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

Janani Balasubramanian[1], Elisa Crocioni[2], Mattia Frattini[2], Scott Hill[1], Darryen Sands[1], 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 a novel coding approach through a Matlab interface 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 generated PC score curves from the Raman spectrum. 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. 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 spectrum which could be extended to handle all kinds of datasets with simultaneous spatial and chemical details.

Keywords: OpenPCA; Raman mapping; Cannabinoids; Trichomes.

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]. These cannabinoids are secreted from trichome structures which are the specialized hairs covered on female inflorescences. Three kinds of trichomes were observed with different morphologies, namely capitate-stalked, capitate-sessile, and bulbous [3]. A basal cell, secretory cells, several stalk cells, and large sub-cuticular storage cavity together constitute the abundantly present capitate-stalk of trichome that contains highest cannabinoid levels [2]. Phytocannabinoids, a unique group of terpenophenolics with three leading cannabinoids tetrahydrocannabinol (THC), cannabinol (CBN) and cannabidiol (CBD) are naturally synthesized in the trichomes. Δ^9-THC is considered as an illicit drug as it possesses psychotropic effect and hence 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 on the spatial distribution of the chemical species in the plant [7, 8]. However, the high spatial resolution of confocal Raman micro-spectroscopy, as proved for

instance in ref. [9] , may provide a way to obtain the chemical distribution of the cannabinoid substances at the micrometric scale, by probing directly samples of Cannabis, such as the trichomes. 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. [7, 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 applied in several fields. 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 i.e. Δ^9-THC in hemp is ≤ 0.2 % and marijuana 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 that comprises 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, the image generated with relation to the scanned area and

intensity is devoid of interference. The drawback behind carrying out this method in regular software includes the limited information associated with produced image that corresponds only to the selected wavenumber, low signal-noise ratio and loss of signal. Key information encompassed within Raman spectrum matrix is usually widely distributed throughout the dataset [13]. Principal component analysis (PCA) is an effective statistical method to handle a complex large data matrix by reducing the dimensionality and still preserving the most critical features [14]. However, PCA being a modern data analysis tool remain as a black box that is widely used but poorly understood. In this work, a novel coding approach is presented to introduce PCA in Matlab where the background process is observable, open and 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 analysis of a novel dataset of *C. sativa* trichomes by the implementation of new white-box approach. This work demonstrates 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 its distribution. Nevertheless, this technique is not limited only to *Cannabis*, but could be extended widely 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, 10 mg of pure CBD was directly used without any solvent to carry out 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 obtain the Raman spectra. The area of the trichome scanned was over a grid of 21 x 21 points, with 1 μm spacing. Here, each single point Raman analysis had a duration of 10 sec with 1 accumulation (referred as sample 1) and 10 accumulations each (referred as 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, χ_{ov}:

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

Where $\langle \boldsymbol{X}_{ov} \rangle$ 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:

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

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 ($\mathbf{\Sigma}_{vv}$) can be introduced as follows:

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

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) $\mathbf{\sigma}_{ss}$ is the diagonal matrix of the eigenvalues of $\mathbf{\Sigma}_{vv}$ and $\mathbf{L}_{vs}$ is the orthogonal matrix of the eigenvectors of $\mathbf{\Sigma}_{vv}$. The orthogonality of $\mathbf{L}_{vs}$ 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 $\mathbf{L}_{sv}$, 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 $\mathbf{\Sigma}_{vv} = \boldsymbol{\chi}_{vo}\boldsymbol{\chi}_{ov}$ / (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 (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 analysed area does not belong to Δ^9-THC, therefore the obtained Raman data set was processed by PCA to get 2D Raman mapping image. Cluster analysis is 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 is 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 literature 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 [7, 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] represent the similar strong peaks close to 1002, 1087 and 1605 cm^{-1}. These vibrations of most enhanced peaks at 1002 and 1087 cm^{-1} belong to the benzene ring and the alkyl chain of this drug. The band observed in the range of 1600 to 1670 cm^{-1} belong to the Benzene ring stretch (sym) and oxygen atoms of the hydroxyl group [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 physical states of the samples, the different crystalline forms 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} depicts CH deformation, 1032 cm^{-1} depicts C-C stretching, 1437 cm^{-1} depicts CH_3 twist and bend [21].

The CBD sample is a pure crystalline powder and provides a very neat FT-Raman spectrum with sharp and well-defined peaks. Out of the whole spectra, speaks at 1433 and 1662 cm^{-1} in the fingerprint region and peak at 2927 cm^{-1} in the high frequency region is predominant and could be recognized clearly. 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 the most similar chemical structures as they are distinguished by two benzene rings with different number of hydroxyl group (OH). Compared to Δ^9-THC, CBD has a strong peak at 1433 cm^{-1} which can be used to distinguish this analog 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* which are the site of production of metabolites. In this study, we examine the chemical composition and spatial distribution of one such important metabolite, cannabinoids at the microscopic scale using micro-Raman spectroscopy. Here, we aim to differentiate the structure of secretary vesicles in the whole trichome using Raman spectroscopy based on the fact of variation in the levels of cannabinoids. According to Livingston et al. [6] some regions 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 a trichome sample 1 and 2 is depicted in Figure 2a and d. The Raman chemical map generated with the most intense peak of the Raman spectra of trichome sample 1 and 2 i.e. 1295 cm^{-1} is depicted in Figure 2b and e. This Raman map image demonstrates the information obtained from direct conversion at a particular wavenumber without any advanced dataset processing analysis. Average spectrum obtained from Raman spectroscopy of several points of a flower trichome sample 1 and 2 is represented in Figure 2 c and f.

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

The PCA of the Raman spectra of the map was performed in order to visualize similarities between spectra and to identify 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, screeplot in the logarithmic scale is obtained to inspect the relevant principal components (PCs) as shown in Figure 3a (screeplot of sample 1). The screeplot is a representation of total variance of multivariate dataset by denoting the principal variance of PCs (eigenvalues) in decreasing order with regard to PC index. It is noticed that how quickly the intensity of the principal components decreases along the screeplot. For this reason, 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 the 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 depicted in Fig. 3f. The loadings of PC 1 to PC3 conveys some relevant chemical information, whereas, PC4 has mostly an undulatory behaviour with unclear peaks that is attributed to noisy signal. The PC1 loadings PC1 clearly display an intense fluorescence background, which is almost missing in the loadings of PC2. 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 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 that are more Raman active. Since the fluorescence signal is less strongly related to the chemical structure of the compound than the Raman signal, one may expect to get chemical information out of PC2. The corresponding scoremaps of PC1 to PC4 was obtained from Matlab PCA scripts. In the scoremap, each single measurement point in the Raman mapping experiment is represented in separate pixel along with different shades of the color to identify the value of the score. The lack of homogeneity in scoremaps is a representation of variation in how much of the Raman spectrum associated with each point. This indicates the local changes in levels of cannabinoids in the trichome sample. In the first scoremap (Fig. 3f), we could observe three dark spots (consider negative loading). The corresponding loading of PC1 displays same trend and super imposes perfectly with the average spectrum. Hence, this is regarded as the overall strength of the Raman signal. To be more precise, this indicates the overall variation that arise from the different point to point focusing on the curved bulb of the trichome surface without any chemical details. In the second scoremap, we can identify central dark region. However, since it is a complex data set with the combination of several chemical species, it is difficult to interpret the structure more precisely. Therefore, using the PCA scripts in Matlab, the range of wavenumber is selected from 1580-1700 cm^{-1} that contains a characteristic peak of cannabinoids observed in both THC and CBD. The Fig. 3g represents scoremap and loadings of PC1 and PC2 at specific 1580-1700 cm^{-1} wavenumber range. As mentioned earlier, the scoremap of PC1 does not include any specific chemical information as the loading looks similar to the average spectra (Fig. 3d). This may be associated with the florescence background over the Raman spectra. Noticing the scoremap of PC2, three dark

circular spots (consider negative loading) are observed that indicates high concentration of cannabinoids, while the bright regions corresponds to low concentration. The dimensions of the bright spots is about 8-12 µm. This is compared with the dimension of the vesicles in the *C. Sativa* trichome. In general, THC is accumulated in this specific region called vesicles. The dimension of trichomes of about 65 µm was analysed and assuming it contains an average 8 vesicles, the dimension of the vesicle is deduced between 6 and 10 µm. The length of the dark region in scoremap 2 is compatible with the size of a vesicle. 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. The PC1 is ignored owing to the same reason. Based on the observation of PC2, we can observe the maximum and minimum spectral change with respect to average spectrum follows a pattern where the peak get more narrow and intense. The corresponding scoremap displays the three bright circular regions (consider positive loading) that confirms the accumulation of substances in the trichome that is attributed to secretary vesicles.

The screeplot of trichome sample 2 from PCA analysis of the overall spectra is depicted in Fig. 4a. PC1 to PC4 loadings and scoremap of Raman analysis of the sample 2 is given in Fig. 4f. The first scoremap (Fig. 4f), we could observe entire black region. The corresponding loading of PC1 displays its due to the fluorescence background of the Raman signal. In the second scoremap, we can identify few circular dark spots (consider negative loadings) that could depict the overall presence of chemical species. To extract the precise details belonging to the specific chemical species, the range of wavenumber that contains a characteristic peak of cannabinoids at range 1600-1700 cm^{-1} was analysed. 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. Although PCA is useful to visualize regions with different

chemical composition, a detailed interpretation of chemical information brought by the PCs is not always successful since the vibrational modes of the chemical compounds are not always represented in the PCs loadings maintaining the same profile of the chemical Raman spectra, for example, negative peaks can appear since the shape of the bands are the results of the mathematical elaboration and do not reflect the exact vibrational spectral bands [17]. As mentioned earlier, the scoremap of PC1 does not include any specific chemical information and may be associated with the florescence background over the Raman spectra. Noticing the scoremap of PC2, a large dark spot was observed. Since the dimensions of the observed dark area is 20 µm, this would be ideally compared with the trichome structure where the presence of chemical species make it Raman active. The PC3, the maximum and minimum difference with respective to average spectrum is devoid of any fluorescence. We can observe three dark spots with area of around 8-10 µm that fitted well with vesicles size. The same pattern was observed for the PCA analysis of the region with high intense peak 1270-1290 cm^{-1} (Fig. 4c, e, and h). The reproducible pattern with different spectral range 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.

References

[1] M. Hesami, M. Pepe, M. Alizadeh, A. Rakei, A. Baiton, A.M.P.J.I.C. Jones, Products, Recent advances in cannabis biotechnology, 158 (2020) 113026.

[2] P. Rodziewicz, S. Loroch, Ł. Marczak, A. Sickmann, O.J.P.S. Kayser, Cannabinoid synthases and osmoprotective metabolites accumulate in the exudates of Cannabis sativa L. glandular trichomes, 284 (2019) 108-116.

[3] P.L. Carretero, A. Pekas, L. Stubsgaard, G.S. Blanco, H. Lütken, L.J.B.C. Sigsgaard, Glandular trichomes affect mobility and predatory behavior of two aphid predators on medicinal cannabis, 170 (2022) 104932.

[4] Y. Liu, H.-Y. Liu, S.-H. Li, W. Ma, D.-T. Wu, H.-B. Li, A.-P. Xiao, L.-L. Liu, F. Zhu, R.-Y.J.T.T.i.A.C. Gan, Cannabis sativa bioactive compounds and their extraction, separation, purification, and identification technologies: An updated review, (2022) 116554.

[5] G. Micalizzi, F. Vento, F. Alibrando, D. Donnarumma, P. Dugo, L.J.J.o.C.A. Mondello, Cannabis Sativa L.: A comprehensive review on the analytical methodologies for cannabinoids and terpenes characterization, 1637 (2021) 461864.

[6] S.J. Livingston, T.D. Quilichini, J.K. Booth, D.C. Wong, K.H. Rensing, J. Laflamme-Yonkman, S.D. Castellarin, J. Bohlmann, J.E. Page, A.L.J.T.P.J. Samuels, Cannabis glandular trichomes alter morphology and metabolite content during flower maturation, 101 (2020) 37-56.

[7] L. Sanchez, D. Baltensperger, D.J.A.c. Kurouski, Raman-based differentiation of hemp, Cannabidiol-rich hemp, and Cannabis, 92 (2020) 7733-7737.

[8] L. Ramos-Guerrero, G. Montalvo, M. Cosmi, C. García-Ruiz, F.E.J.T. Ortega-Ojeda, Classification of Various Marijuana Varieties by Raman Microscopy and Chemometrics, 10 (2022) 115.

[9] A. Badou, S. Pont, S. Auzoux-Bordenave, M. Lebreton, J.-F.J.J.o.S.B. Bardeau, 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, 214 (2022) 107854.

[10] L.-L. Tay, J. Hulse, R.J.C.J.o.C. Paroli, FTIR and Raman Spectroscopic Characterization of Cannabinoids, (2022).

[11] E. von der Esch, A.J. Kohles, P.M. Anger, R. Hoppe, R. Niessner, M. Elsner, N.P.J.P.o. Ivleva, TUM-ParticleTyper: A detection and quantification tool for automated analysis of (Microplastic) particles and fibers, 15 (2020) e0234766.

[12] Z. Sobhani, X. Zhang, C. Gibson, R. Naidu, M. Megharaj, C.J.W.r. Fang, Identification and visualisation of microplastics/nanoplastics by Raman imaging (i): Down to 100 nm, 174 (2020) 115658.

[13] C. Fang, Y. Luo, X. Zhang, H. Zhang, A. Nolan, R.J.C. Naidu, Identification and visualisation of microplastics via PCA to decode Raman spectrum matrix towards imaging, 286 (2022) 131736.

[14] J.E. Halstead, J.A. Smith, E.A. Carter, P.A. Lay, E.L.J.E.P. Johnston, Assessment tools for microplastics and natural fibres ingested by fish in an urbanised estuary, 234 (2018) 552-561.

[15] H.J.J.o.e.p. Hotelling, Analysis of a complex of statistical variables into principal components, 24 (1933) 417.

[16] J.R. Schott, Matrix analysis for statistics, John Wiley & Sons, 2016.

[17] S.J.L.V.M.P.D. Fedchak, USA, Presumptive field testing using portable raman spectroscopy, (2014) 1-59.

[18] S.K. Islam, Y.P. Cheng, R.L. Birke, M.V. Cañamares, C. Muehlethaler, J.R.J.C.P. Lombardi, An analysis of tetrahydrocannabinol (THC) and its analogs using surface enhanced Raman Scattering (SERS), 536 (2020) 110812.

[19] J. Leonard, A. Haddad, O. Green, R.L. Birke, T. Kubic, A. Kocak, J.R.J.J.o.R.S. Lombardi, SERS, Raman, and DFT analyses of fentanyl and carfentanil: Toward detection of trace samples, 48 (2017) 1323-1329.

[20] M. Wahadoszamen, A. Rahaman, N.M. Hoque, A. I Talukder, K.M. Abedin, A.J.J.o.S. Haider, Laser Raman spectroscopy with different excitation sources and extension to surface enhanced Raman spectroscopy, 2015 (2015).

[21] S. Yüksel, A.M. Schwenke, G. Soliveri, S. Ardizzone, K. Weber, D. Cialla-May, S. Hoeppener, U.S. Schubert, J.J.A.C.A. Popp, Trace detection of tetrahydrocannabinol (THC) with a SERS-based capillary platform prepared by the in situ microwave synthesis of AgNPs, 939 (2016) 93-100.

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

Figures

Figure 1

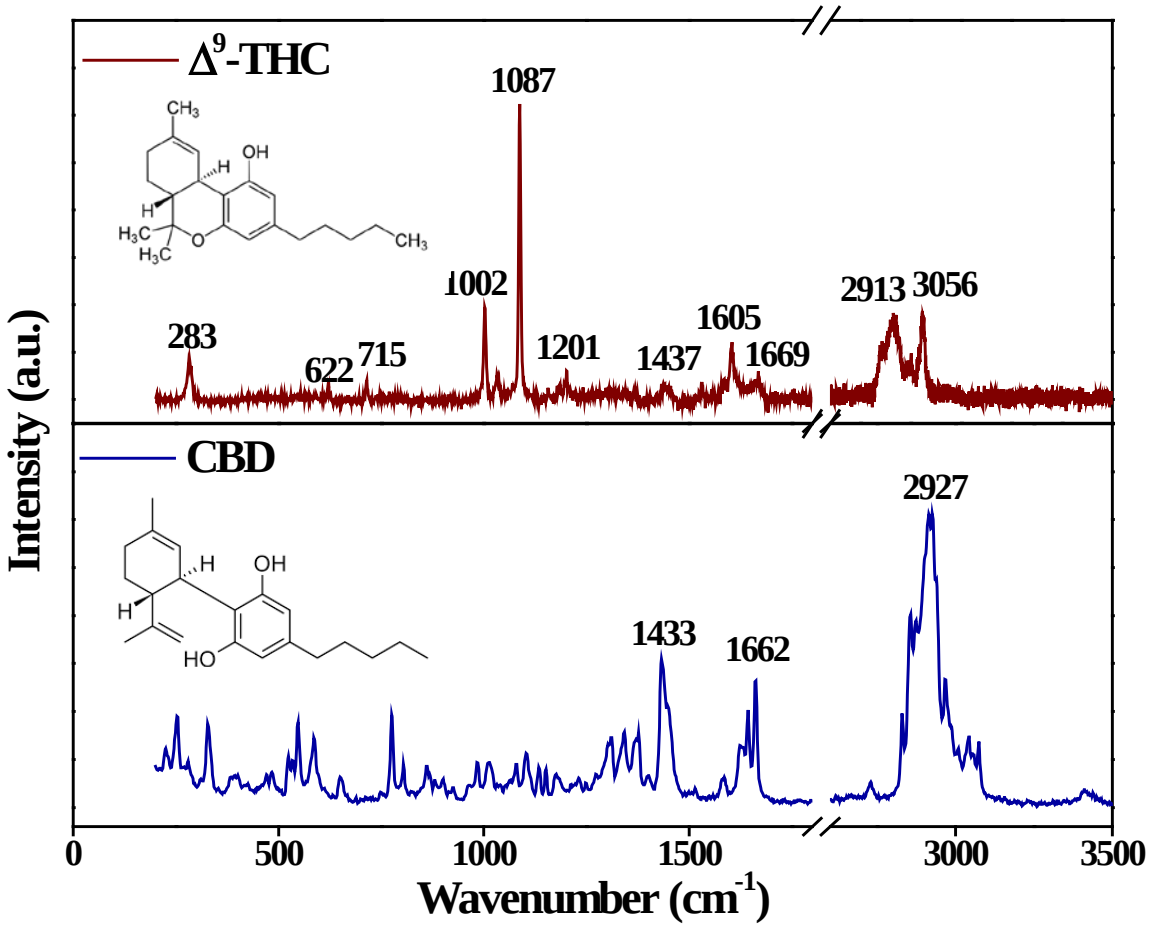

Figure 1. Pure cannabinoid spectra: the Raman spectrum of the ephemeral crystal of Δ9-tetrahydrocannabinol and the Raman spectrum of the pure cannabidiol; (inset - chemical structure of Δ9-THC and CBD).

Figure 2

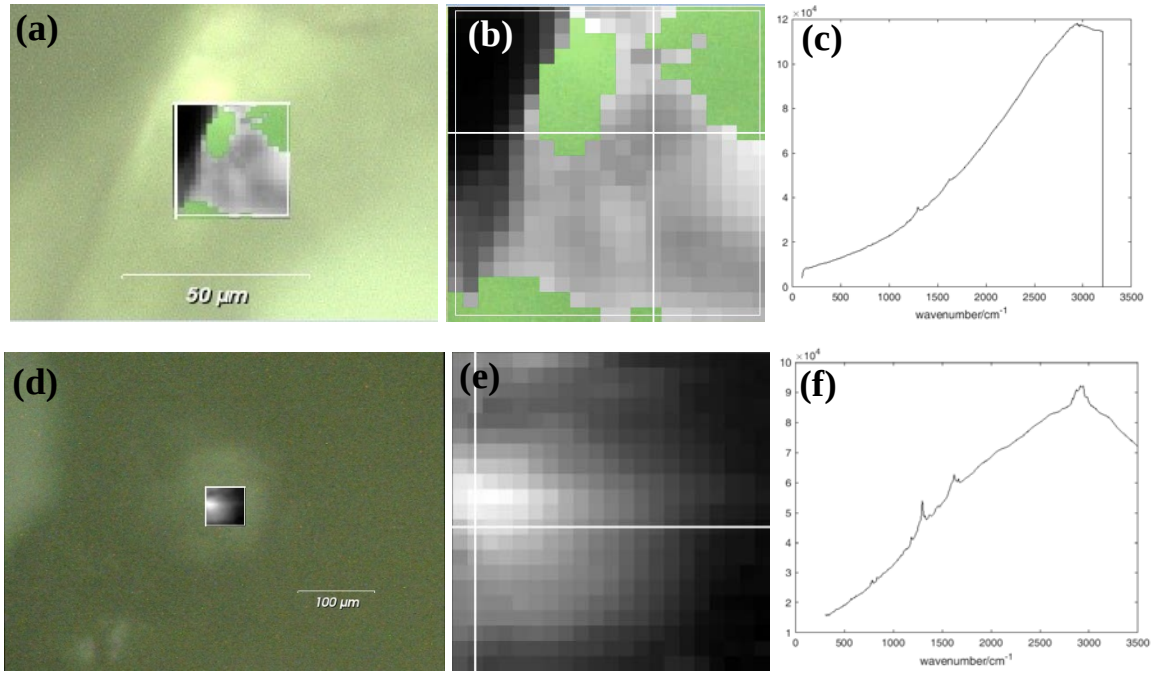

Figure 2. Trichome sample 1: (a) Bright field image of a trichome, (b) Raman chemical map of the Raman intensity at 1295 cm^{-1} and (c) average spectrum obtained from Raman spectroscopy of several points of a flower trichome. Trichome sample 2: (d) Bright field image of a trichome, (e) Raman chemical map of the Raman intensity at 1295 cm^{-1} and (f) average spectrum obtained from Raman spectroscopy of several points of a flower trichome.

Figure 3

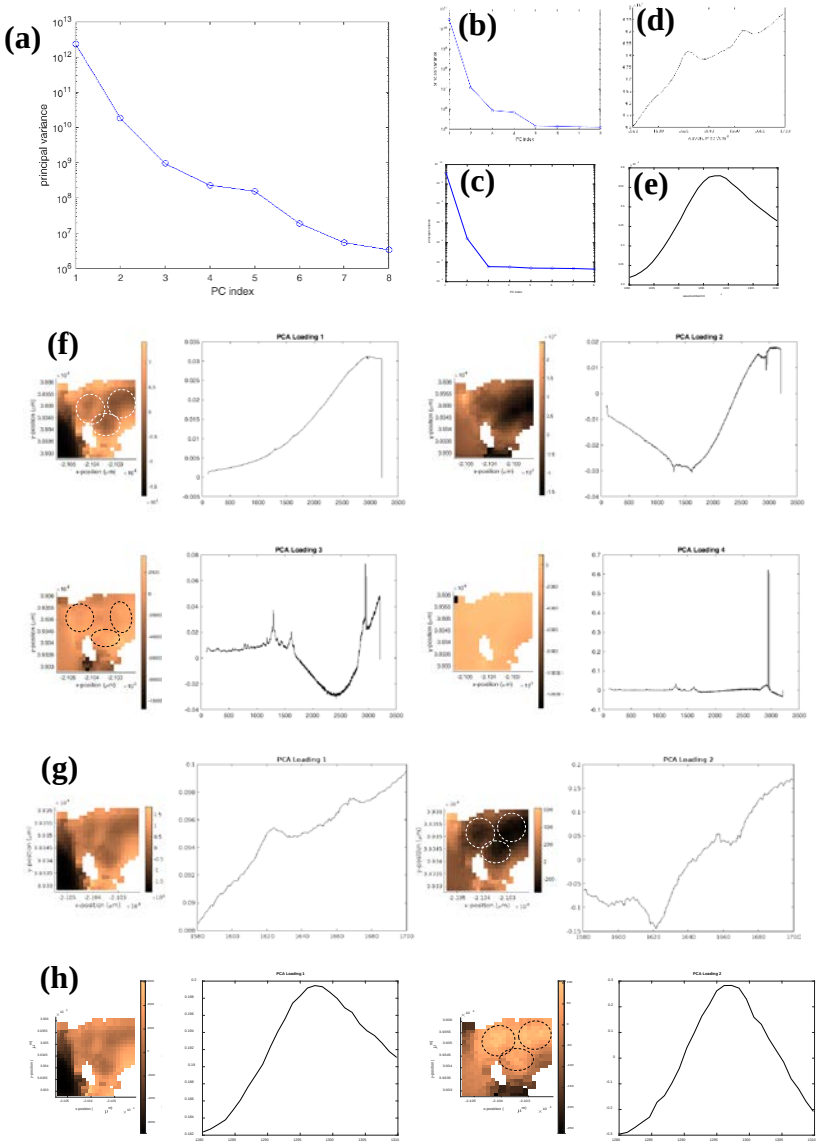

Figure 3. PCA analysis of trichome sample 1: (a) Screeplot of the Principal Components in the logarithmic scale on the y axis: variance of the dataset as a function of the PC index(s), (b) Screeplot of the filtered dataset in the spectra range 1580-1700 cm^{-1}, (c) Screeplot of the filtered dataset in the spectra 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 spectra range 1580-1700 cm^{-1} and (h) Scoremaps and loadings of PC1 and PC2 of the filtered dataset in the spectra range 1280-1310 cm^{-1}.

Figure 4

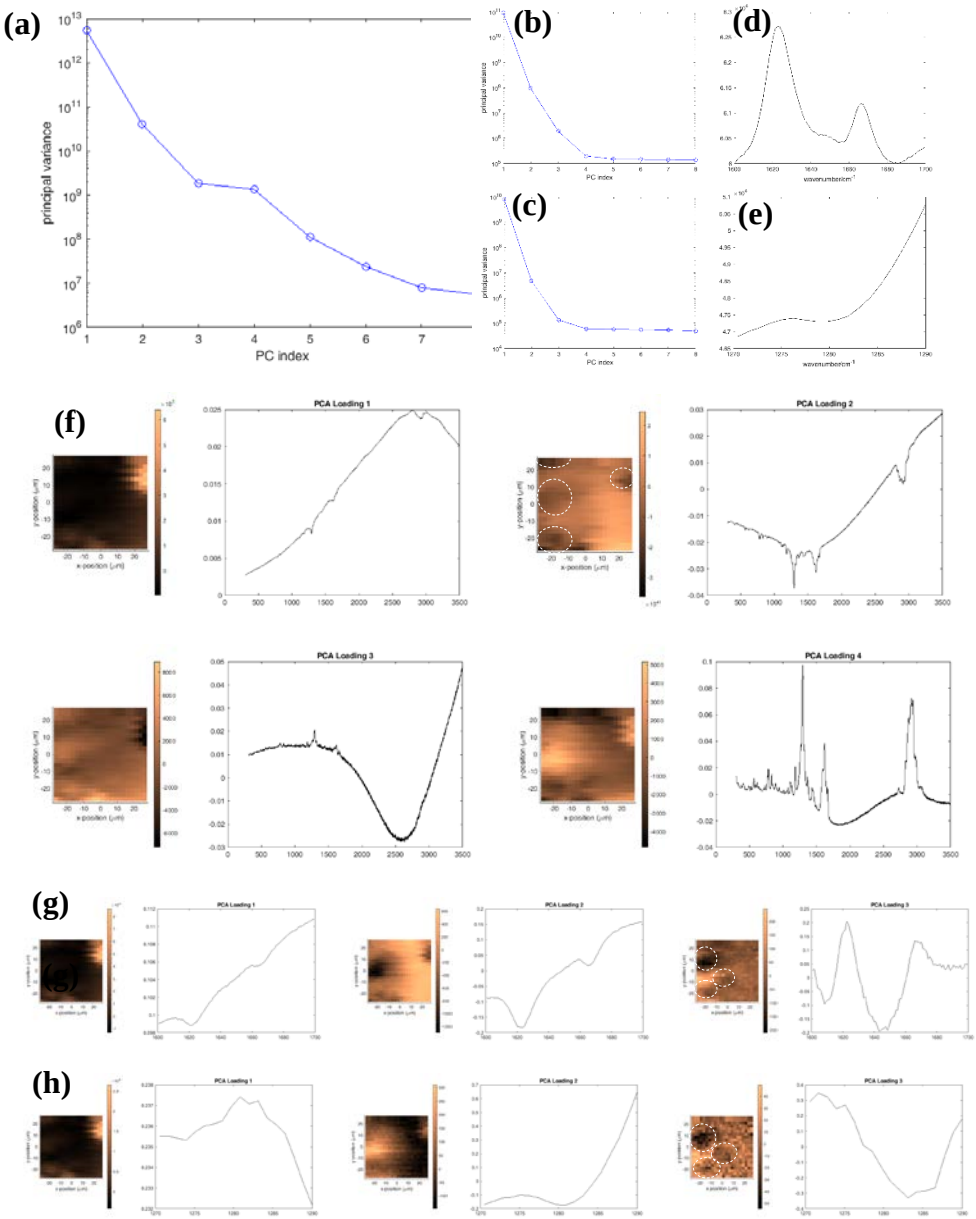

Figure 4. PCA analysis of trichome sample 2: (a) Screeplot of the Principal Components in the logarithmic scale on the y axis: variance of the dataset as a function of the PC index(s), (b) Screeplot of the filtered dataset in the spectra range 1600-1700 cm^{-1}, (c) Screeplot of the filtered dataset in the spectra 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 spectra range 1270-1290 cm^{-1}.