Gautam et al. EPJ Techniques and Instrumentation (2015) 2:8
DOI 10.1140/epjti/s40485-015-0018-6

EPJ Techniques and Instrumentation
a SpringerOpen Journal

REVIEW Open Access

Review of multidimensional data processing approaches for Raman and infrared spectroscopy

Rekha Gautam[1], Sandeep Vanga[1], Freek Ariese[1,2] and Siva Umapathy[1,3*]

* Correspondence:
umapathy@ipc.iisc.ernet.in
[1]Department of Inorganic and Physical Chemistry, Indian Institute of Science, Bangalore 560012, India
[3]Department of Instrumentation and Applied Physics, Indian Institute of Science, Bangalore 560012, India
Full list of author information is available at the end of the article

Abstract

Raman and Infrared (IR) spectroscopies provide information about the structure, functional groups and environment of the molecules in the sample. In combination with a microscope, these techniques can also be used to study molecular distributions in heterogeneous samples. Over the past few decades Raman and IR microspectroscopy based techniques have been extensively used to understand fundamental biology and responses of living systems under diverse physiological and pathological conditions. The spectra from biological systems are complex and diverse, owing to their heterogeneous nature consisting of bio-molecules such as proteins, lipids, nucleic acids, carbohydrates *etc.* Sometimes minor differences may contain critical information. Therefore, interpretation of the results obtained from Raman and IR spectroscopy is difficult and to overcome these intricacies and for deeper insight we need to employ various data mining methods. These methods must be suitable for handling large multidimensional data sets and for exploring the complete spectral information simultaneously. The effective implementation of these multivariate data analysis methods requires the pretreatment of data. The preprocessing of raw data helps in the elimination of noise (unwanted signals) and the enhancement of discriminating features. This review provides an outline of the state-of-the-art data processing tools for multivariate analysis and the various preprocessing methods that are widely used in Raman and IR spectroscopy including imaging for better qualitative and quantitative analysis of biological samples.

Keywords: Preprocessing; Baseline removal; Principal component analysis; Linear discriminant analysis; Classification models; Clustering; Partial least squares; Cross validation; Receiver operating characteristic

Introduction

Raman and IR spectroscopies provide detailed chemical information and are routinely used in various application areas including pharmaceutical, polymers, forensic, environmental, food sciences *etc.* [1–8]. Often the identification and quantification of components in biological samples by spectroscopic methods alone is hindered because of the sample's diverse nature. The spectra from heterogeneous bio-systems such as cells, tissues, biofluids *etc.*, consisting of a large number of bio-molecules, are complex. Partly, this is also due to the development of more sophisticated instruments which can provide high-resolution data and data from samples with their native matrices containing many

interfering substances. In addition to that, the differences from one sample to another in different pathological conditions are very small and difficult to observe in raw spectra. Therefore, to obtain meaningful information and for deeper insight we need to process and analyze the data. The data analytical methods that deal with only one variable at a time are called univariate methods. Univariate data analysis methods such as first and second order derivatives, curve fitting, difference (*e.g.* diseased minus normal) spectral analysis, and various bands intensity/area under the curve ratios facilitate the visualization of band shifts, peak broadening, change in intensities *etc.* [2, 9–11]. However, none of the spectroscopic measurements depend on a single variable. Spectroscopic data consists of thousands of variables (wavenumbers) and measurements (objects/observations). To utilize the complete information of the complex spectra and to handle the large data set, multivariate analysis is needed. Multivariate data analysis refers to data analytical methods that deal with more than one variable at a time. The main aim of these statistical analysis techniques is to perceive the relationship between the variables. This is based on the idea of considering many non-selective variables instead of just one variable and then ultimately combining them in a multivariate model. The application of multivariate statistical methods to chemistry and biology is also called Chemometrics [12]. Multivariate analysis tools are used for the efficient processing of huge datasets and to align their informative features [13, 14]. It helps in data analysis, especially in cases where large amounts of data are generated, like in NMR, FTIR, Raman, and GC-MS [12–14]. Using multivariate analytical tools, patterns in the data could be modeled and these models can be used routinely to predict the newly acquired data of a similar type. Various data mining methods such as principal component analysis (PCA), linear discriminant analysis (LDA), multiple linear regression (MLR), cluster analysis (CA) and partial least squares regression (PLS) to name a few are employed in the field of Raman and IR spectroscopy [14–16]. The multivariate statistical methods are very useful for processing of Raman and IR spectral data because of their ability to analyze the vast spectral distribution and thoroughly discriminate between spectra of different samples that show only very minor changes [15, 16]. The effective implementation of these chemometric methods requires the pretreatment of data. The preprocessing of raw data helps to eliminate noise (unwanted signals) and to enhance requisite signals such as discriminating features. Sometimes, chemometrics itself assists in data pre-processing, to reduce and correct for interferences such as overlapping bands, baseline drifts, scattering and mainly to analyze spectral variations [17, 18]. Subsequent sections in this paper discuss some of the basic definitions, provide an overview of the various preprocessing methods and state-of-the-art data processing tools for multivariate analysis that are widely used in Raman and IR spectroscopy.

Basic concepts and definitions

The basic definitions pertaining to statistical analysis are important when dealing with more complex multivariate data structures [19–21]. Some of them are listed below.

Let's assume variable 'V' (whose values can be intensities at a given wavenumber over the different observations or spectra in the data set) as defined below:

$$V = (\nu_1, \nu_2, ..., \nu_M)$$

where M is the number of observations (spectra)

Mean **(*μ*)**: the average intensity value of the variable:

$$\mu = \frac{\sum_{i=1}^{M} v_i}{M}$$

Median: the middle value of the variable when the data is aligned either in increasing or decreasing order.

Mode: The intensity that occurs most frequently among the values of the variable.

Standard Deviation **(*SD*)**: measure of the spread of the intensities of the variable:

$$SD = \sqrt{\frac{\sum_{i=1}^{M} (v_i - \mu)^2}{M-1}}$$

Variance: the square of the standard deviation, which is another measure for the spread of the intensities of the variable:

$$Variance = SD^2 = \frac{\sum_{i=1}^{M} (v_i - \mu)^2}{M-1}$$

Covariance **(*cov*)**: a measure of the linear association between two variables (*e.g.* at different wavenumbers); say U and V. Covariance can be positive as well as negative. A large absolute value of covariance means that there is a strong linear dependence between the two variables and vice versa. A large positive value of covariance indicates that the values of both variables are either increasing or decreasing together. The covariance is negative if the values of both variables are moving in opposite directions. Close to zero covariance means that the two variables do not show any pattern *i.e.* independent of each other. As there could be many such variables in a data set, a covariance matrix can be obtained by calculating the covariance between all pairs of variables:

$$cov(U, V) = \frac{\sum_{i=1}^{M} (u_i - \mu_u)\,(v_i - \mu_v)}{M-1}$$

Correlation **(*r*)**: this is a more practical measure to compare the linear dependencies of mixed variables (variables with different units or scales). In such cases covariance fails to depict the real picture. Correlation (also called Pearson's correlation coefficient) is a unitless, scaled covariance measure. Pearson's correlation coefficient 'r' is defined below:

$$r = \frac{\sum_{i=1}^{M} (u_i - \mu_u)\,(v_i - \mu_v)}{\sqrt{\sum_{i=1}^{M} (u_i - \mu_u)^2}\sqrt{\sum_{i=1}^{M} (v_i - \mu_v)^2}}$$

Correlation = 0, means there is no correlation between the variables.

Correlation = +1, means there is an exactly linear positive correlation between the variables.

Correlation = −1, means there is an exactly linear negative correlation between the variables.

'r^2' is the most common measure of the fraction of the total variance that can be modeled by this linear association measure. However, most of the times several different variables contribute simultaneously, which requires a multivariate modeling of the property (outcome being modeled through multivariate analysis). A correlation between the property and variable close to unity indicates that the property depends mostly on one variable called the 'selective variable'. Multivariate data analysis usually deals with 'non-selective' variables, which means that several different variables contribute simultaneously. In that case a multivariate modeling of the property can help by means of dimension reduction.

Eigenvectors and Eigenvalues: An eigenvector is a special non-zero vector (say 'x') of a square matrix 'A'. Multiplying matrix 'A' with such a non-zero vector results in stretching/compression of the vector but the direction of the eigenvector remains constant. The eigenvalue of an eigenvector is the quantity (scalar) by which the original eigenvector scales after multiplication by the matrix 'A':

$$Ax = \lambda x$$

where λ is a nonzero scalar, also called eigenvalues

There could be multiple eigenvectors of a matrix 'A'. All of those eigenvectors are orthogonal to each other, and therefore linearly independent, if the square matrix 'A' is a real-valued - symmetric matrix. Eigenvectors of the covariance matrix of the data set are extremely important as they can represent underlying correlation patterns compactly. It is important to note that the covariance matrix is a real-valued symmetric square matrix. Later in this review it will be discussed in detail along with dimensionality reduction.

Distance Metrics: In many multivariate algorithms the distance between observations (spectra) is an important part in defining the objective function of the algorithm. Some of the commonly used distance metrics are mentioned below. Let us say "A" and "B" are two spectra with intensities (A_1, A_2, A_3 … A_N) and (B_1, B_2, B_3 … B_N) respectively.

$$Euclidean: \qquad dist = \sqrt{\sum_{i=1}^{N}(A_i - B_i)^2}$$

N is the of number of variables (spectral data points)

$$Manhattan: \qquad dist = \sum_{i=1}^{N}|A_i - B_i|$$

$$Mahalanobis: \qquad dist = \sqrt{(A-B)^T S^{-1}(A-B)};$$

where S is the covariance matrix and (A-B) is the transpose of (A-B)

$$Minkowski: \qquad dist = \left(\sum_{i=1}^{N}|A_i - B_i|^p\right)^{\frac{1}{p}}; where\ p \geq 1$$

p= 1 gives the Manhattan distance & p = 2 gives the Euclidean distance

$$Cosine\ Similarity\text{: } dist = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} {A_i}^2}\sqrt{\sum_{i=1}^{N} {B_i}^2}}$$

Signal to Noise Ratio **(*SNR*)**: Signal to noise ratio quantifies the amount of desired signal relative to the noise. This metric is used to measure signal strength and detectability. In vibrational spectroscopy, the intensity at a particular wavenumber is used as indication for "Signal". The standard deviation of intensities at dead regions *i.e.* without any peak (varies with sample type) is generally considered as "Noise".

Review

Preprocessing

Data processing applied prior to univariate/multivariate analysis is known as preprocessing. Preprocessing is required to eliminate effects of unwanted signals such as fluorescence, Mie scattering, detector noise, calibration errors, cosmic rays, laser power fluctuations, signals from cell media or glass substrate *etc.* and also to enhance subtle differences between different samples [22, 23]. As Raman and IR spectroscopy are based on two different phenomena, the signal and background noise (unwanted signals) are also different and different pretreatment steps are required. As spectra of the same material could have been recorded over several days/months, it is very difficult to calibrate the Raman instrument precisely in order to have the same Raman shift axis. Also, different gratings provide different spectral resolutions. Therefore, spectra need to be aligned to a common axis before applying any pre-processing method. Apart from spectral alignment, Raman spectra should also be corrected for Cosmic ray events (CREs) before further pre-processing is applied. CREs are generated due to high-energy particles passing through the charge coupled device (CCD) and generating many electrons, which the CCD interprets as signal. These are totally random, appear as very sharp emission lines and usually affect only one pixel at a time.

In recent years, Vibrational Spectroscopy has been extensively used in the field of biology and medicine. The main challenge in using Raman spectroscopy (which is inherently a weak phenomenon) for biological samples is the strong intrinsic fluorescence from many biomolecules. The fluorescence background is often many times more intense than the weak Raman signals. Therefore to extract the informative Raman signals we need to process the raw spectrum to remove the fluorescence. Also ambient light and detector thermal noise may contribute to the background. Various processing methods such as polynomial fitting, first and second order differentiation, frequency domain filtering *etc.* have been used for this purpose. This problem is also tackled by involving hardware which includes the use of longer excitation (NIR) wavelengths, time gating, wavelength shifting *etc.* [22, 24, 25]. Similarly, the mid-infrared spectrum from cells and/or tissues is hampered by Mie scattering as the cells and cell-organelles within the samples under investigation are of comparable size with the radiation wavelengths (2.5-25 μm) used. This contributes to a broad and undulating background to the FTIR spectrum which in turn gives rise to distorted band shapes, intensities and positions [26, 27]. This needs to be corrected before interpretation of the data and Extended Multiplicative Scattering Correction (EMSC) can be employed for this purpose. The Raman and FTIR spectra are also affected by detector noise and intensity fluctuations

of the radiation source used. The SNR can be improved by increasing the integration time or by using smoothing filters. The source/environment fluctuations add some variability to the spectra that is not related to the actual differences (chemical or structural) in the samples. Various normalization methods are used to surmount these variations. These normalization methods also overcome the variations in the FTIR spectra due to inconsistent sample thickness. As the sources of the above described contributions to the raw spectra are different in case of Raman and FTIR spectroscopy, the preprocessing methods used to overcome these issues are quite different and some of them are described below in detail.

Preprocessing in Raman spectroscopy

Spectral axis alignment

All the analysis techniques expect to have the same Raman shift axis across all the spectra. So, it is very important to align all spectra to have a common spectral axis. Different local regression methods can be used to calculate intensities at a pre-defined common spectral axis using the intensities at an existing spectral axis [28].

Cosmic ray/spike removal

Usually, the spike elimination from the raw spectrum is done by collecting two extra spectra for each experiment and by comparing them on a pixel by pixel basis. If the difference exceeds the expected detector noise variance of the less intense pixel then the greater count is replaced by the smaller count. Generally, spikes are sharper (lower FWHM) compared with genuine Raman bands. Although it is easy to detect such spikes based on thresholding on maximum intensity value, it is not straightforward to correct such spikes. Usually, local interpolation based methods are used to repair spike affected regions [29]. Particularly in Raman imaging, information from the Raman spectra corresponding to the unaffected adjacent pixels can also be used to correct spikes.

Background correction

Background correction/baseline removal is a very important part of preprocessing. Various phenomena explained in the previous section like fluorescence *etc.* induce uneven amplitude shifts across different wavenumbers (Raman shifts). These amplitude shifts have to be compensated before proceeding with further analysis. In the literature many such techniques have been compared and evaluated in detail [22, 30]. Some common methods employed for baseline removal of Raman spectra are discussed here:

a) Median Window based methods
   This is a moving window based method where at each point (Raman shift) only a few intensity values (length of the window) around the point are used for estimating the baseline value at that point. The median of such a local window of intensity values at each point is calculated first. This series of median values is convolved with a Gaussian function to make sure the estimated baseline is free from sharp discontinuities [31]. Although this method is model free (non-parametric), it is primitive for handling Raman signals.

b) Differentiation based methods
Generally, the baseline has broad bands and low frequency components compared with genuine Raman bands. Differentiation of the raw Raman signal amplifies higher frequency (sharp) components and lower frequency components such as background fluorescence are suppressed. However, the noise present in the raw Raman signal also contains very high frequency components and in turn they also get enhanced due to differentiation along with genuine Raman bands. To suppress these noisy components generally a smoothing operation is employed as the post-processing step following differentiation. In the literature many such methods exist, including Savitzky-Golay (SG) filter based derivatives [32] and kernel smoothing based derivatives [33].

c) Polynomial Fitting based methods
This is by far the most commonly used method for baseline removal of Raman spectra. In this method certain points in the spectrum are chosen as base points and a polynomial is fitted through these points. This polynomial is subsequently subtracted from the Raman spectrum to eliminate background effects. For a simple Raman signal like a Raman spectrum of a non-fluorescent solid compound, a straight line fit through a couple of points could be sufficient enough, whereas for a complex Raman signal like a Raman spectrum of tissues/cells one might need many base points along with a fifth-order or higher polynomial fit [24]. Selecting the appropriate polynomial order is extremely important as an incorrectly chosen higher order polynomial may estimate some important Raman bands as background. Also, higher order polynomial fitting may be affected by high frequency noise and hence the background estimates are inconsistent. Some modified multi-polynomial fitting based background correction techniques are proposed to handle these issues [34, 35].

d) Asymmetric Least Squares based methods
Here, a smoothed signal is estimated from a given raw Raman signal as baseline. The residual signal between the raw Raman signal and the estimated baseline is the corrected Raman signal. This can be achieved by the ordinary least squares method. In ordinary least squares an objective function, defined as the sum of the squared difference between the raw Raman signal and the baseline to be estimated, is minimized iteratively. In ordinary least squares equal priority is given to the negative and positive residual errors. However, as a baseline should always be yielding positive residual errors to make sure all the important Raman bands are intact, ordinary least squares can be altered to have a bias towards positive residual errors. This method is called Asymmetric Least Squares (ALS) baseline correction [36]. Here, we can add a second-order derivative as another term to the objective function to make sure the estimated baseline is as smooth as possible. There are many simple algorithms like gradient descent which estimate such a baseline that minimizes the objective function. This method is relatively fast even for complex signals, requires fewer parameters and turns out to be effective for Raman spectra.

e) Frequency Domain Analysis based methods
As mentioned earlier, the baseline is defined as a component of the raw Raman signal having broad bands. The baseline varies at much lower rates compared with genuine Raman bands. In other words, the baseline is the low frequency component of the raw Raman spectrum and the Raman bands are much narrower. Similar to

differentiation, frequency domain based methods try to exploit these properties of the baseline and Raman bands to separate them out from each other. Here, we consider Raman spectra as a time series of frequencies (Raman Shifts) and the output from frequency domain analysis based methods contains information about the underlying variations (frequencies) in the time series of Raman Shifts. Broadly, we can categorize frequency domain based methods in to two categories as mentioned below.

FT based methods: Fourier Transform (FT) is widely used in conventional signal processing and telecommunications. FT uses sinusoids and cosinusoids as basis functions to extract frequency information present in the Raman signal. Each sinusoid or cosinusoid present in the basis function set represents a unique frequency. FT decomposes the Raman signal into linear combinations of such sinusoid or cosinusoid waves; their amplitude represents the contribution towards the Raman signal. Now, it is easy to threshold amplitudes corresponding to low frequency sinusoids or cosinusoids by replacing them with zero and thus nullify the baseline. The baseline corrected Raman signal can be reconstructed by applying Inverse Fourier Transform (IFT) on these modified amplitudes [25, 30]. Fast Fourier Transform (FFT) is an efficient algorithm which can be used to implement both FT and IFT.

Wavelet based method: Wavelet based denoising techniques are widely used in various fields including image processing, chemometrics *etc.* Wavelets are functions which are localized both in time or space as well as frequency. As cosinusoids or sinusoids are localized only in frequency, when FT is applied more terms are required to represent the same Raman signal compared with Wavelets. This is due to the fact that quite a large number of cosinusoids or sinusoids of increasing frequency have to be used to cancel out each other when applied to discrete signals like Raman signals which are defined only for a limited set of values. There are many Wavelet families available in the literature such as Mexican hat, Haar, Daubechies, Symmlet, triangular *etc.* and different Wavelet families have different mother Wavelets. For example in case of the Haar Wavelet family, a square function is the mother Wavelet. All the other Wavelets in the given Wavelet family are shifted and scaled versions of the corresponding mother Wavelet. Using these Wavelets as basis functions (Wavelet Transform), we can extract frequency-like information from the Raman signal. Here, the Raman signal is decomposed into different scales (multi resolution). Each scale (resolution) gives different frequency-related information contained in the Raman signal. As baseline (low frequency) and noise (high frequency) related frequencies are different compared with genuine Raman bands (mid frequency), at an optimum resolution appropriate thresholds can be applied to eliminate both baseline and noise simultaneously. After thresholding (removing) the baseline, the corrected Raman signal can be obtained by the Inverse Wavelet Transform [30, 37]. Moreover, these Wavelet based methods can be combined with polynomial and differentiation based methods to get superior results [38].

Smoothing

Baseline removal eliminates effects of broad bands or low frequency components present in the Raman spectra. However, the high frequency component of the Raman

signal, which typically has much lower FWHM compared with genuine Raman bands, needs to be removed too. Smoothing is often employed for the removal of high frequency components, and SG (Savitzky Golay) filtering is one of the commonly used smoothing techniques. The SG filter is a moving window based local polynomial fitting procedure [32], which needs to be fed with parameters like the size of moving window, polynomial order *etc.* As the moving window size increases, some of the genuine Raman bands with lower FWHM may disappear. Therefore, it is very important to choose an appropriate polynomial order and moving window size to retain all the important Raman bands.

Apart from the SG filtering technique, other local regression methods like LOWESS (Locally Weighted Scatterplot Smoothing) can be used for smoothing [39]. Other spatial smoothing techniques like Gaussian blurring can also be used for this purpose. In the discrete domain, these filters use predefined coefficients to convolute with Raman signals [23]. Again it is very important to understand that, as all of these methods are applied locally based on a moving window, underlying parameters have to be chosen carefully such that none of the important Raman bands are eliminated during smoothing.

Normalization

Normalization is a very important part of preprocessing, as different spectra of the same material may have been recorded at different times and under different instrument conditions such as alignment and laser power levels. So, spectra from the same material could have different intensity levels. Normalization is the process which takes care of disparity in intensity levels by making sure that the intensity of a given Raman band of the same material is as similar as possible across the spectra recorded under the same experimental parameters but slightly different conditions. There are numerous normalization techniques available in the literature [22, 40]. Based on various underlying factors and the problem to be solved, one particular normalization technique may be more suitable than others. Some of the commonly used normalization techniques and a brief description of each are given below.

Let us assume the spectrum to be normalized is defined as a vector 'S' and the normalized spectrum as a vector '*SN*' where

$$S = (s_1, s_2, ..., s_N)$$

N is the number of Raman shifts (spectral data points)
and

$$SN = (sn_1, sn_2, ..., sn_N)$$

Here, each element of the vector represents the Raman intensity at a given Raman shift.

(a) Vector normalization

In vector normalization, first of all the '*norm*' of the spectrum, which is defined as the square root of the sum of the squared intensities of the spectrum, is calculated. Further, each of the Raman intensities corresponding to a Raman shift is divided by the '*norm*' to obtain the normalized spectrum:

$$norm = \sqrt{s_1^2 + s_2^2 + ... + s_N^2}$$

$$sn_i = s_i/norm\ ; i = 1, 2, \ldots, N$$

(b) Min-max Normalization

Here, the *'maximum'* and *'minimum'* values of all the intensities of the given spectrum are calculated first. Then, each Raman intensity corresponding to a Raman shift is replaced by a new intensity obtained from subtracting *'minimum'* and dividing by *'range (maximum-minimum)'*:

$$s_{max} = \max(s_1, s_2, \ldots, s_N)$$

$$s_{min} = \min(s_1, s_2, \ldots, s_N)$$

$$sn_i = (s_i - s_{min})/(s_{max} - s_{min}); i = 1, 2, \ldots, N$$

(c) Standard Normal Variate (SNV) Normalization

This technique is similar to Min-max normalization except that instead of *'minimum'* *'mean'* and instead of *'range'* *'standard deviation (SD)'* is used:

$$mean = (s_1 + s_2 + \ldots + s_N)/N$$

$$SD = \sqrt{\left((s_1 - mean)^2 + (s_2 - mean)^2 + \ldots + (s_N - mean)^2\right)/(N-1)}$$

$$sn_i = (s_i - mean)/(SD); i = 1, 2, \ldots, N$$

(d) Peak Normalization

In peak normalization, the intensity corresponding to the central frequency of a particular Raman band is used as reference. Let's define it as *'P'*. Now, each Raman intensity of the spectrum is divided by *'P'* to obtain the normalized spectrum.

$$sn_i = s_i/P; i = 1, 2, \ldots, N$$

This method is not recommended when there is a possibility of a shift in the band position across the spectra from different samples under investigation. For example, if we want to compare native versus denatured protein samples which are known to cause a shift in the amide I and amide III regions of Raman spectra, the peak normalization with respect to those bands is not recommended.

Outlier removal

Due to factors like instrumental artifacts, variations in the sample *etc.*, some of the spectra from the same sample diverge from the group. These spectra can be considered as unwanted spectra or outliers. It is extremely important to omit these spectra before applying multivariate techniques to get desired results. In some cases it is possible to use some kind of Signal to Noise (SNR) based thresholding to detect such spectra. These SNR thresholding methods will be discussed in detail in FTIR preprocessing.

Also, one can apply thresholding methods in the compressed domain based on the SD of data in the compressed domain to eliminate such outliers [41]. If a given spectrum is very different from all the other spectra in the data set with respect to a particular variable(s) in the compressed domain *i.e.* the intensity value of a spectrum is less (or greater) than the defined threshold for a particular variable(s), then it can be considered as an outlier. This compressed domain is generally found by using factor

analysis methods such as Principal Component Analysis (PCA). PCA is discussed in detail in subsequent sections of this review.

In the compressed domain, mostly the dominant axes (variables) are used to judge the outliers. In the case of PCA, outliers can be found by using dominant axes such as PC1, PC2 and PC3 *etc.* Here, a simple thresholding algorithm is applied to the data set consisting of Raman spectra [10] using PC1 and PC2 to demonstrate the outlier removal method. After projecting the original data onto the compressed domain, score values with respect to PC1 and PC2 (PC1 and PC2 are new axes in the compressed domain) can be obtained. If a particular spectrum has a PC1 or PC2 score value greater (or lesser) than the corresponding mean by 2.58 (99 % confidence level) times the SD of PC1 or PC2 respectively, it can be marked as an outlier. Figure 1 is an illustration of the outliers through a scatter plot of score values of different principal components (PCs). Each point in the scatter plot represents a Raman spectrum, recorded from *Drosophila* muscles at 785 nm excitation wavelength, which is focused onto the muscles using a 50x (NA = 0.75) objective for an integration time of 150 s. Here, the

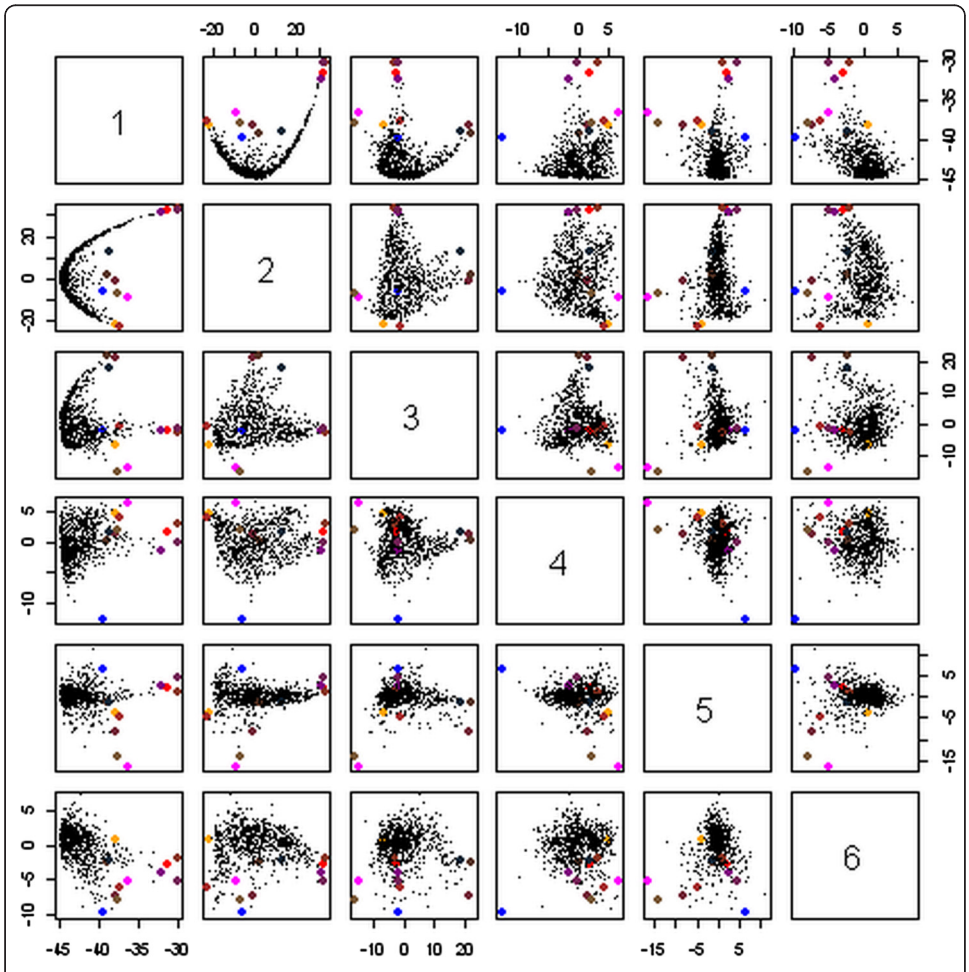

Fig. 1 Scatter Plots of score values of different Principal Components (PCs). Each point in the scatter plot represents a Raman spectrum, recorded from *Drosophila* muscles at 785 nm excitation wavelength. Score values corresponding to retained spectra are marked in black, and score values of outliers are marked in other colors. The first row indicates PC1 through PC6 and similarly the first column indicates PC1 through PC6. Score values of outliers are far apart from the other score values in each scatter plot

retained spectra are marked black in color and outliers are marked in colors other than black. The first row indicates PC1, the second row indicates PC2 and so on. Similarly, the first column indicates PC1 and so on. As we can clearly see from Fig. 1, the outlier score values are certainly far apart from the mean score values of each PC. Though PC1 and PC2 are used to determine the outliers, the score values of these outliers with respect to other PCs (PC3, PC4, PC5, and PC6) are also very different from their corresponding mean values as shown in Fig. 1. This indicates the effectiveness of using compressed domain techniques (particularly PCA) for finding and eliminating outliers.

Apart from the methods discussed above, particularly in the case of imaging data, spectra that are recorded outside the Region of Interest (ROI) are also considered as outliers and need to be removed. For example, in the case of single cell Raman imaging, spectra from the area devoid of cells (containing only buffer and substrate) need to be eliminated before proceeding with the analysis. Here, different clustering techniques can be used to segment out the cell region from the non-cell region [42, 43]. In the multivariate data analysis section such clustering methods are explained in detail.

Preprocessing in FTIR spectroscopy

Those methods that are suitable for both Raman and IR data and covered in previous section are not repeated here.

Background correction

Due to various instrumental and scattering effects, an actual FTIR spectrum gets superimposed on top of a background. Similar to Raman spectroscopy, the background in FTIR signals consists of broad bands (low frequency regions). These background correction methods could be as simple as subtracting an offset (DC shift) or removing a piecewise constructed baseline by selecting a few points and joining those points through straight lines [44]. Various complex background correction/baseline removal techniques explained earlier for Raman spectroscopy can also be used for FTIR spectra, but some of those techniques are more suitable than others. Such techniques include polynomial fitting and differentiation based on SG filters [32]. Generally, lower order polynomials (second or third order) are well suitable for FTIR spectroscopy. In fact, this technique can be combined with certain normalization techniques such as Multiplicative Scatter correction (MSC) to perform background correction and normalization simultaneously. More details about this technique are given below. As second order derivatives can efficiently remove the background present in FTIR spectra, SG-based second order differentiation techniques are very popular for background correction of FTIR spectra [45]. As the SG filter is a nonlinear weighted smoothing function, it makes sure that high frequency noise amplified during second order differentiation is well suppressed.

Normalization

Normalization is a very important part of preprocessing as it attempts to minimize the effects of source power fluctuations (MIR radiation source), scattering, variations in sample thickness *etc.* Normalization attempts to simultaneously correct for various shifting and scaling effects caused by the above mentioned phenomena. Some of the normalization methods used for Raman data are quite effective for FTIR spectroscopy too. Such techniques include Min-max normalization, SNV normalization *etc.* These

methods are applied individually to each spectrum, so these methods can be classified as 1-way methods [46]. Certain normalization methods considering more than one spectrum (say 'n' number of spectra) at a time while building the model can be classified as n-way methods. For example, techniques like multiplicative scatter correction (MSC) and extended MSC (EMSC) are very popular in FTIR spectroscopy. While MSC attempts to normalize FTIR spectra, EMSC also takes care of baseline removal along with normalization.

(a) Multiplicative Scatter Correction (MSC)

MSC tries to eliminate effects of amplification (multiplicative) and constant offset (additive). Correction coefficients of each spectrum are calculated by regressing it onto an ideal sample spectrum (a representative spectrum of the group of spectra under consideration in a completely noise free environment). In other words, each spectrum is fitted to the ideal sample spectrum (generally the average spectrum) as closely as possible using least squares [47]. Let us assume $S^1, S^2, \ldots, S^M$ are the FTIR spectra under consideration and S^μ is the average spectrum of the data set. Each spectrum is a vector of FTIR absorption values (intensities):

$$S = (s_1, s_2, \ldots, s_N)$$

N is the number of spectral data points

MSC tries to represent each spectrum in terms of average spectra by the following equation:

$$S^i = a^i + \left(b^i * S^\mu\right) + E^i \; ; \; i = 1, \; 2, \; \ldots, \; M$$

where E^i is the corresponding residual spectrum, which represents unique chemical information present in the spectrum *'i'* on top of the average spectrum S^μ. Once these parameters (scalar values) a^i, b^i are calculated for each spectrum by fitting the whole data onto the average spectrum in least squares sense, corrected spectra are obtained by the following equation:

$$S\_msc^i = \frac{\left(S^i - a^i\right)}{\left(b^i\right)} = \left(E^i / b^i\right) + S^\mu \; ; \; i = 1, \; 2, \; \ldots, \; M$$

where S_msc^i be the corrected spectrum of corresponding raw spectrum (S^i). As MSC is a set dependent transformation, it is sensible to apply MSC separately for different classes.

(b) Extended Multiplicative Scatter Correction (EMSC)

In the EMSC model a polynomial is also included along with MSC. Hence, it is called extended MSC [48, 49]. Generally, a second order polynomial is used in the EMSC model. Let us assume $S^1, S^2, \ldots, S^M$ are the FTIR spectra and S^μ is the average spectrum and each spectrum is a vector of FTIR absorption values (intensities):

$$S = (s_1, s_2, \ldots, s_N)$$

N is the number of spectral data points

Also, s_ν^i is an intensity value (scalar) at a particular wavenumber 'ν' on the FTIR axis of a given spectrum S^i. EMSC models each spectrum in terms of the average spectrum and a polynomial by the following equation:

$$s_\nu^i = a^i + \left(b^i * s_\nu^\mu\right) + \left(c^i * \nu\right) + \left(d^i * \nu^2\right) + \epsilon_\nu^i ; i = 1, 2, \ldots, M ; \nu = 1, \ldots, N$$

Here, a second order polynomial is used for baseline correction. One can choose to include a higher order polynomial if required. Once the parameters a^i, b^i, c^i, d^i are calculated then the corrected spectrum is given by the following equation:

$$S\_emsc^i = \left(E^i / b^i\right) + S^\mu ; \; i = 1, 2, \ldots, L$$

where $E^i = (\varepsilon_1^i, \varepsilon_2^i, \ldots, \varepsilon_N^i)$ is the corresponding residual spectrum and S_emsc^i the corrected spectrum of S^i. Furthermore, in the literature some innovative modifications are proposed as part of EMSC such as including a representative spectrum of common sources of interference such as water vapor and paraffin [50]. EMSC is by far the most commonly and widely used technique for preprocessing of FTIR spectra as it gives flexibility to model various interference sources, background and scattering effects together.

Exclusion of Low SNR signals

In the case of FTIR imaging it is extremely important to carry out quality tests prior to any other preprocessing steps to make sure poor quality (low SNR) signals are eliminated. This is due to the fact that in some of the preprocessing techniques like EMSC where the entire gamut of spectra is used, the presence of these poor quality spectra can prevent the results of multivariate analysis to be carried out further. The quality tests can be performed either by defining thresholds on certain absorbance values or SNR as explained below.

a) Both upper and lower thresholds can be applied on FTIR absorbance values of a specific vibration mode, for example the amide I region, which generally indicates inconsistent sample thickness regions [46, 51]. For example, a low sample thickness is indicated by the presence of noise in the case of FTIR imaging data. Pixels (FTIR spectra) corresponding to these regions ought to be removed before proceeding with further steps. A threshold on the absorbance of amide I band can identify these regions quite effectively [44]. Similarly, a threshold on the area under such bands may also be applied to eliminate unwanted spectra.
b) Alternatively, SNR thresholding can also be applied to detect outliers. For example, in case of biological samples, the absorbance value of amide I (1620–1690 cm^{-1}) can be considered as signal and the absorbance values in the dead region or signal free zone (1800–1900 cm^{-1}) can be considered as noise (background) [46, 51]. A threshold can be applied on the SNR calculated as explained above and will remove unwanted spectra quite effectively.

Practically, Raman/FTIR preprocessing involves a subset of the above explained steps. Based on the application, preprocessing consists of a combination of sequentially executed steps in the same order as mentioned above. Sometimes, this sequence is aided by some special procedures, for example water vapor correction in case of FTIR [46] or

representative cell media subtraction in case of Raman image clustering of a single cell [42, 43] *etc.* Although there are many comprehensive studies available in the literature [22, 30, 46] on optimal preprocessing steps, the art of preprocessing is yet to be standardized. Most of the time, the right preprocessing steps depend on problem statement, observations, prior experience and intuition of the researcher.

Multivariate data analysis

The spectroscopic data can be displayed in the form of a matrix where the columns represent the wavenumbers/Raman Shifts (variables) and rows represent observations (spectra) *i.e.* each spectrum is represented by a row in the matrix [19–21, 52, 53]. In the case of hyperspectral images, data are represented in the form of a hypercube with two spatial dimensions (pixel coordinates × and y) and the third dimension is the spectral dimension. In the data matrix each pixel is represented by a row and wavenumbers in columns.

The spectroscopic measurements or observations consist of two parts:

$$\textbf{Observation} = \textbf{Relevant Signal} + \textbf{Noise}$$

Here, the relevant signal is considered as the actual representation of the underlying chemical information, which is correlated with the property of interest. The noise part is everything else that is irrelevant to the property of interest, including spectral noise. For example, if one would like to measure using spectroscopy the concentration of one component (C1) in a mixture that also contains C2 and C3, then the signals from C2 and C3 can be considered as noise along with instrumental noise. One of the most important objectives of multivariate analysis is to separate the relevant signal from the noise part by using intrinsic variable correlations in a given data set. The concept of variance is very important as "directions with maximum variance" are almost directly related to the structural part of the relevant signal [52–54].

There are many multivariate data analysis techniques available and for an appropriate selection the goal of the analysis should be clearly defined. The three main objectives of multivariate data analysis are defined below:

1. Data description and explorative data structure modeling of any generic data matrix. Principal Component Analysis (PCA) is frequently used for this purpose
2. Discrimination, Classification, Clustering deal with dividing a data matrix into two or more groups of measurements (objects).
3. Regression and Prediction: Regression is a method for relating two sets of variables by quantifying them with respect to each other.

The multivariate analysis methods are broadly divided into two groups

Unsupervised methods: These methods are used when there is no supervising guidance (labeling) available *e.g.* PCA. Unsupervised methods are very useful to find hidden structures in the unlabeled data and are often used as precursor to supervised methods when working on huge data sets. Various cluster analysis algorithms like K-means, Hierarchical Cluster Analysis (HCA) *etc.* are also considered as unsupervised methods.

***Supervised methods*:** These methods differ from unsupervised methods due to the fact that they label the classes to be discriminated. Unlike unsupervised methods, there are two important phases. The 'training phase' is considered as a passive modeling stage, which uses a 'training data set' (which is labeled) to find the patterns in the data. The model parameters learned during the training phase are stored for further validation. The second phase, called the 'prediction phase' (testing phase) is the active stage where the unseen data (the data which were not part of the training set) are validated using the model parameters learned in the first phase, using for instance Discriminant Analysis (DA), Multiple Linear Regression (MLR), Principal Component Regression (PCR), Partial Least Squares (PLS), Support Vector Machines (SVM).

The main disadvantage of these supervised methods is their dependency on labeling data. When the data set has a large number of observations it is very cumbersome to label each and every one of these observations. So unsupervised methods are often used in conjunction with supervised methods to solve this problem. Here, out of many observations, only a few observations are labeled in the beginning. Initially all of these observations are fed to unsupervised methods like cluster analysis. Based on the clusters obtained, each unlabeled observation in a particular cluster is labeled with the dominant class label of already labeled observations in the same cluster. At the end of this procedure, all the observations would have been labeled. Then, the labeled data are fed to supervised methods for further classification. This type of algorithms (methods) is called semi-supervised methods.

As mentioned earlier, some combinations of variables (wavenumbers) of a given data set are highly correlated with each other. If one can capture these underlying correlation patterns, it is useful in representing the data set compactly with fewer variables. These variables are linear combinations of the original variables. Obtaining these new variables in lower dimensional space is a well known problem in multivariate data analysis and is known as "Dimensionality Reduction". It also helps in separating out relevant signals from unwanted noise. Moreover, many classification and clustering algorithms are quite expensive in computational complexity. It makes more sense to transform the original variables to a lower-dimensional space before feeding the data to these classification algorithms. PCA is one such widely used dimensionality reduction technique.

Principal component analysis (PCA)

PCA is an unsupervised data transformation procedure of complex data sets. PCA is used for projecting a higher dimensional data matrix "X" onto a low component subspace. It reduces a set of variables into a smaller set of orthogonal, and therefore independent, principal components (PCs) in the direction of maximal variation *i.e.* it reduces the dimensionality and retains the most significant information for further analysis. PCA tries to decompose the data matrix "X" with m object rows and n variable columns ($m \times n$ matrix) into a structured part (S) and a noise part (E). The m objects are different observations (spectra) and the n variables are the measurements (wavenumbers) for each object. The n variables jointly characterize each of the m objects. The n-dimensional co-ordinate system consists of orthogonal axes with a common origin for the variables, called the 'variable space'. The number of independent basis vectors *i.e.* the number of independent sources of variation within the data matrix may

often be lower than n which is the backbone of dimensionality reduction. Assume that we have a matrix "X" with 500 spectra from different samples (objects) and recorded over 1000 wavenumbers (700–1700 cm^{-1}) *i.e.* variables. So, each spectrum is a point in this "1000- dimensional" variable space, and we have m different observation points for each variable. Vibrational spectra (Raman and IR) contain the cumulative chemical information of all the biochemical molecules present in the sample in terms of intensities at these different wavenumbers (bands). Among these, several features have the same origin of variation, which results in strong correlations between few variables (wavenumbers) and sets the stage for dimensional reduction.

The PCs obtained from PCA can be defined as variance-scaled vectors in the variable space. As mentioned earlier, the main objective of PCA is the transformation of a coordinate system from n 'variable space' into a new and more relevant 'PC coordinate space' while simultaneously dropping "the noise part". These PCs are obtained by calculating the eigenvectors and eigenvalues of the covariance matrix obtained from the data matrix (X). The eigenvector with the highest eigenvalue gives rise to the first PC (PC1) *i.e.* the direction of greatest variance in the data. "PC1" is the direction (axis) that maximizes the longitudinal (along axis) variance or the axis that minimizes the squared projection (transverse) distances. PC1 demonstrates the maximum variance in the data and the second principal component (PC2) illustrates the largest residual variance along a direction orthogonal to PC1 and so forth. These PCs are completely uncorrelated and independent, leaving no further scope for dimensionality reduction.

For an X-matrix ($m \times n$), the largest number of PCs can be either one less than the number of objects (m-1) or equal to the number of variables (n) depending on whichever is smaller. The higher-order PCs (directions) are progressively thought of as noise directions which accounts for noise component. Despite thousands of spectral channels the relevant information (spectral variance) can be explained by the first few dominant PCs (say 'k') as repeated information is present in various spectral channels. This new coordinate system consists of only a few orthogonal PCs and the optimal number of the PCs ('k') to be retained depends on the eigenvalues of the PCs. Higher eigenvalues represent PCs with less noise, but as the eigenvalues decrease the SNR of the PCs also decreases. There are some well known techniques such as "scree plots" and "percentage of variance explained" which are used to determine the optimal 'k' [55, 56]. Eigenvalues corresponding to the PCs are sorted in descending order and plotted against the PC number to obtain a scree plot. A scree plot looks like a steep curve initially and as the number of PCs increase the scree plot tends to get flattened. The optimal 'k' is around the point where the scree plot begins to level off. Similarly, we can plot the percentage of variance explained against the PC number to determine the optimal 'k'. Usually there is a steep increase in percentage of variance explained, followed by a flat line.

The origin of the PC-coordinate system can be obtained by translation of the origin in variable space to the average object ("centre of gravity" of the group) and this common origin of the PCs is called the mean centre. Each PC can be represented as a linear combination of the n unit vectors of the variable space. Each PC is also called a "loading" and the coefficients in the linear combination representing the PC indicate the contributions of each variable (wavenumber) in the original variable space. The loading matrix (U_K) consists of the k PCs that have been retained and acts as the transformation matrix between the original variable coordinate system and the new

PC-system. Each column in the matrix "U_K" represents a PC *i.e.* the loadings. The values of each object in the new coordinate system are called PC scores of the object, *i.e.* the projections of an object '*i*' onto the PC1, PC2, PC3 and so on give the corresponding scores *yi1*, *yi2*, *yi3* and so forth. So, object '*i*' corresponds to a point in the new PC-coordinate system with scores (coordinates) *yi1*, *yi2*, *yi3* and so forth. The number of scores is the same as the number of PCs *i.e.* '*k*'. The score matrix "Y" constitutes all the scores for all the objects and the scores of each object make up a row.

$$X = YU_K{}^T + E = Structure + Noise;$$

$$Y = XU_K$$

where X is the original m x n data matrix

Y is the $m \times k$ scores matrix and U_K is then $n \times k$ loadings matrix

The PC-model is the structure part *i.e.* YU_K^T and $\boldsymbol{E}$ is the measure of lack-of-fit of the model; a smaller $\boldsymbol{E}$ represents a better model. The score vectors *i.e.* columns of the "Y" matrix are orthogonal and each column represents the scores for one particular PC. These score vectors are the footprints of the objects projected down onto the PCs. Two pairs of score vectors plotted against each other (called score plot/scatter plot) are used as a 2D window into PC space, which depicts the relation between the objects with respect to those PCs.

In a PCA model each pixel (spectrum) of an image is represented by a small number of PC scores instead of the full range of wavenumbers (full spectrum) *i.e.* variables that depict the complete information about its chemical composition. The pixels of similar composition in an image are expected to have similar score values and this is used for image clustering. The 2D scatter plot in the PC-space displays the pixel clusters, which can be visually identified. So, these score plots can be used for outlier detection, identification of trends, groups, exploration of patterns *etc.* [38, 41, 45, 47, 57–59]. This is illustrated by performing PCA analysis (Software used in this work: PLS and MIA Tool Box from Eigenvector Inc. USA with MATLAB (Ver. 11A) from MATHWORK) on the raw and preprocessed (as explained earlier) FTIR images of mouse liver tissue at zero, half, three and six hours post acetaminophen (paracetamol) treatment (Fig. 2). The first major three PCs were selected (contribution higher than 5 % with total cumulative value of above 90 %) for observation. The first three columns indicate the PC score images of the raw FTIR data. There are notable differences in the PC scores within the same image. This can be attributed to thickness changes within the same image, scattering effects and DC shift. Also, there is no significant change in PC scores across the different time points. In a nutshell, due to the high variance of PC scores within the same image, it is very difficult to differentiate between different time points using PCA of raw data.

As illustrated in the last three columns of Fig. 2, preprocessing certainly enhanced the uniformity of the PC scores within the same image and the discriminating capability across the time points is also improved. Although there is a lot of discrepancy in the tissue samples as they are collected from different mice from different regions of the liver, PCA along with preprocessing shows the potential to discriminate between different time points effectively.

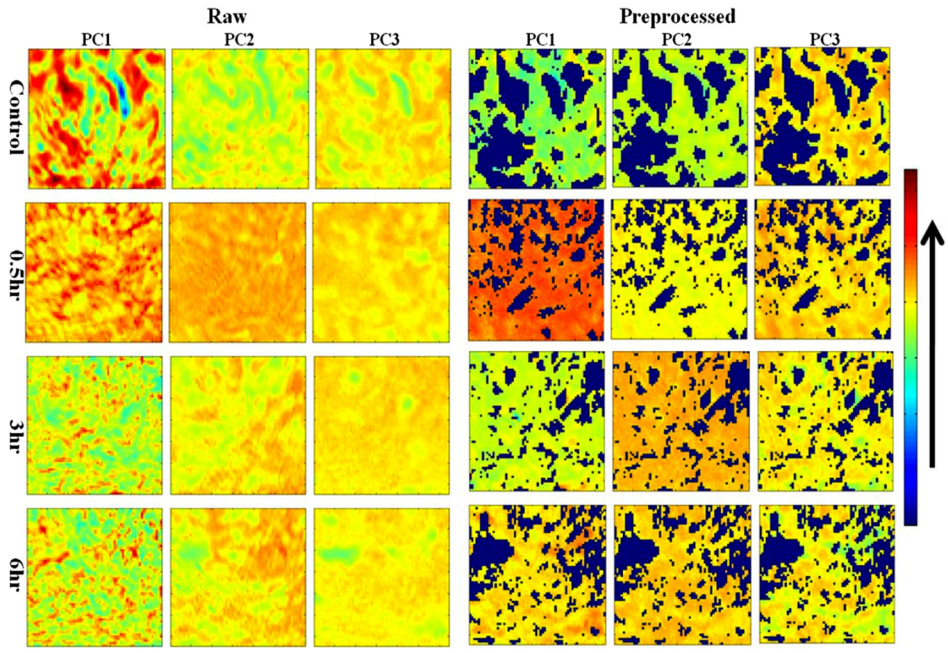

Fig. 2 Principal Component Scores of raw and preprocessed FTIR images of mouse liver exposed to acetaminophen. The rows represent control, 0.5 h, 3 h and 6 h images respectively. The first three columns represent PC1, PC2, PC3 score images of the raw FTIR spectra. The last three columns represent PC1, PC2, PC3 score images of the preprocessed FTIR spectra. Dark blue regions in the preprocessed images (4^{th}, 5^{th} and 6^{th} columns) indicate the presence of glass substrate. Uniformity within the same image and differentiation capability across different time points is enhanced by preprocessing the data

In the case of spectroscopic data which can have up to 1000 variables the 2D loadings plot is usually quite complex and difficult to interpret. So it is better to plot a one-dimensional vector (taking into consideration one PC at a time) also known as a loading spectrum. These spectra are used for the assignment of the most important variables (bands) contributing to the structural part of the data matrix [60–63]. Although the loading spectra are completely uncorrelated, they cannot be directly associated with a single chemical compound always. In other words, there is a significant difference in the mathematical properties of a loading (PC) and its chemical interpretation.

Undoubtedly PCA is capable of identifying some important structural information in the data but it has less discrimination power due to the fact that it is an unsupervised procedure *i.e.* it does not try to model patterns which are important for classifying one group with another or quantifying the expected outcomes in terms of measured variables, rather it models patterns which compactly represent the data. Often, interpretation of the complex biochemical information obtained through vibrational spectroscopic techniques requires further data analysis using supervised procedures like LDA, HCA, PLS, PCR *etc.* As discussed earlier, each of these methods is meant to solve different problems. Some of the important problems, the corresponding multivariate data analysis techniques and their applications to vibrational spectroscopy/chemical imaging are explained in detail below.

Classification models

In statistics, classification is defined as the categorization of given objects (observations) into two or more types. In vibrational spectroscopy classification models are extensively used in a wide range of applications from forensics to medicine [64, 65]. These classification models are useful for early diagnosis and understanding the mechanism of disease progression [66–69]. Some of the important and widely used classification techniques are explained below.

a) Linear Discriminant Analysis (LDA)

The main aim of Discriminant analysis (also called Fisher's linear Discriminant analysis) is to find the "Discriminant axes" which optimally classify the data into two or more classes. LDA is closely related to PCA as both of them look for latent axes which compactly explain variance in the data. The main difference between PCA and LDA is that LDA is a supervised method and PCA is an unsupervised method. PCA looks for projections to maximize variance and LDA looks for projections that maximize the ratio of between-class to within-class scatter as depicted in Fig. 3.

Data can be projected into the new dimensional space using these axes found with LDA. In the new dimensional space, each observation would have fewer variables (dimensionality reduction) and at the same time observations belonging to the same class will form lumps (clusters) and each cluster would be clearly differentiated from the other [41, 70].

Let us assume that the original data set "X" is labeled with two different classes, where "X_1" represents data of class **1** and "X_2" represents data of class **2**. Each observation has "n" variables and there are "m" such observations out of which "m_1" belong to class **1** and "m_2" belong to class **2**. In other words, "X" is a data matrix with size "m × n", "X_1" is a data matrix with size "m_1 × n" and "X_2" is a data matrix with size "m_2 × n". Also, let us

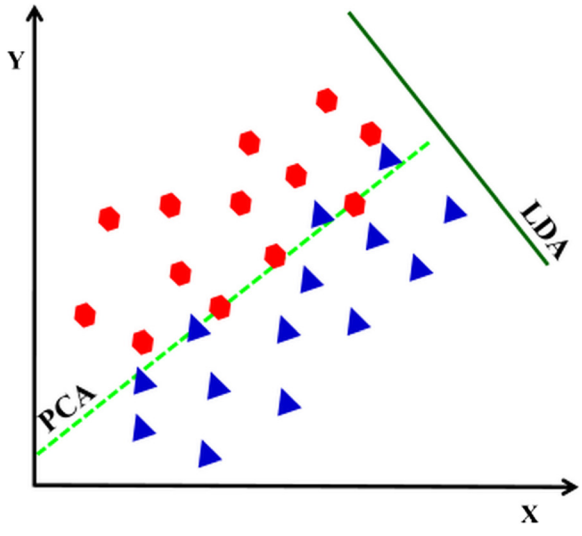

Fig. 3 Schematic representation of the axes found using PCA and LDA for a two-class data set

define "SC_w" which represents within-class scatter and "SC_b" which represents between-class scatter.

$$SC_w = \sum_{i=1}^{C} SC_i$$

where C is the number of classes
and

$$SC_i = \sum_{j=1}^{m_i} (S_j - \mu_i)(S_j - \mu_i)^T$$

where m_i is total number of observations of the i^{th} class
S_j is one such observation (spectrum) and μ_i is the mean of all such observations of the i^{th} class

$$SC_b = \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T$$

where μ_i is the mean of class $\boldsymbol{i}$ and μ is the mean of all such means
LDA tries to find the axes 'W' that maximize the objective function (ratio of between-class scatter to within-class scatter) "$J(W)$" defined as below:

$$J(W) = \frac{|W^T SC_b W|}{|W^T SC_w W|}$$

where, $W = [w_1 | w_2 | \ldots \ldots \ldots w_L]$ and L is the number of solutions (projections).
The solution to this optimization problem is given by solving the generalized Eigenvalue problem given below.

$$SC_b w_i - \lambda_i SC_w w_i = 0 \,; i = 1,\ 2,\ 3,\ \ldots\ldots\ldots,\ L$$

where each w_i (eigenvector) gives a unique projection and λ_i is the corresponding eigenvalue
Here, we may get either "*C-1 (Number of classes-1)*" or "*n (number of variables in the original data set)*" solutions, whichever is lowest. Generally, "*n*" tends to be much higher than "*C-1*" in the context of spectroscopic data. So, LDA gives "*C-1*" projections *i.e.* $L = C\text{-}1$. As explained earlier, these "*C-1*" projections (also called Linear Discriminants - LDs) not only help to achieve dimensionality reduction but also efficiently discriminate all the other classes from each other.

It is very useful to combine both PCA and LDA approaches (called PC-LDA model), which improves the efficiency of classification as it automatically finds the most diagnostically significant features. Another advantage of the PC-LDA model is that it is easy to visualize the clusters in three dimensional space using LD scores. Here, first PCA is applied to the original data set "X" and only the first few principal component scores are retained for further analysis. "X" is an "m × n" matrix and let us assume that the PC score matrix is "Y":

$$Y = X * U_k \,;$$

where U_k is an "n × k" matrix with first k principal components (PC loadings) as columns.

So, the resulting principal component scores matrix "Y" is of size "m × k". Now, LDA is applied on matrix "Y" to obtain the LD score Matrix "Z" as below.

$$Z = Y * W;$$

where U_k is an n × k matrix with first k principal components (PC loadings as columns)

So, the resulting Linear Discriminant scores matrix "Z" is of size "m × (C-1)". This matrix "Z" represents compactly the original data "X" and differentiates one class from another very efficiently. Similar to PCA, loading analysis can also be performed using the PC-LDA model. Here, each LD loading can be represented as a linear combination of PC Loadings.

Let us say that the LD loadings matrix is defined as "V" and can be obtained as:

$V = U_k * W$; where V is of size "n × (C-1)"

Here, each column of loading matrix "V" represents a particular loading and these loadings can be used to understand the role of a particular wavenumber (or band) in differentiating one class from others. It can be used to understand and identify the critical vibrational bands causing the differences between classes. In the literature, many studies have been conducted where the PC-LDA model is used for classification purposes [38, 45, 47, 58, 71]. We have also performed a PC-LDA analysis on the preprocessed Raman spectra from *Drosophila* muscles to differentiate between 2- and 12-days old flies using MatLab (Math Works, 2010) and R (Team, 2012) [9]. Raman spectra were recorded using a commercial Raman micro-spectrometer (Renishaw, InVia system) at 785 nm excitation wavelength, which was focused onto the muscles using 50x (NA = 0.75) objective for an integration time of 150 s. As a first preprocessing step cosmic ray removal was done after acquiring each spectrum using Renishaw WiRE 3.2 software. Other spectral preprocessing steps such as band alignment using local regression (LOWESS), baseline correction using the ALS method and smoothening using a SG filter with a window width of 11 and polynomial order of 5 were performed. Further all spectra were normalized using a SNV transformation and mean centering across was performed before applying PCA. As shown in Fig. 4 preprocessing is certainly needed to reduce the variability of the Raman data and thus enable the detection of minor differences. All spectra from the mutants $upheld^1$ (up^1) and $upheld^{101}$ (up^{101}) the control *Canton-S* (*CS*) were subjected to outlier removal as discussed earlier. PCA was done on the remaining valid Raman spectra for dimensionality reduction and the most significant 60 PC scores (~95 % of variance) were used to perform LDA for further classification. The 3D scatter plot of scores of LDFs illustrates a good separation between the 2- and 12-days old samples of mutants (Fig. 5) which depicts that PC-LDA model is able to differentiate between the early stage of muscle degeneration (2^{nd} day samples) and almost completely degenerated muscles (12^{th} day samples). However, the 2- and 12-days old samples of control (*CS*) flies grouped together.

b) Soft Independent Modeling of Class Analogy (SIMCA)

This approach is normally used to model each class locally. First of all, PCA is applied to the original data of each class separately to model the particular class and only

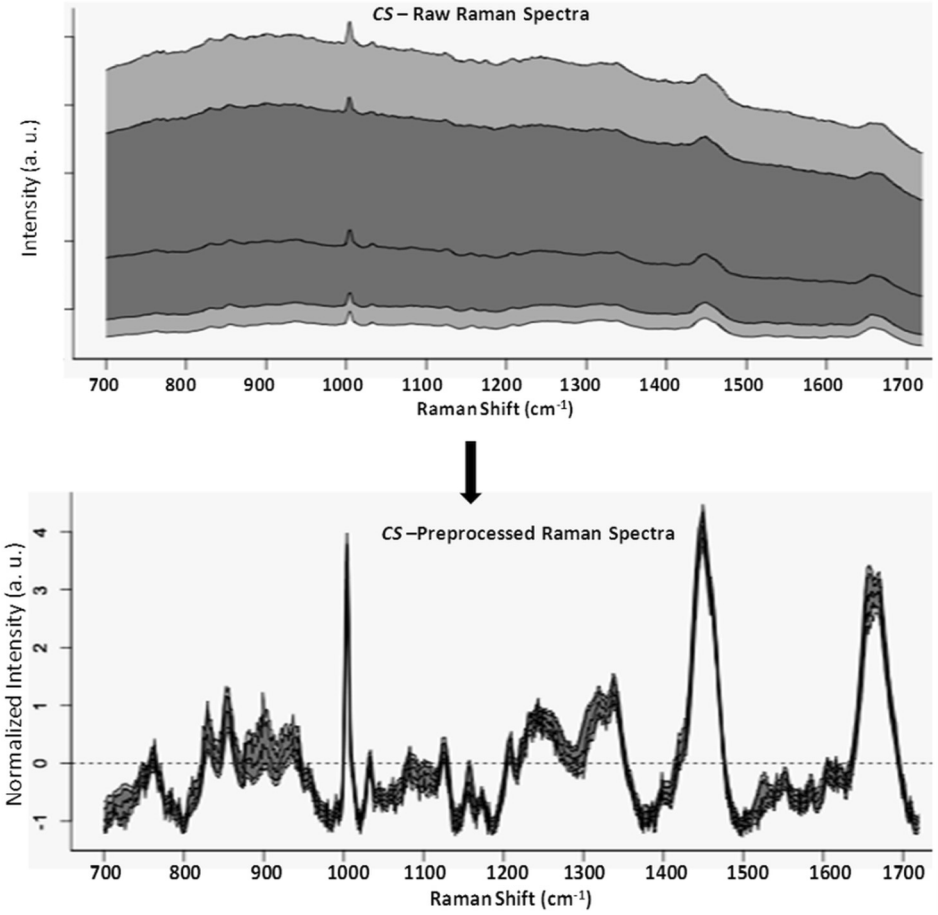

Fig. 4 a Raw and **b** preprocessed Raman spectra of muscles from control *Canton-S* (CS) 2 days old flies. The dark shaded areas correspond with the average Raman intensity +/− one standard deviation. The differentiating Raman features are enhanced by preprocessing the data, which reduces the variability among the spectra from the same group

few significant PCs are retained. The number of PCs can vary from one class to the other and this can be determined using cross validation which will be discussed later in this review. So, based on the optimal number of PCs retained, we can easily calculate the average residual variance (variance which is not explained by the optimal number of PCs) of each class. When making membership decisions, SIMCA takes into account the fact that the unknown sample (test sample) will be similar to the other samples in its true representative class in the lower dimensional space (PC scores). So, an unknown sample is projected onto every PC model (each class has a PC model) and the residual variance of the unknown sample with respect to the current PC model is compared against the average residual variance of the current PC model calculated for the training set. This comparison is used as goodness of fit to make membership decisions. The advantage of the SIMCA model is that a given unknown sample is not classified as any of the classes if the residual variance of the unknown sample exceeds a particular threshold for each PC model. Such unknown samples are considered as outliers. However, a given unknown sample could get labeled with more than one class if the residual variance of the unknown sample is less than a particular threshold for more than one class. The other disadvantage of SIMCA is that it is highly sensitive to the quality

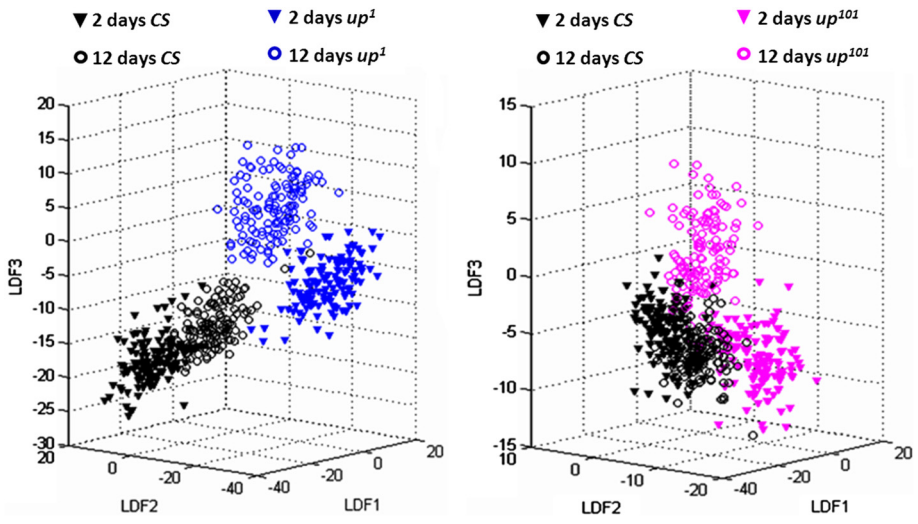

Fig. 5 PC-LDA scores plot of 2- and 12 days old flies with (left) *upheld*^{1} (*up*^{1}) mutation in comparison with control flies *CS*; (right) *upheld*^{101} (*up*^{101}) mutants in comparison with control flies *CS*. The 2- and 12 days old flies of mutants' *up*^{1} and *up*^{101} are well separated from each other, while for the control group *CS* the 2- and 12 days old flies almost merged together

(SNR) of the data used in the training phase (PC modeling phase). In spite of this, SIMCA has been applied successfully to spectroscopic data to solve many classification problems [72–75].

c) Artificial Neural Networks (ANN)

Artificial Neural Networks are computational models inspired by the functionality of the central nervous system of the human brain. Here, many artificial nodes (counterparts of neurons) are arranged in layers and each node is connected to all other nodes in the adjacent layers. Typically these layers are categorized as input layer, output layer and hidden layers. In a given neural network setup, there is only one input and output layer, but there could be multiple hidden layers. The more hidden layers, the deeper is the neural network. The strength of a neural network lies in its connections. A typical illustration of an ANN is given in Fig. 6; the graph shows only one hidden layer but there could be several hidden layers between the input and the output layers. The input layer represents the variables in a given observation, for *e.g.* intensities at all wavenumbers in a given spectrum. All of these intensities are fed as input to every node in the hidden layer as shown in Fig. 6. There are many models that represent artificial nodes in the hidden layers. Different models apply different nonlinear functions (activation functions) to the weighted sum of input values. Let us say an input observation "S" has values s_1, s_2 …., s_N and the output of the node in the hidden layer is 'h':

$$h = f\left(\sum_{i=1}^{N} w_i s_i\right);$$

where "f" represents a non linear function (activation function)

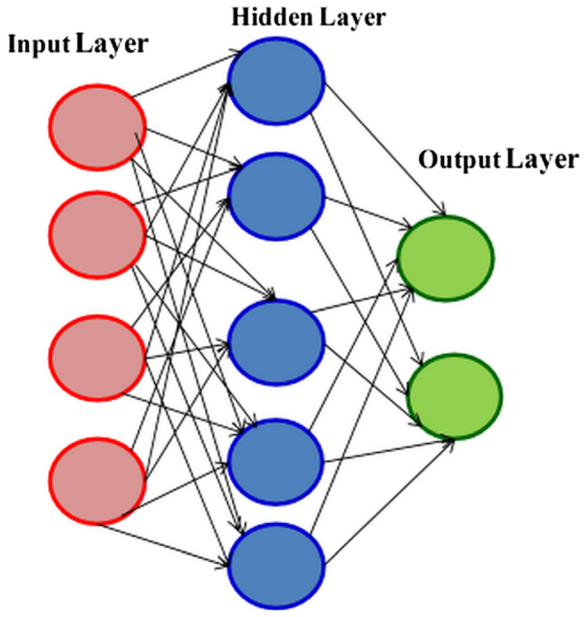

Fig. 6 Schematic representation of an ANN, in this case with a single hidden layer. The data flow from input layer to output layer *via* the hidden layer(s). Black arrows represent the connections between the artificial nodes and each node is connected to all other nodes in the adjacent layers

and "w_i" represents the weight of the edge connecting the i^{th} input node to the in the hidden layer.

One of the most commonly used activation function is "sigmoid" also called "logistic function", which transforms a value ranging from -∞ to ∞ to a value between 0 and 1 as follows:

$$S(x) = \frac{1}{1 + e^{-x}}$$

So, when "logistic function" is applied to the weighted sum of input values, we get the following equation.

$$h = \frac{1}{1 + e^{-\left(\sum_{i=1}^{N} w_i s_i\right)}}$$

The output of such nodes in the hidden layer is passed from one layer to the other through the connections between them. The absolute weight of these edges (connections) indicates the importance (strength) of the particular connection. These weights are learned in the training phase to produce the desired final output. Based on the type of node in the output layer, the ANN can be modeled as either regressor or classifier. If the node in the output layer is similar to the nodes in the hidden layer (*i.e.* with sigmoid activation function) then the ANN acts as classifier. ANN is a very powerful tool and can represent extremely complex structures in the data as it can approximate many patterns locally using different hidden layers. ANN is also used to solve classification problems in the field of vibrational spectroscopy [76, 77].

d) Support Vector Machines (SVM)

Support Vector Machines identify the decision boundaries (hyperplanes) between the classes in an optimal way *i.e.* they try to separate the observations of different classes by a clear and widest gap in *N*-dimensional space (say, each observation has 'N' variables) according to an optimization criterion. In the training phase, SVM tries to find such an optimal set of hyperplanes from the training data and when a new observation is projected onto these hyperplanes it can be easily classified with one of the class labels used in the training phase.

The schematic represents the various hyperplanes (shown in magenta, orange and green color) - in two-dimensional spaces a hyperplane becomes a line - to classify observations (in two-dimensional spaces) into two categories (Fig. 7). SVM finds the optimal hyperplane (shown in green color) that has the maximum margin from the boundaries of observations belonging to both categories. In vibrational spectroscopy many studies have already been conducted using SVM as the classifier [38, 78–80].

Spatial clustering models

The automated grouping of the pixels in an image having the same characteristic bands is called spatial clustering. This is done by considering two important criteria: (i) pixels in the same group are as similar as possible and (ii) pixels in different groups are as dissimilar as possible. Various multivariate methods (unsupervised) such as K-means Cluster Analysis (KMCA), Agglomerative Hierarchical Cluster Analysis (AHCA), Principal Component Analysis (PCA), Fuzzy C Means Cluster Analysis (FCMCA), Vertex Component Analysis (VCA), and Divisive Correlation Cluster Analysis (DCCA) are being widely used for cluster analysis of Raman and IR images [45, 81, 82]. Spatial clustering algorithms can be divided into two main categories:

Hard Clustering: Each pixel (object) belongs to only one of the clusters. Many clustering algorithms like KMCA, HCA are hard clustering algorithms.

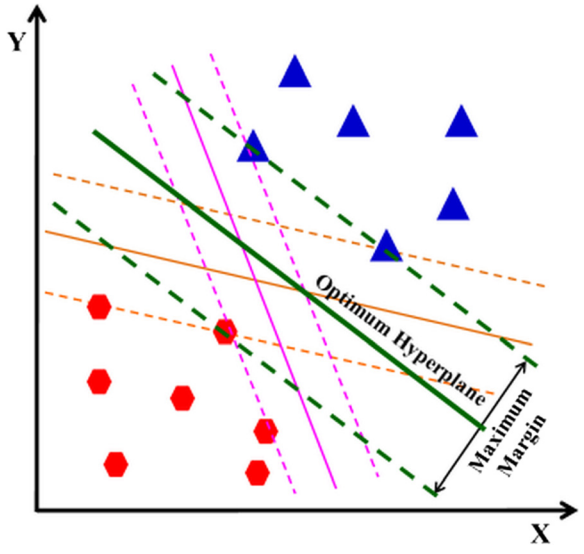

Fig. 7 Schematic representation of Support Vector Machine (SVM), identifying the optimal hyperplane for a two-class data set

Soft Clustering: Each pixel (object) belongs to some extent to each cluster *i.e.* the algorithms capture the probabilities with which a given pixel belongs to each class and these probabilities should sum to '1'. Algorithms such as FCMCA belong to the soft clustering category.

Hard clustering can be described as soft clustering with the probability of a pixel belonging to a particular cluster is '1' and all other remaining clusters is '0'.

Clustering algorithms can also be distinguished based on the approach and underlying statistical methodologies [83]. Some of the categories of cluster algorithms are listed below:

Hierarchical Approaches: These methods use recursive approaches either in top-down or bottom-up fashion in partitioning a given data set. Algorithms like AHCA, DHCA belong to this category.

Partitioning Approaches: These methods iteratively keep shuffling the cluster labels from one pixel to other until a particular criterion (objective function) is minimized, for *e.g.* KMCA.

Graph Based Approaches: Graph theory related techniques like minimal spanning tree and max-flow min-cut *etc.* are used for clustering. Here, data is converted to Graph by representing pixels (objects) as nodes and the distance between the pixels as edge weights. Based on the particular type of problem the definition of distance (edge weights) varies. For example if we want to increase the probability of pixels adjacent to each other falling in the same cluster, we can incorporate the spatial location information in the distance calculation to make sure adjacent pixels have less distance from each other.

Density Based Approaches: These methods are somewhat similar to partitioning based approaches, but they calculate the probability of pixels belonging to each cluster. Here, it is assumed that the probability of a given pixel belonging to each cluster can be modeled by a particular probability distribution function. The overall distribution of the data is assumed to be a mixture of several such distributions. Algorithms like Gaussian Mixture Models (GMM) can be categorized under these approaches.

However, when it comes to spectroscopy only few of the above mentioned algorithms are being widely used. Some of those algorithms are explained in detail later in the review. One common problem with most of these clustering algorithms is to find the optimal number of possible clusters for the data set. Although there are a few rules of thumb for the selection of the ideal number of clusters, it depends mostly on the problem at hand and the intuition of the chemometrician. Nevertheless, there are a few simple and useful guidelines to identify an ideal number of clusters and these techniques vary from method to method. More details about the techniques specific to each method are explained below.

a) K-means Cluster Analysis (KMCA)

The K-means algorithm is the simplest and one of the most widely used clustering algorithms in vibrational spectroscopy [45, 47, 57]. Along with observations, KMCA expects additional input parameters like the number of clusters (K) and initial cluster centers. KMCA is an iterative algorithm and two steps are performed in every iteration:

Step-1: Find the distances from a given observation to each cluster centre and label the observation with the label of the cluster centre which is the closest. At the end of this step each observation would have been labeled with a new label.

Step-2: Recalculate the cluster centers based on the new labels obtained in *step*-1.

KMCA repeats the above explained steps until a convergence criterion is met. One of the most commonly used convergence criteria is that the distance between the current (iteration) cluster centers and the previous (iteration) cluster centers is less than a selected threshold. As mentioned earlier, KMCA requires the initial cluster centers to be fed as input before running the algorithm. Generally there are two ways of assigning the initial cluster centers. In the first approach, all the observations are randomly labeled with the possible number of labels (K) and subsequently cluster centers are calculated. In the second approach, from all observations in the data set, 'K' observations are selected randomly as cluster centers. Irrespective of the method used for calculating the initial cluster centers, each time KMCA may converge to an entirely different solution because of the randomness in allocating initial cluster centers. This problem can be solved by running KMCA many times (say 100) and finally selecting one of those solutions based on some disparity criterion. Disparity is calculated as the sum of distances from the final cluster centers to the observations in the given cluster as shown below. Let us say we have 'K' clusters named as C_1, C_2, C_3 … C_K with cluster centers as μ_1, μ_2, μ_3 … μ_K respectively:

$disparity = \sum_{i=1}^{K} \sum_{x \in C_i} dist(x, \mu_i)$; where "dist" represents the distance metric

Here, generally Euclidean distance is used as distance metric. Many clustering algorithms including KMCA have a dependency on the number of clusters "K". Most of the time the optimal "K" is subjective as it depends on the problem being solved. In some of the earlier studies conducted in the field of vibrational spectroscopy, KMCA was run by varying "K" from 2 to a very high number (say 20). Then, the optimal number of "K" was chosen by comparing the KMCA clustered images with histopathological images [82]. As "K" increases underlying spatial patterns are revealed and beyond a certain number of clusters these patterns will not yield any extra information. As these observations are done qualitatively, the optimal number of "K" varies from problem to problem. There are also some quantitative approaches available to determine the optimal number of "K". One such method is called the "elbow method". Here, for each "K" starting from 2, the disparity measure is calculated. As "K" increases the disparity decreases. But, beyond a certain "K" the change in disparity is very minimal. So, based on the percentage change in disparity an optimal "K" can be obtained.

b) Hierarchical Cluster Analysis (HCA)

HCA is one of the most powerful and frequently used methods in chemical (Raman and IR) imaging. In HCA many of the distance metrics, explained earlier in the beginning of the review, can be used, of which Euclidean and Mahalanobis distance metrics are the most common. One of the biggest advantages of HCA is that it does not have to run again and again as the number of clusters varies. There are mainly two different approaches in HCA, firstly a "bottom up" approach called AHCA and secondly a "top down" approach called DHCA.

Agglomerative hierarchical cluster analysis (AHCA)

In AHCA, all the observations in the data set are considered to be belonging to different clusters. In other words, each observation belongs to a different cluster. At each step of AHCA, two clusters are merged based on a particular distance criterion until all the clusters are merged into a single cluster. Many different distance criteria are used to merge the clusters; some of the most popular methods are listed below.

Single-linkage: The distance between two clusters is defined as the minimum of all distances between the observations in one cluster and the other cluster:

$$single\text{–}linkage\ distance = min\{dist(x,y) : x \in X, \quad y \in Y\};$$

where X and Y are two different clusters

Complete-linkage: The distance between two clusters is defined as the maximum of all distances between the observations in one cluster and the other cluster:

$$complete\text{–}linkage\ distance = max\{dist(x,y) : x \in X, \quad y \in Y\};$$

where X and Y are two different clusters

Average-linkage: The distance between two clusters is defined as the average of all distances between the observations in one cluster and the other cluster:

$$average\text{–}linkage\ distance = \frac{1}{N_X N_Y} \sum_{x \in X} \sum_{y \in Y} dist(x,y);$$

where X and Y are two different clusters;

N_X,N_Y are the corresponding numbers of observations

Ward's Criterion: Here, at every step two clusters that yield a minimum increase in total within-cluster variance (which is similar to disparity of single cluster) are merged [39, 40, 43].

This whole hierarchical process is captured as a "Dendrogram". A dendrogram is a tree structure used to illustrate the merging process and corresponding distances at each step. Based on the optimal distance we can split the data set into a particular number of clusters or based on the number of clusters required an optimal distance is chosen to split the data set into clusters using "Dendrogram".

Divisive hierarchical cluster analysis (DHCA)

The DHCA process starts with a single cluster and ends up with the same number of clusters as the number of observations in the data set. At every step, a chosen cluster is split in two based on a particular criterion. AHCA has less computational complexity and is the most popular technique for chemical (Raman and IR) imaging [39, 40, 77, 84, 85].

Here, we demonstrate the use of AHCA for spatial clustering of a HeLa cell, the most commonly used human (cervical cancer) cell line. Raman images were generated by raster-scanning the laser beam (633 nm) over the Hela cell with a step size of 1 μm (Fig. 8). The excitation line was focused using a 63x (NA = 0.9) water immersion objective for an integration time of 15 s using an inverted microscope system. The Raman image is classified into four clusters using AHCA to visualize regions with high Raman spectral similarities. The white light image is shown in Fig. 8a. Since significant portions of the image represent only substrate and buffer (background), a high SNR filter is employed to make sure such unwanted signal is removed prior to multivariate

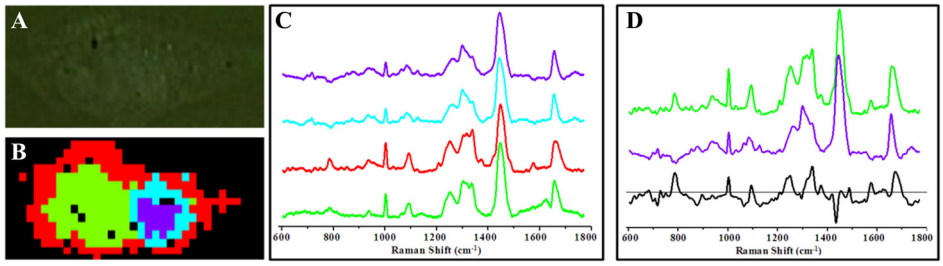

Fig. 8 Hierarchical Cluster Analysis of Raman image from a HeLa cell, the most commonly used human (cervical cancer) cell line. **a** White-light image; **b** corresponding Raman image of four clusters obtained by using AHCA multivariate technique; black regions indicate low SNR spectra (mostly buffer and substrate), which were removed prior to AHCA during the preprocessing stage; **c** average Raman spectra of four clusters; and **d** the difference spectrum (black) obtained by subtracting the average spectrum of the purple region from the average spectrum of the green region

analysis. The black regions in Fig. 8b represent the pixels which were removed before clustering. Also, subsequently each spectrum in the image was baseline corrected, smoothed and normalized to eliminate effects of instrumental drift. In order to reduce dimensionality, eliminate noise and to improve computational efficiency PCA was performed prior to cluster analysis [39, 40]. The thirty most significant PC scores were retained and further fed to AHCA for clustering and the corresponding average spectra from each cluster were plotted (Fig. 8c). The average spectrum from the green cluster belongs to the nuclear region due to the presence of Raman bands at 785, 810, 1093, 1338 and 1576 cm^{-1} which are assigned to nucleic acids. The positive nucleic acid bands in the difference spectrum *i.e.* the average spectrum of the green cluster minus the average spectrum of the purple cluster also illustrates the high nucleic acids content in the green region (Fig. 8d). Importantly, the negative bands at 717, 1064, 1440 and 1740 cm^{-1} in the difference spectrum indicate that the purple cluster is dominated by vesicular lipids. The Raman image of HeLa cell was recorded 2 h after isolating from the media, which must have resulted in the formation of vesicles due to stress [86]. This is also clear from the granular structures seen in the white light image. Overall, we could divide the Raman image according to the chemical components and their relative intensities effectively using multivariate analysis.

a) Fuzzy C Means Cluster Analysis (FCMCA)

Unlike KMCA and HCA, FCMCA is a soft clustering algorithm. In other words, each observation belongs to more than one cluster with a given probability (membership value). Otherwise, FCMCA is very similar to KMCA except that these membership values have to be included in the objective function. In vibrational spectroscopy FCMCA is often used for solving the soft clustering problem [73, 84, 87].

Multivariate regression models

Unlike classification or clustering, regression is used in quantification of particular dependent variables (expected outcome of the experiment). In the case of regression methods two matrices are used, one with dependent variables Y and the other is our

previously introduced "X" matrix. Using these data matrices an underlying regression model (coefficients) is learned during the training phase. This model is used to predict Y-values from new measurements of X. This solves the problem of estimating new measurements (Y) which may be expensive, difficult, time consuming, dangerous, ethically undesirable *etc.* to obtain experimentally [58, 78, 79, 88, 89]. There are many multivariate regression methods available and some of the commonly used methods in vibrational spectroscopy are as follows.

In case of *Multiple Linear Regression* (*MLR*), a linear relationship is assumed between independent and dependent variables. The regression of one Y-variable is done on a set of independent X-variables using the Least Squares Criterion [78, 79]. In spectroscopic applications the predictor variables (measurements at different frequencies) are highly correlated with each other, which leads to an ill-conditioned least squares problem. This can be solved by projecting the original variables/measurements into a lower dimensional space where the latent predictor variables are not correlated with each other. *Principal Component Regression* (*PCR*) and *Partial Least Squares* (*PLS*) are two such methods.

In the case of *Principal Component Regression* (*PCR*), X-variables are first subjected to PCA and then the Y-variable/s is/are regressed onto this decomposed X-matrix [52, 54]. A major shortcoming of PCR is that although the latent variables obtained from PCA maximize the variance in predictor variables, they may not be optimal for predicting the response as covariance between the predictor and response variables is not considered when calculating these latent variables.

Partial Least Squares (*PLS*) is an improvement over PCR where limitations of MLR are also overcome. In PLS covariance between predictor and response variables *i.e.* X^TY is subjected to Singular Value Decomposition (SVD) in contrast to X^TX as in the case of PCR to obtain the latent variables which are further used to predict the response variable [20, 21, 52, 54]. Also in the case of MLR and PCR, correlations between dependent variables are not considered. In other words, it is assumed that dependent variables are independent of each other. In the real world it is quite possible that the dependent variables which are being estimated from the same pool of independent variables are correlated with each other. The PLS method can quite effectively handle one or more co-varying dependent variables. This is done by projecting both X and Y into latent variable spaces T and U respectively, such that T and U are coupled, and chosen to maximize covariance between predictor and response variables *i.e.* X^TY. Subsequently a linear regression function between the latent variables, T and U is also learned. Given a new observation, first they are projected into latent space defined by T and further using linear regression the latent response variables are predicted which are in the space defined by U. Now, the actual response variables are obtained by back transforming the predicted latent response variables defined by U.

PLS is one of the most widely used analytical techniques along with vibrational spectroscopy to estimate and quantify the signature of various components in a given sample. The applications range from forensics to medicinal research [90–92]. Harvey Lui et al. performed principal component with general discriminant analysis (PC-GDA) and PLS on Raman spectroscopic data to distinguish malignant from benign skin lesions with good diagnostic accuracy [91]. This application has been commercialized jointly by Verisante Aura™ to aid medical professionals in the detection of cancer. Infrared

absorption spectroscopy along with multivariate analysis tools is also used for noninvasive *in vivo* glucose sensing for human subjects [92].

Performance metrics and validation of statistical models

Multivariate statistical methods, and in particular classification techniques, may perform excellently on the training data where the parameters of the model are learned. However, they may not adapt very well to the new unseen data supplied in the testing phase. So it is very important to understand the robustness of the model built during the training phase before moving to the validation phase. This also helps in evaluating the prediction strength of the model. Cross Validation is one such technique that determines the robustness of the statistical model. Before moving on with Cross Validation it is important to define some performance metrics of the classification model. Results obtained from a classification model can be tabulated in the form of a truth table. A simple example of a truth table (also called confusion matrix) for a two-class (say, positive and negative classes) classifier is shown in Table 1.

True Positive (*TP*): Number of observations which are originally positive and also classified as positive.

True Negative (*TN*): Number of observations which are originally negative and also classified as negative.

False Positive (*FP*): Number of observations which are originally negative but classified as positive. These errors are called Type-I errors.

False Negative (*FN*): Number of observations which are originally positive but classified as negative. These errors are called Type-II errors.

Based on the values in the truth table some of the important performance metrics are:

Sensitivity (*True Positive Rate – TPR or recall*):

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity (*True Negative Rate – TNR*):

$$Specificity = \frac{TN}{TN + FP}$$

Precision (*Positive Predictive Value – PPV*):

Table 1 Truth table for a two-class classifier

		Ground truth (desired output)	
		Ground truth – positive (Class 1)	Ground truth – negative (Class 2)
Classifier Result (Statistical model Output)	Classifier result – Positive (Class 1)	True Positive (TP)	False Positive (FP) Type–I Error
	Classifier result – Negative (Class 2)	False Negative (FN) Type–II Error	True Negative (TN)

$$Precision = \frac{TP}{TP + FP}$$

False Positive Ratio – FPR:

$$FPR = \frac{FP}{FP + TN}$$

Negative Predictive Value – NPV:

$$NPV = \frac{TN}{TN + FN}$$

Total accuracy:

$$Total\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Cross validation

As mentioned earlier, cross validation of the model is important (i) to avoid over fitting or under fitting the model due to an inappropriate selection of components used, and (ii) to determine the associated prediction error for future work (For *e.g.* in the testing phase). This is done by dividing the original data set into a training and a validation dataset, where the training set is used for learning the parameters and the validation set is used for evaluating the performance of the classification technique [72, 93, 94]. There are many different versions of cross validation, but some of the methods which are more popular are listed below.

Leave One Out Cross Validation (LOOCV): Here, one observation is excluded at a time from the training set and the resulting model is evaluated on this left out observation. This is repeated for all the observations in the data set and the average performance across those iterations is considered as the performance of the classification model.

K-Fold Cross Validation (KFCV): Here, the original data are divided randomly into K-sub folds (subsets). Each subset is excluded at a time from the training set and the resulting model is evaluated on the left out subset. This is repeated 'K' times till all the subsets are validated. The average performance across the 'K' iterations is considered as the performance of the classification model.

Receiver operating characteristics (ROC) curves

ROC curves are used to determine the ability of a classification model to discriminate negative from positive test results. Most of the classification models give an output such as the probability that a particular observation belongs to each class. This probability can be discretized by applying a threshold on the probability values. For a two-class problem, if the probability value is greater than the threshold it is considered as a positive outcome, otherwise as a negative outcome. Obviously it makes more sense to keep this threshold as 0.5. It is easy to visualize that if this threshold is '0' then all the observations are classified as positive. Similarly, if the threshold is kept as '1' then all the observations are classified as negative. So, it is interesting to observe what happens to performance metrics like "sensitivity" and "specificity" as this threshold is varied from '0' to '1'. The ROC curves capture this behavior by plotting "sensitivity" against "FPR (1- specificity, which indicates false alarms)" for the different possible probability

thresholds of a classification model [9, 38, 68, 95]. It exhibits the tradeoff between "sensitivity" and "specificity" *i.e.* an increase in "sensitivity" will be accompanied by a decrease in "specificity". A model, with 100 % sensitivity and 100 % specificity *i.e.* with no overlap in the two distributions, has an ROC plot that passes through the upper left corner (Fig. 9). As the curve approaches the 45° diagonal line (also called "Line of No Discrimination") the accuracy of the model decreases. If the classifier is randomly allocating the class labels the ROC of such a classifier should be the same as the 45° diagonal line. As a result the area under the curve (AUC) of the ROC is an important measure of the predictability of a classification model. A strong and robust classifier should have an AUC close to '1'. If the AUC of a classifier is less than '0.5' then it is performing worse than the "Line of No discrimination" *i.e.* the classifier exhibits the opposite of the desired behavior.

Negative control studies

This is used in evaluating the supervised classifier to understand its behavior when purposefully wrong labels are fed to the classification model [96]. Here, actual labels (ground truth) of the observations in the training set are randomized (shuffled) and fed to the classification model. Subsequently performance metrics like "PPV", "sensitivity" and "total accuracy" are collected by repeatedly performing the study as explained above. Each time when the study is performed, a new set of randomized training set labels (which are not the same as ground truths) are fed to the classifier. For a robust classification model it is expected to get very low values for the above mentioned performance metrics (say, 0-50 %). If it is not the case then the classifier is very sensitive to potential confounding variables and correlations in the data set, making it highly volatile particularly when subjected to new test data.

Conclusion

Various multivariate data preprocessing, analysis methods, and the validation criteria, which are commonly used for Raman and IR spectroscopy, are described in this review.

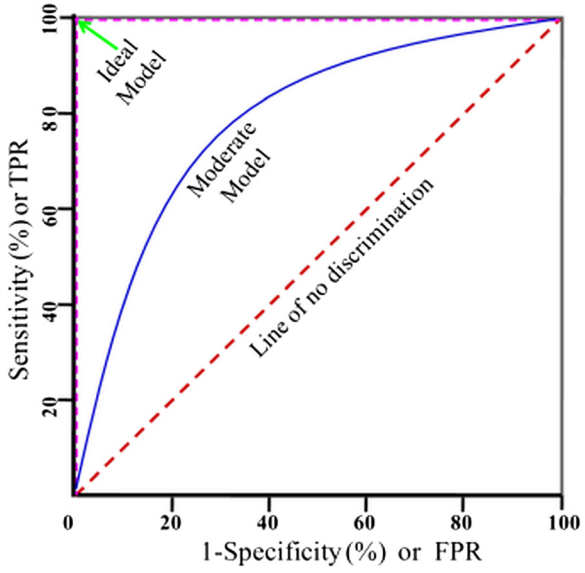

Fig. 9 Illustration of a receiver operating characteristic (ROC) plot and three levels of performance

Data preprocessing is the crucial step in Raman and IR data analysis to extract the accurate information. Most widely used data preprocessing methods are discussed but it is still active research area as we need to optimize the steps according to the sample and its native matrix. Depending on the objective of the study and depending on the technique (Raman or FTIR) used, one (or more) of the data preprocessing and analysis methods can be applied for effective interpretation of the data. One has to select a subset of the data analysis methods for solving the specific problems.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RG designed and performed the experiments, analyzed the data and wrote the manuscript; SV designed the algorithm, performed coding, data analyses and wrote part of the manuscript; FA contributed to the text, data analyses and interpretation; SU designed, conceptualized and supervised the research program, contributed to the data analyses and text. The manuscript was edited by FA and SU, and then read and approved by all authors.

Acknowledgements
The authors gratefully acknowledge Prof. D. Nandi (BC, IISc) and Dr. U. Nongthomba (IPC, IISc) for providing the samples. The authors would also like to thank the Indian Institute of Science (IISc), Council of Scientific & Industrial Research (CSIR), Department of Biotechnology (DBT) and Department of Science and Technology (DST), government of India, for their financial support. Prof. S. Umapathy thanks DST for the award of a JC Bose fellowship.

Author details
[1]Department of Inorganic and Physical Chemistry, Indian Institute of Science, Bangalore 560012, India. [2]LaserLaB, VU University Amsterdam, Amsterdam, the Netherlands. [3]Department of Instrumentation and Applied Physics, Indian Institute of Science, Bangalore 560012, India.

Received: 1 November 2014 Accepted: 5 May 2015
Published online: 02 June 2015

References

1. Rafferty DW, Koenig JL. FTIR imaging for the characterization of controlled-release drug delivery applications. J Control Release. 2002;83:29–39.
2. Gautam R, Chandrasekar B, Deobagkar-Lele M, Rakshit S, Kumar BNV, et al. Identification of Early Biomarkers during Acetaminophen-Induced Hepatotoxicity by Fourier Transform Infrared Microspectroscopy. PLoS ONE. 2012;7(9), e45521.
3. Gautam R, Samuel A, Sil S, Chaturvedi D, Dutta A, et al. Raman and Infrared Imaging: Applications and Advancements. Curr Sci. 2015;108:341–56.
4. Blout ER, Fields M. Absorption spectra. VII. The infra-red spectra of some nucleic acids, nucleotides, and nucleosides. J Biol Chem. 1949;178:335–43.
5. Diem M, Mazur A, Lenau K, Schubert J, Bird B, et al. Molecular pathology via IR and Raman spectral imaging. J Biophotonics. 2013;6:855–86.
6. Singh B, Gautam R, Kumar S, Kumar BNV, Nongthomba U, et al. Application of vibrational microspectroscopy to biology and medicine. Curr Sci. 2012;102:232–44.
7. Davis R, Mauer LJ. Fourier tansform infrared (FT-IR) spectroscopy: A rapid tool for detection and analysis of foodborne pathogenic bacteria. Curr Res Technol Educ Topics Appl Microbiol Biotechnol. 2010;2:1582–94.
8. Herrero AM. Raman Spectroscopy for Monitoring Protein Structure in Muscle Food Systems. Crit Rev Food Sci. 2008;48:512–23.
9. Gautam R, Vanga S, Madan A, Nongthomba U, Umapathy S. Raman Spectroscopic Studies on Screening of Myopathies. Anal Chem. 2015;87:2187–94.
10. Barret TW, Peticolas WL, Robson RC. Laser-Raman light scattering observations of conformational changes in myosin induced by inorganic salts. Biophys J. 1978;23:349–58.
11. Gautam R, Deobagkar-Lele M, Majumdar S, Chandrasekar B, Victor E, et al. Molecular profiling of sepsis in mice using Fourier Transform Infrared Microspectroscopy. J Biophoton. 2015. doi:10.1002/jbio.201400089.
12. Deming SN. Chemometrics: an overview. Clin Chem. 1986;32:1702–6.
13. Svante W. Chemometrics, why, what and where to next. J Pharm Biomed Anal. 1991;9:589–96.
14. Ritz M, Vaculíková L, Plevová E. Application of infrared spectroscopy and chemometric methods for the identification of selected minerals. Acta Geodyn Geomater. 2011;8:47–58.
15. Horton RB, Duranty E, McConico M, Vogt F. Fourier transform infrared spectroscopy and improved principal component regression (PCR) for quantification of solid analytes in microalgae and bacteria. Appl Spectrosc. 2011;65:442–53.
16. O'Connell ML, Ryder AG, Leger MN, Howley T. Qualitative analysis using Raman spectroscopy and chemometrics: a comprehensive model system for narcotics analysis. Appl Spectrosc. 2010;64:1109–21.
17. Reisner LA, Cao A, Pandya AK. An integrated software system for processing, analyzing, and classifying Raman spectra. Chemometr Intell Lab. 2011;105:83–90.
18. Robin JS, Gavin J, Molly MS. Noninvasive analysis of cell cycle dynamics in single living cells with Raman microspectroscopy. J Cell Biochem. 2008;104:1427–38.

19. Adams MJ. Chemometrics in Analytical Spectroscopy. 2nd ed. Cambridge: The Royal Society of Chemistry; 2004.
20. Mark H, Workman J. Chemometrics in Spectroscopy Elsevier B.V. 2007.
21. Esbensen KH. Multivariate Data Analysis-in practice, 5th Edition. CAMO; 2000.
22. Bocklitz T, Walter A, Hartmann K, Rösch P, Popp J. How to pre-process Raman spectra for reliable and stable models? Anal. Chim Acta. 2011;704:47–56.
23. Srinivasan GK. Vibrational Spectroscopic Imaging for Biomedical Applications, first edition. McGraw-Hill Professional; 2010.
24. Lieber CA, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra. Appl Spectrosc. 2011;57:1363–7.
25. Mosier-Boss PA, Lieberman SH, Newbery R. Fluorescence rejection in Raman-spectroscopy by shifted-Spectra, edge-detection, and FFT filtering techniques. Appl Spectrosc. 1995;49:630–8.
26. Bambery KR, Wood BR, McNaughton D. Resonant Mie scattering (RMieS) correction applied to FTIR images of biological tissue samples. Analyst. 2012;137:126–32.
27. Mohlenhoff B, Romeo M, Diem M, Woody BR. Mie-Type Scattering and Non-Beer-Lambert Absorption Behavior of Human Cells in Infrared Microspectroscopy. Biophys J. 2005;88:3635–40.
28. Burger J, Gelad P. Hyperspectral NIR image regression part I: calibration and correction. J Chemom. 2005;19:355–63.
29. Li S, Dai L. An improved algorithm to remove cosmic spikes in Raman spectra for online monitoring. Appl Spectrosc. 2011;65:1300–6.
30. Schulze G, Jirasek A, Lu ML, Lim A, Turner RFB, Blades MW. Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation. Appl Spectrosc. 2005;59:545–74.
31. Friedrichs MS. A model-free algorithm for the removal of baseline artifacts. J Biomol NMR. 1995;5:147–53.
32. Savitzky MG. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem. 1964;36:1627–39.
33. Wand M, Ripley B. KernSmooth. Functions for Kernel Smoothing for Wand & Jones (1995), R Package Version 2.23-10; 2013.
34. Zhao J, Lui H, McLean D, Zeng H. Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy Appl. Spectrosc. 2007;61:1225–32.
35. Esmonde-White FWL, Esmonde-White KA, Morris MD. Minor Distortions with Major Consequences: Correcting Distortions in Imaging Spectrographs. Appl Spectrosc. 2011;65:85–98.
36. Eilers PHC. Boelens HFM. Baseline correction with asymmetric least squares smoothing: Leiden University Medical Centre Report; 2005.
37. Ramos PM, Ruisanchez I. Noise and background removal in Raman spectra of ancient pigments using Wavelet transform. J Raman Spectrosc. 2005;36:848–56.
38. Zhang Z, Chen S, Liang Y, Liu Z, Zhang Q, et al. An intelligent background correction algorithm for highly fluorescent samples in Raman spectroscopy. J Raman Spectrosc. 2010;41:659–69.
39. Bussian B, Härdle W. Robust Smoothing Applied to White Noise and Single Outlier Contaminated Raman Spectra. Appl Spectrosc. 1984;38:309–13.
40. Randolph TW. Scale-based normalization of spectral data. Cancer Biomark. 2006;2:135–44.
41. Harvey TJ, Hughes C, Ward AD, Faria EC, Henderson A, et al. Clasification of fixed urological cells using Raman tweezers. J Biophotonics. 2009;2:47–69.
42. van Manen HJ, Kraan YM, Roos D, Otto C. Single-cell Raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes. Proc Natl Acad Sci U S A. 2005;102:10159–64.
43. Hartsuiker L, Zeijen NJ, Terstappen LW, Otto C. A comparison of breast cancer tumor cells with varying expression of the Her2/neu receptor by Raman microspectroscopic imaging. Analyst. 2010;135:3220–6.
44. Krafft C, Sobottka SB, Geiger KD, Schackert G, Salzer R. Classification of malignant gliomas by infrared spectroscopic imaging and linear discriminant analysis. Anal Bioanal Chem. 2007;387:1669–77.
45. Bhargava R. Levin IW. Wiley-Blackwell: Spectroscopic Analysis using Infrared Multichannel Detectors; 2005.
46. Lasch P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. Chemometr Intell Lab. 2012;117:100–14.
47. Geladi P, MacDougall D, Martens H. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Appl Spectrosc. 1995;39:377–562.
48. Krafft C, Diderhoshan MA, Recknagel P, Miljkovic M, Bauer M, Popp J. Crisp and soft multivariate methods visualize individual cell nuclei in Raman images of liver tissue sections. Vib Spectrosc. 2011;55:90–100.
49. Martens H, Nielsen JP, Engelsen SB. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Anal Chem. 2003;75:394–404.
50. Wolthuis R, Travo A, Nicolet C, Neuville A, Gaub M, et al. IR Spectral Imaging for Histopathological Characterization of Xenografted Human Colon Carcinomas. Anal Chem. 2008;80:8461–9.
51. Lasch P, Petrich W. Data Acquisition and Analysis in Biomedical Vibrational Spectroscopy. RSC Analytical Spectroscopy Series. 2011;11:192–225.
52. Geladi P. Chemometrics in spectroscopy. Part 1. Classical chemometrics. Spectrochim. Acta B. 2003;58:767–82.
53. Miller JN, Miller JC. Statistics and chemometrics for analytical chemistry. 6th edition, Prentice Hall; 2010.
54. Geladi P, Sethson B, Nyström J, Lillhonga T, Lestander T, et al. Chemometrics in spectroscopy. Part 2. Examples. Spectrochim. Acta B. 2004;59:1347–57.
55. Goel PN, Singh SP, Krishna CM, Gude RP. Investigating the effects of Pentoxifylline on human breast cancer cells using Raman spectroscopy. J. Innov. Opt. Health Sci. 2015;8(1550004):1–11.
56. Mahesh S, Jayas DS, Paliwal J, White NDG. Comparison of Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) Methods for Protein and Hardness Predictions using the Near-Infrared (NIR) Hyperspectral Images of Bulk Samples of Canadian Wheat. Food Bioprocess Tech. 2015;8:31–40.
57. Klein K, Gigler AM, Aschenbrenner T, Monetti R, Bunk W, et al. Label-Free Live-Cell Imaging with Confocal Raman Microscopy. Biophys J. 2012;102:360–8.
58. Ellis DI. Rapid identification of closely related muscle foods by vibrational spectroscopy and machine learning. Analyst. 2005;130:1648–54.

59. Walsh MJ, German MJ, Singh M, Pollock HM, Hammiche A, et al. IR microspectroscopy: potential applications in cervical cancer screening. Cancer Lett. 2007;246:1–11.
60. Bonnier F, Byrne HJ. Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems. Analyst. 2012;137:322–32.
61. Hori R, Sugiyama J. A combined FT-IR microscopy and principal component analysis on softwood cell walls. Carbohyd Polym. 2003;52:449–53.
62. Dupuy N, Duponchel L, Huvenne JP, Sombret B, Legrand P. Classification of edible fats and oils by principal component analysis of Fourier transform infrared spectra. Food Chem. 1996;57:245–51.
63. Pichardo-Molina JL, Frausto-Reyes C, Barbosa-García O, Huerta-FrancoR, González-Trujillo JL, *et al.* Raman spectroscopy and multivariate analysis of serum samples from breast cancer patients. Lasers Med. Sci. 2007;22 http://rd.springer.com/article/10.1007/s10103-006-0432-8.
64. Parker FS. Applications of infrared, Raman, and resonance Raman spectroscopy in biochemistry. Springer; 1983
65. Vandenabeele P. Raman spectroscopy. Anal Bioanal Chem. 2010;397:2629–30.
66. David EI, Goodacre R. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. Analyst. 2006;131:875–85.
67. Krafft C, Steiner G, Beleites C, Salzer R. Disease recognition by infrared and Raman spectroscopy. J Biophotonics. 2009;2:13–28.
68. Haka AS, Shafer-Peltier KE, Fitzmauric M, Crowe J, Dasari RR, Feld MS. Diagnosing breast cancer by using Raman spectroscopy. Proc Natl Acad Sci U S A. 2005;102:12371–6.
69. Mahadevan-Jansen A, Richards-Kortum RR. Raman spectroscopy for the detection of cancers and precancers. J Biomed Opt. 1996;1:31–70.
70. Steiner G, Shaw A, Choo-Smith LP, Abuid MH, Schackert G, et al. Distinguishing and Grading Human Gliomas by IR Spectroscopy. Biopolymers. 2003;72:464–71.
71. Ghita A, Pascut FC. Cytoplasmic RNA in Undifferentiated Neural Stem Cells. A Potential Label-Free Raman Spectral Marker for Assessing the Undifferentiated Status. Anal Chem. 2012;84:3155–62.
72. Sikirzhytski V, Virkler K, Lednev IK. Discriminant Analysis of Raman Spectra for Body Fluid Identification for Forensic Purposes. Sensors. 2010;10:2869–84.
73. Heraud P, Wood BR, Beardall J, McNaughton D. Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells. J Chemom. 2006;20:193–7.
74. Muik B, Lendl B, Molina-Díaz A, Ortega-Calderón D, Ayora-Cañada MJ. Discrimination of olives according to fruit quality using Fourier transform Raman spectroscopy and pattern recognition techniques. J Agr Food Chem. 2004;52:6055–60.
75. Krafft C, Shapoval L, Sobottka SB, Geiger KD, Schackert G, Salzer R. Identification of primary tumors of brain metastases by SIMCA classification of IR spectroscopic images. BBA-Biomembranes. 1758;2006:883–91.
76. Podshyvalov A, Sahu RK, Mark S, Kantarovich K, Guterman H, et al. Distinction of cervical cancer biopsies by use of infrared microspectroscopy and probabilistic neural networks. Appl Opt. 2005;44:3725–34.
77. Kneipp J, Beekes M, Lasch P, Naumann D. Molecular Changes of Preclinical Scrapie Can Be Detected by Infrared Spectroscopy. J Neurosci. 2002;22:2989–97.
78. Effendi W, Zheng W, Huang Z. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. Int J Oncol. 2008;32:653–62.
79. Brudzewski K, Kesik A, Kołodziejczyk K, Zborowska U, Ulaczyk J. Gasoline quality prediction using gas chromatography and FTIR spectroscopy: An artificial intelligence approach. Fuel. 2006;85:553–8.
80. Rösch P, Harz M, Schmitt M, Peschke KD, Ronneberger O, et al. Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations. Appl Environ Microb. 2005;71:626–1637.
81. Miljkovic M, Chernenko T, Romeo MJ. Label-free imaging of human cells: algorithms for image reconstruction of Raman hyperspectral datasets. Analyst. 2010;135:2002–13.
82. Lasch P, Haensch W, Naumann D, Diem M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. BBA-Mol Basis Dis. 2004;1688:176–86.
83. Maimon OZ. Lior R (Eds). Data mining and knowledge discovery handbook: Springer; 2005.
84. Steller W, Einenkel J, Horn LC, Braumann UD, Binder H, et al. Delimitation of squamous cell cervical carcinoma using infrared microspectroscopic imaging. Anal Bioanal Chem. 2006;384:145–54.
85. Fabian H, Lasch P, Boese M, Haensch W. Infrared microspectroscopic imaging of benign breast tumor tissue sections. J Mol Struct. 2003;661–662:411–7.
86. Dinh PX, Beura LK, Das PB, Panda D, Das A, Pattnaik AK. Induction of Stress Granule-Like Structures in Vesicular Stomatitis Virus-Infected Cells. J Virol. 2013;87:372–83.
87. Mansfield JR, Sowa MG, Scarth GB, Somorjai RL, Mantsch HH. Analysis of spectroscopic imaging data by fuzzy C-means clustering. Anal Chem. 1997;69:3370–4.
88. Paradkar MM, Joseph I. Rapid determination of caffeine content in soft drinks using FTIR–ATR spectroscopy. Food Chem. 2002;78:261–6.
89. Berger AJ, Itzkan I, Michael SF. Feasibility of measuring blood glucose concentration by near-infrared Raman spectroscopy. Spectrochim Acta A. 1997;53:287–92.
90. McLaughlin G, Doty KC, Lednev IK. Discrimination of human and animal blood traces via Raman spectroscopy. Forensic Sci Int. 2014;238:91–5.
91. Lui H, Zhao J, McLean D, Zeng H. Real-time Raman Spectroscopy for In Vivo Skin Cancer Diagnosis. Cancer Res. 2012;72:2491–500.
92. Liakat S, Gmachl CF, Michel AP, Bors K. Noninvasive mid-infrared in vivo glucose sensor. 2014;U.S. Patent Application 14/470,386.
93. Stone N, Kendall C, Shepherd N, Crow P, Barr H. Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers. J Raman Spectrosc. 2002;33:564–73.

94. Molckovsky A, Kee-Song LW, Shim MG, Marcon NE, Wilson BC. Diagnostic potential of near-infrared Raman spectroscopy in the colon: differentiating adenomatous from hyperplastic polyps. Gastrointest Endosc. 2003;57:396–402.
95. Dingari NC, Barman I, Saha A, McGee S, Galindo LH, et al. Development and comparative assessment of Raman spectroscopic classification algorithms for lesion discrimination in stereotactic breast biopsies with microcalcifications. J Biophotonics. 2012;6:371–81.
96. Soares JS, Barman I, Dingari NC, Volynskaya Z, Liu W. Diagnostic power of diffuse reflectance spectroscopy for targeted detection of breast lesions with microcalcifications. Proc Natl Acad Sci U S A. 2013;110:471–6.