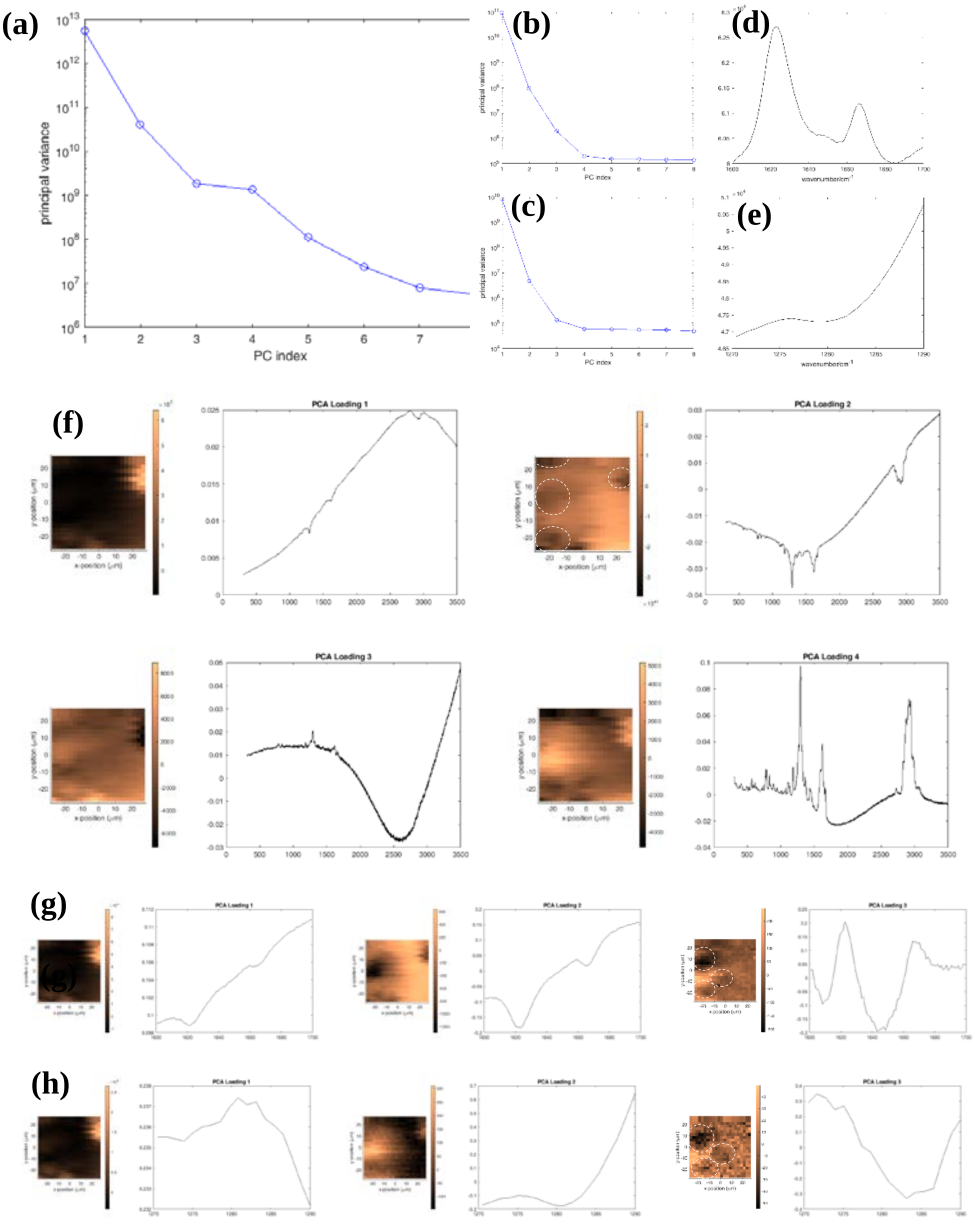

Figure 4. PCA analysis of trichome sample 2: (a) Screeplot of the Principal Components in the logarithmic scale on the y axis: variance of the dataset as a function of the PC index(s), (b) Screeplot of the filtered dataset in the spectral range 1600-1700 cm^{-1}, (c) Screeplot of the filtered dataset in the spectral range 1270-1290 cm^{-1}; (d) average spectrum in the region between 1600-1700 cm^{-1}, (e) average spectrum in the region between 1270-1290 cm^{-1}; (f) Scoremaps and loadings of PC1, PC2, PC3 and PC4; (g) Scoremaps and loadings of PC1, PC2 and PC3 of the filtered dataset in the spectra range 1600-1700 cm^{-1}; and (h) Scoremaps and loadings of PC1, PC2 and PC3 of the filtered dataset in the spectral range 1270-1290 cm^{-1}.