Quantitative Tablet Characterization Based on Multi-component Image Analytics & Pattern Matching

Subhrangshu Bit

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah
Pin - 711 202, West Bengal

A thesis submitted to
Ramakrishna Mission Vivekananda Educational and Research Institute
in partial fulfillment of the requirements for the degree of
MSc in Big Data Analytics
2018

Dedication

Dedicated to my parents, who instilled in me the virtues of perseverance and commitment and relentlessly encouraged me to strive for excellence.

Acknowledgements

I take this opportunity to express my hearfelt gratitude to the supervisor of the work, Mr. Nishit Mittal, Engagement Lead — Data Science, DPEX, Dr. Reddy's Laboratories for his wise guidance and incessant ecouragement throughout the entire period of work.

The cooperation and help from the professors of Department of Computer Science, Ramakrishna Mission Vivekananda Educational and Research Institute is thankfully acknowledged.

I am specially thankful to Swathy Prabhu Maharaj, Head, Department of Computer Science and Mrinmay Maharaj, Lecturer, Ramakrishna Mission Vivekananda Educational and Research Institute, for their constant support and arrangements to guide my transition into the corporate environment from academics.

I extend my sincere thanks to Chandrakant Saha, my mentor in Dr. Reddy's Laboratories, for his active help in this project.

Finally, I would like to express my best regards to my parents and other elders of my family for their ceaseless blessings and inspiration.

Ramakrishna Mission Vivekananda Educational and Research Institute, Belur Math, West Bengal

Subhrangshu Bit

July 21, 2021

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled *'Quantitative Tablet Characterization Based on Multi-component Image Analytics & Pattern Matching'* submitted by *Mr. Subhrangshu Bit*, who has been registered for the award of MSc in Big Data Analytics degree of Ramakrishna Mission Vivekananda Educational and Research Institute, Belur Math, Howrah, West Bengal is absolutely based upon his own work under the supervision of myself and that neither his thesis nor any part of the thesis has been submitted for any degree or any other academic award anywhere before.

Mr. Nishit Mittal
Engagement Lead - Data Science
Digital and Process Excellence
Dr. Reddy's Laboratories
Bachupally, Hyderabad, 500090, Telengana

Abstract

In recent years, Raman spectroscopy has become very popular due to the advancement in instrumentation have allowed us to focus on their application rather than on the operation and limitations of the instrument. The large scale information provided by a single Raman spectrum includes the molecular structure, qualitative and quantitative information of the analyte. This work leverages this data to quantify the amount of constituents in the analyte and also simultaneously generate a spatial distribution of the same thereby providing a agile reverse engineering process. The study is based on simple linear models and matrix computations like linear regression, multivariate curve resolution which have been moulded according to the requirement of the problem. We further use statistical F-test and statistics such as R^2 in order to evaluate the results and filter the unwanted from the data. The proposed pipeline relies on the basis of Beer-Lambert law however without any reliability on any form of calibration data. Moreover, our framework allows the quantification and spatial distribution to be specific to a particular layer or region of the analyte. All the work will be consolidated into an UI enabling users to analyze with a few clicks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years the pharmaceutical industry has been obliged to comply with increasingly stringent regulatory requirements enforced by the Food and Drugs Administration (FDA). These requirements have been enforced in order to ensure that the risks associated with pharmaceutical products to public health are minimized. The FDA has encouraged pharmaceutical companies to invest resources into developing advanced methodologies that improve their understanding and subsequently the control of their manufacturing process.

In order to improve the quality of their products and understand the characterizations a step deeper a number of pharmaceutical companies have embraced the use of various spectroscopic imaging techniques, such as near infrared (NIR), infrared (IR) and Raman. These techniques are being explored as potential methods for advanced characterization of quality of products [31], [13]. Amongst several evolving applications of spectroscopy in the pharmaceutical industry include producing chemical images of constituent components on the surfaces of the tablets [17] and determination of content uniformity [11].

Raman spectroscopy has become more popular in recent years because improvements in instrumentation have allowed us to focus on their applications rather than on the operation and limitations of the instrument. A single Raman spectra can provide a large scale information about the sample. Its many well-resolved spectral features often provide good specificity for qualitative analysis and good analyte selectivity for quantitative analysis. Raman spectroscopy is more often associated with the determination of molecular structure and with qualitative analysis than it

is with quantitative analysis. Applications of Raman spectroscopy for quantitative analysis of sample composition have, however, been described extensively in technical literature. The ability of Raman spectra to uniquely identify and generate a fingerprint corresponding to each of the components encourages it's application into newer heights. The results of quantitative analysis is a value specifying the amount of something in a sample of interest, an estimate of uncertainty in that value and if possible, independent information that tests the validity of the attained results. The trend towards automating the entire process without any human intervention in order to address very large data sets and to reduce the impact of operator bias on the analytical results places greater importance on robustness and validity testing. A computer aided statistical/mathematical analysis in the absence of Raman experts is the final goal.

Since the introduction of the mathematical formalization and information theory over five decades ago, scientists have found applications of the theory in many diverse fields of science and technology. Various methods developed have proved to be particularly powerful when applied to model instrumental measurements [3], [9]. The use of information theory for analytical chemistry has been the subject of explicit development since the early 1970s. Methods for determining the chemical composition of various materials and composites were developed using signals bearing such information. This lead to the generation of an amalgam of analytical chemistry and statistical/mathematical analytics popularly known as chemometrics. With the advent of data science the field of chemometry has been enriched further with more data driven models generated from spectroscopy. The role of chemometrics coupled with data science in the analysis of Raman spectroscopy data is becoming increasingly important for many different application areas. It allows the quantification and qualification of very complex systems and the usage of developed multivariate techniques ensures full exploitation of the entire spectral data unlike univariate techniques involving much human intervention and feature engineering.

Keeping in mind the above the project aims to build a multi-component image analytics platform that can identify, discriminate & quantify fraction of each component in a tablet using Chemical Imaging. Hence providing a spatial distribution of the various components and particle statistics to understand the formulation and process design space.

Chapter 2

Literature Review

The phenomenon of inelastic light scattering is known as Raman radiation and was first documented by Raman and Krishnan in 1928 [28]. Raman spectroscopy is becoming one of the most popular analytical measurement tools for pharmaceutical applications ranging from verification of raw materials to process monitoring of drug production to quality control of products. Similar to infrared spectrum, Raman spectrum consists of a wavelength distribution of peaks corresponding to molecular vibrations specific to the sample being analyzed. Chemicals, such as drugs, can be identified by the frequency and quantified by the intensity of the peaks.

2.1 Analysis of pharmaceutical formulations

Several pharmaceutical forms have already been studied by Raman spectroscopy. One of the cited useful features of Raman spectroscopy is its ability to carry out direct measurement in solids. Among the parameters that influence the intensity of the Raman signal detected from solids are particle size and packing density.

Rodriguez [29] has described various experimental Fourier Transform (FT) Raman imaging procedures and their ability to both obtain and spatially resolve chemical information in the analysis of formulated tablets of pharmaceutical interest. Experimental analytical procedures using the imaging techniques are outlined.

Breitenbach et al. [5] used con-focal Raman spectroscopy to examine solid dispersion of the anti-inflammatory agent ibuprofen. The group investigated the physiochemical stability of the formulation under stress conditions together with the content and the homogeneity of the drug distribution in the formulation matrix. The

method was found to be promising for monitoring the spatial distribution of drugs in solid dispersion. The authors have stated that Raman spectroscopy can investigate different layers (e.g. coatings on a tablet), areas or simply the quality of mixing in a manufacturing process, which is of great industrial importance.

It is very common to organize and archive Raman spectroscopy data as a 3-dimensional hyperspectral image. Analysis of such huge hyperspectral images can be time consuming and rigorous. Andrew et al. [2] have described two curve resolution methods namely Principal Factor multivariate curve resolution (PF-MCR) and orthogonal projection multivariate curve resolution (OP-MCR) for analysis of three-way Raman image data. The results from MCR analysis using either method provides the number of chemical species present in the sample, the spectrum of each species for identification, and the concentration image for each species. A discussion is given addressing rapid analysis aspects of OP-MCR and the relative merits and drawbacks of the technique in comparison to PF-MCR.

Findlay and Bugay [12] described how variable temperature (VT)-Raman spectroscopy can be used to study the dynamics of crystallization of menthol from a solvent (ethanol). The authors stated that Raman spectroscopy can also be applied as a quantitative technique, but some criteria must first be considered critically, e.g. homogeneous sample mixing, particle size, and instrument variability and reproducibility. They concluded that Raman spectroscopy can be used in the pharmaceutical analytical laboratory in a variety of ways. Traditional drug substance characterization is enhanced with additional information provided by Raman spectral data, and quantitative polymorph assays can be developed. Raman spectroscopy can also be used qualitatively and semi-quantitatively to support pharmaceutical development.

According to Langkilde et al. [20], [21], differences can be seen between Raman spectra from different crystal forms of a compound, or between crystalline and amorphous forms. These investigators showed that the possibility of minimal sample preparation and the sensitivity to polymorphism make Raman spectrometry ideal for the study of crystal forms of pharmaceutical compounds, as they observed different FT-Raman spectra from the two polymorphs of a compound. They identified a frequency shift that leads to well resolved bonds and found that, for mixtures of A and B, intensity of the two bonds were proportional to the amount of the A and B forms.

2.2 Quantitative/Qualitative Analysis Approaches

2.2.1 Pre-processing

Data processing prior to univariate/multivariate modelling is a necessary and important step. Pre-processing is primarily required to eliminate effects of unwanted signals such as fluorescence, detector noise, calibration errors, cosmic rays, laser power fluctuations, signals from glass substrate etc. and also to enhance subtle differences between different samples [4].

The cosmic spike elimination from the raw spectrum is generally 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 compared with genuine Raman bands. Usually, local interpolation based methods are used to repair spike affected regions [23]. Whitaker et al. [33] have presented a despiking algorithm based on the calculation of modified Z scores (based on median of first order detrended spectra) to locate spikes and a simple moving average filter to remove the located spikes. They have stated that the algorithm is computationally efficient and inexpensive compared to collection of multiple spectra for locating spikes.

Background correction/baseline removal is a very important part of pre-processing. Various phenomenon explained such as fluorescence etc. induce uneven amplitude shifts across different wavenumbers. These amplitude shifts have to compensated before further analysis. In literature many such techniques have been compared and evaluated in detail [4]. Some of the common methods employed for baseline removal are:

- a) Median based Window Methods
  A moving window based method where at each point only a few intensity values (length of the window) in it's neighbourhood are used to estimate the baseline value at that point. The median of such a local window of intensity values at each point is calculated first, followed by convolving with a Gaussian function to make sure that the estimated baseline is free from sharp discontinuities [14].

- b) Derivative based Methods
  In general, the baseline has broad bands and low frequency components compared to genuine Raman bands. Derivative (numerical) of the Raman signal

amplifies the higher frequency components whereas the lower frequency components such as the background fluorescence is suppressed.

- 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. Initially polynomial fitting typically required user intervention and thus is time consuming and prone to variability. However, Lieber et al. [24] automated the method of fluorescence subtraction based on a modification to the leasts-squares polynomial curve fitting. Zhao et al. [35] further improved this technique with the addition of a peak removal procedure during the first iteration and a statistical method to account for signal noise effects thereby improving the performance in real-time in vivo applications and low signal-to-noise ratio environments. This was further worked upon to completely remove any form of user intervention. Zhang et al. [34] developed a novel algorithm named adaptive iteratively reweighted Penalized Least Squares (airPLS) which works by iteratively changing weights of sum squares errors (SSE) between the fitted baseline and original signals, and the weights of the SSE are obtained adaptively using the difference between the previously fitted baseline and the original signals.

Although baseline removal eliminates the effects of large band or low frequency components in the Raman signal, it still suffers from high frequency component and needs to be removed. 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 may disappear. Therefore, it is very important to choose an appropriate polynomial order and moving window size to retain all the important Raman bands.

2.2.2 Univariate Data Analysis

The goal of univariate data analysis is to find a mathematical relationship between the metric extracted from a Raman spectra, such as peak width, peak intensity etc. and the desired property, such as analyte concentration. Several articles have compared and analyzed the results of univariate techniques for quantification of Raman spectra [19], [27].

Ivana Durickovic [8] has described a univariate methodology for analysis of a particular analyte. The author has described that the position of the peak defined by its maximum corresponds to the vibration frequency of the chemical species. Since each chemical bond has its own characteristic vibrations, the position of the peaks lead to the identification of the chemical species. The peak intensity is related to the corresponding chemical species concentration. In order to determine this parameter, it is necessary to use normalization of the integrated intensity of the Raman line as the peak intensity is also sensitive to the laser power.

The mathematical relationship between the Raman metric and the desired property of the analyte is called an analytical model. While it is theoretically possible to calculate the analytical model, in practice several necessary constants such as absolute Raman cross-section and optical collection efficiency are rarely known. As a result, analytical models are almost always created by measuring Raman spectra of known samples (the training set) and empirically relating the Raman metric to the known property [25].

A novel method consisting of automatic decomposition of Raman spectra and a model for quantitative analysis was developed for the analysis of components of natural gas in [26]. In this work the concentration of the unknown component was determined using the area of the vibration peaks in the spectra. However, the quantification of components in this pipeline requires a calibration set which is computationally expensive.

2.2.3 Multivariate Data Analysis

The spectroscopic data can be displayed in the forms of a matrix described in detail in 4.1 in Chapter 4. The spectroscopic measurements consists of two parts -

$$Observation = Relevant\ \ Signal + Noise$$

Here, the relevant signal is considered as the actual representation of the underlying chemical information, which is correlated with the property of interest. There are many multivariate data analysis techniques available and for an appropriate selection the goal of the analysis should be clearly defined. The three major objectives for analysis are -

- Explorative data structure modelling and dimensionality reduction. Principal Component Analysis (PCA) is frequently is used for this purpose.
- Discrimination, Classification, Clustering deal with dividing a data matrix into two or more group of measurements.
- Regression and prediction for quantifying a set of variables with respect to another.

Unlike classification or clustering since our primary goal is to determine the quantity of components in a sample, we focus on the existing literature in multiple linear regression based data analysis techniques. Two matrices are used. in general, X representing the hyperspectral image data and Y represents the pure spectra of the dependent variables i.e. the components. In case of multiple linear regression, there is an assumption of linear dependency of the regressor variables on the independent variables. The regression of X is performed on Y using the least squares criterion [6].

A major caveat arises due to high correlation between the independent variables i.e. strong correlation between the spectra of the different components may cause an ill-conditioned least squares problem. Principal Component Regression (PCR) mitigates this problem by subjecting the independent variables Y through a dimensionality reduction such as PCA and then performing regression of X on the decomposed matrix [15], [16]. However, 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 the covariance between the predictor and response variables are not taken into account during PCA.

Partial Least Squares (PLS) is an improvement over both PCR and multiple linear regression overcoming their limitations. Unlike PCR, instead of the covariance between the dependent variables X^TX the covariance between the predictor and response variables X^TY is subjected to singular value decomposition [10], [15], [16]. The method can quite effectively handle one or more co-varying dependent variables

by projecting both X and Y into latent spaces T and U, 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 is also learned between T and U. However, PLS requires the number of samples/observations in X and Y to be same, which is not the case in our problem and thus fails its application.

Multivariate Curve Resolution (MCR) is the generic denomination of a family of methods meant to solve the mixture analysis problem, i.e., able to provide a chemically meaningful bilinear model of pure contributions from the sole information of an original data matrix including a mixed measurement [30], [22]. Traditionally, MCR was conceived for evolutionary analytical data coming from a process or an analytical measurement [22]. Many analytical measurements, particularly all of those based on spectroscopic methods are particularly suited to be analyzed by MCR, since the underlying analytical model (known as Beer-Lambert law) is formally a bilinear model of pure signal contributions. The application of MCR has grown significantly, and the fields of application have expanded in complexity and diversity. Within the spectroscopic field, the structure in the concentration direction is no longer a requirement. This has allowed analysis of hyperspectral images, in which the image cube has two spatial dimensions and one spectroscopic dimension. To be unmixed, a previous unfolding of the image cube, into a data matrix with rows designating the pixel spectra and columns designating the spectral channels measured, must be carried out. After MCR, an operation of refolding the concentration profiles is needed to recover the spatial structure of the distribution maps of the pure image components [18], [7], [1].

Chapter 3

Methodology

We have developed a multi-level data analysis pipeline that can be curated according to the necessity and demand of the user to generate intensive results and insights. The input 3D hyperspectral cube containing the Raman spectroscopy data is passed through a few pre-processing steps referred in Chapter 2, following which we perform and allow the usage of multiple univariate and multivarite techniques to generate a spatial distribution of the sample of interest along with the relative composition of the components in the sample. In order to gain multiple insights we also allow an interactive 3D visualization of the sample tablet which incorporates the selection of region of interests (ROI) and further extend the study of quantification and spatial distribution focused only in the selected ROI.

3.1 Pre-processing Pipeline

The input 3D hyperspectral cube is analyzed in reference to the spectral library containing the pure spectra of all the components that the sample is composed of. However, since the acquisition of the pure spectra of the components and that of the sample spectra are under varying conditions the spectral length and wavenumbers of the incident laser differ. In order to mitigate this and bring the spectra under similar format both the pure spectra and the acquired spectra are truncated and interpolated within a range provided by the domain expert or the user. In order to perform interpolation we utilize the nearest neighbor algorithm to fill out the missing positions. A typical acquired spectra before and after truncation and interpolation is shown in 3.1. The spectral intensities have also been normalized using the Min-

Max normalization technique [Chapter 2]. However, we also provide the option of normalization using the technique of standard normal variate normalization (SNV).

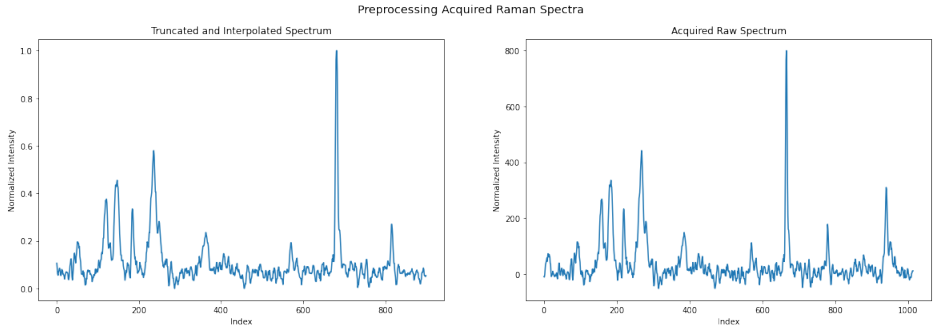

Figure 3.1: Truncated and Interpolated acquired Raman spectra

Prior to addressing the undesirable frequency components in signal we remove the spikes, caused by cosmic rays, using the method of modified Z-scores as described elsewhere [33]. The acquired Raman spectra contains a lot of unwanted high and low frequency components that can cause hindrance to further analysis. Thus we primarily remove the baseline using the method of airPLS [34] since it requires the least amount of human intervention. In order to ensure flexibility, the domain expert or the user can also opt for polynomial fitting techniques [24] and [35]. Following the removal of the low frequency component we then smoothen the high frequency component using the SG filter [32] with a default setting of a window length of 5 consecutive points and the local polynomial order to be 2. A fully preprocessed Raman spectra is shown in 3.2

3.2 Univariate Analysis

The pure spectra is then analyzed by the domain expert to specify the unique peak determining the characteristic of a component. Provided a range of wavenumbers we quantify the peak intensity [Figure:3.3] and the peak area [Figure:3.4] of the Raman spectra at every spatial position along the cross section of the sample. The quantified intensity/area values are then mapped onto a blank canvas containing the white light image of the sample as watermark with the gradient of the color proportional to the value. Generating such maps of all the components that the sample is composed of we overlay all on top of each other to generate a visual representation of the spatial distribution of the components.

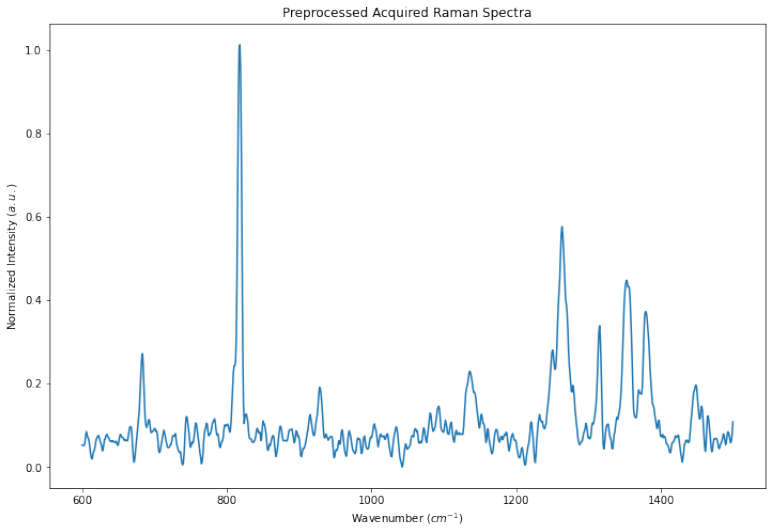

Figure 3.2: Acquired Raman spectra preprocessed using airPLS baseline removal and SG filter smoothening

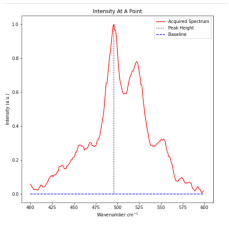

Figure 3.3: Intensity At A Point

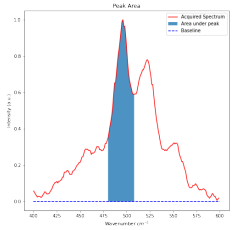

Figure 3.4: Peak Area

In terms of quantification we add the computed intensity/area values for each component separately and generate a relative distribution. However, since even after pre-processing some amount of noise is inherent, we allow the user to filter the intensity/area values further below or above a specified threshold. A good specification of the threshold can generate almost accurate relative quantification of the composition.

3.3 Multivariate Analysis

In order to overcome the problems in univariate methods we include a few existing multivariate methodologies, which has already been described in Chapter 2. We primarily approach in two ways:

3.3.1 Pixel-wise Linear Regression

The idea of linear regression is based on Beer-Lambert's law. The law says that at any given wavelength i, the light absorbance (A) is proportional to the absorbance coefficient of the pure substance (k) at the chosen wavelength i and the concentration of the pure substance (c):

$$A_i = k_i * c \tag{3.1}$$

where the absorbance coefficient ($k_i = a_i * L$) is the product of the path length of the light through the material (L) and the molar attenuation coefficient or absorptivity (a) of the pure substance at the chosen wavelength i, which is a molecular property constituting the 'spectrum' of such molecule.
When there are multiple absorbing components, the total absorbance at any wavelength is the sum of the absorbances, at that wavelength, of all the components in the mixture:

$$A_i = \sum_{j=1}^{j=n} k_{ij} * c_j \tag{3.2}$$

The above can be represented in matrix form as follows:

$$A = Kc + e \tag{3.3}$$

Therefore having the values of the acquired spectra A, and knowing the ones of the pure component spectra making up the mixture K, we can find the concentrations that best determine the acquired spectra. Due to noise it is not possible to obtain the exact solution to the equation, we minimize the sum of squares of error i.e. the Euclidean norm of the error:

$$||A - Kc||_2^2 \tag{3.4}$$

For each component we have a spectral signature as our regressor/independent variable-

$$\mathbf{x} = \begin{bmatrix} intensity_{wv_1} \\ intensity_{wv_2} \\ \vdots \\ intensity_{wv_p} \end{bmatrix} \tag{3.5}$$

The dependent/regressed variable is the acquired spectra at a particular pixel location:

$$\mathbf{y} = \begin{bmatrix} acq_{wv_1} \\ acq_{wv_2} \\ \vdots \\ acq_{wv_p} \end{bmatrix} \tag{3.6}$$

We then perform simple linear regression and try to estimate the regression coefficient:

$$\mathbf{y} = \beta_0 + \beta_1 * \mathbf{x} + \epsilon \tag{3.7}$$

Each pixel of a hyperspectral image has a spectral dimension containing the Raman spectra at that position. Each such spectra is assumed to be generated by a linear combination of the components that constitute the sample. Thereby, treating each of the Raman spectrum as the regressed variable and the pure spectrum of the components as the predictor variables we perform ordinary least squares based linear regression.

At every pixel we store the adjusted R^2 value along with the p-value of the F-test used for testing the significance of the regression. The coefficients $\hat{\beta}$ are also stored corresponding to each pixel. It is important to note that we did not use any bias in the linear model, since it may lead to some discrepancies, which is discussed in Chapter 4.

3.3.2 Matrix Factorization

Pixel-wise linear regression takes a repetitive approach of performing linear regression at every spatial position, however in contrast to that we leverage the idea of multivariate curve resolution of matrix factorization with the added information of the pure spectral signatures of the constituent components. In general, multivariate curve resolution methods permit, with no prior knowledge, to extract simultaneously the spectra of the pure products and their corresponding concentration maps from the experimental data matrix 3.5.

Unlike the univariate methods, there is no need to specify the characteristic spectral zone. We have started by unfolding the hyperspectral cube D. The first unfolding is necessary due to the two way character of the data though the data is stored in three dimensions. Then a classical line by line unfolding method of the spectral data matrix is used in our study. The x first spectra corresponding to the x

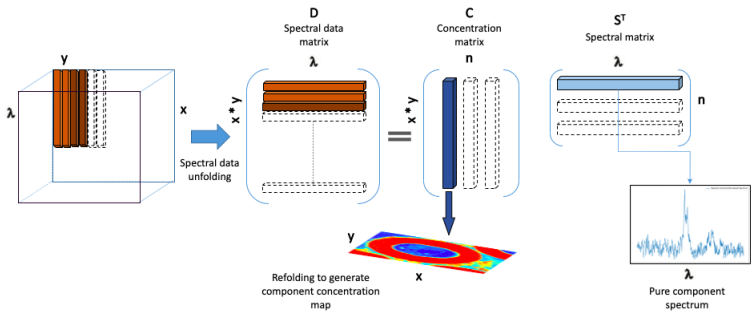

Figure 3.5: Multivariate curve resolution method applied to Raman spectroscopy data.

first pixels of the analyte are placed along the x first lines of the matrix D and so on. Since in our study the number of constituents of the analyte and the pure spectral signatures of the same are known apriory, the spectral matrix S^T and the rank of the matrices C and S are known. The only unknown is the concentration matrix C. Thus, given the design matrix (spectral matrix) S and the unfolded acquired data matrix D the goal is to find the abundance maps C. This is approached as a classical least squares problem by minimizing the sum of squares of errors:

$$argmin_C||D - CS^T||_2^2 \tag{3.8}$$

However, in order to ensure that the concentration matrix C is devoid of any non-negative values we include a contraint and optimize the contrained problem:

$$\begin{gathered} argmin_C||D - CS^T||_2^2 \\ \text{subject to} \\ C_{ij} \geq 0 \quad \forall(i,j) \end{gathered} \tag{3.9}$$

Moreover, since at every pixel the sum of proportions of the constituents are expected to sum to one we further add a constraint the finally optimize the following problem

$$\begin{gathered} argmin_C||A - CS^T||_2^2 \\ \text{subject to} \\ C_{ij} \geq 0 \ \ \forall(i,j) \\ DC = \mathbf{1} \end{gathered} \tag{3.10}$$

Chapter 4

Experimental Evaluation

4.1 Description of Data

We keep the names of the sample tablets and their composition undisclosed owing to confidentiality of the research.

In general, human eye perceives its environment with a colourful rendering. Instead of a single grey-scale image, we visualize a combination of 3 channels red-scale, green-scale and blue-scale. The combination of these three channels generate a plethora of many possible colours and thus an image can be represented as a 3D matrix with ($n_{rows} \times n_{columns} \times 3_{channels}$). However with the advancement in imaging technology we can generate images with and get information regarding wavelengths beyond the visible spectrum (red, green blue) and make sense out of it. In a broad sense, spectral imaging is the parallel acquisition of spatial and their corresponding spectral information in an image space and their combination thereof. The image generated through such spectral imaging techniques are known as an hyperspectral images shown in 4.2.

The hyperspectral images specific to our study is generated from Raman spectroscopy on a sample cross-section of a tablet. We have the hyperspectral images corresponding to three major tablets with the following specifications:

- Product 1 :
  Dimensions: $641 \times 711 \times 1015$
  Wavenumber cm^{-1}: The spectral dimension range from 613.94 to 1723.53

- Product 2 :

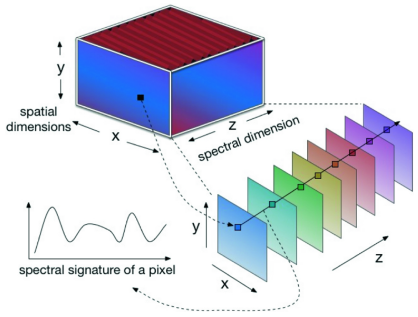

Figure 4.1: An exploration of the hyperspectral image. x and y represent the spatial dimensions containing the frame of the analyte. z represents the channels corresponding to varying wavenumbers of the incident laser.

Dimensions: $254 \times 304 \times 1015$

Wavenumber cm^{-1}: The spectral dimension range from 616.16 to 1721.79

- Product 3 :

  Dimensions: $382 \times 589 \times 1039$

  Wavenumber cm^{-1}: The spectral dimensions range from 482.35 to 1641.35

The format of the data was '.wdf' which is the extension of the file generated by WiRE software associated with the spectrometer manufactured by Renishaw. Along with the spectral data as a 3D hypercube a number of attributes of the file are taken into account for better understanding and visualization. The white light image of the sample containing the cross-section of the tablet is extracted from one of these attributes. Also the wavenumbers corresponding to the spectra is extracted as a separate vector. However, most of the studies associated with quantification of components from Raman spectroscopy are based on calibration data and thus require spectral data of the same sample with varying compositions. We on the other hand attempt to quantify based on a single sample data using the library containing the pure spectra of all the components that the sample is composed of.

4.1.1 Data organization

The extracted 3D spectral hypercube and the library containing pure components' spectra are then organized in a format to enable further statistical analysis as shown in 4.2. The hyperspectral image is read into a 3D matrix with the dimension $x_{pos} \times y_{pos} \times wavenumber$, x and y represent the spatial dimensions and wavenumber represents the spectral dimension. The spectra corresponding to the components in

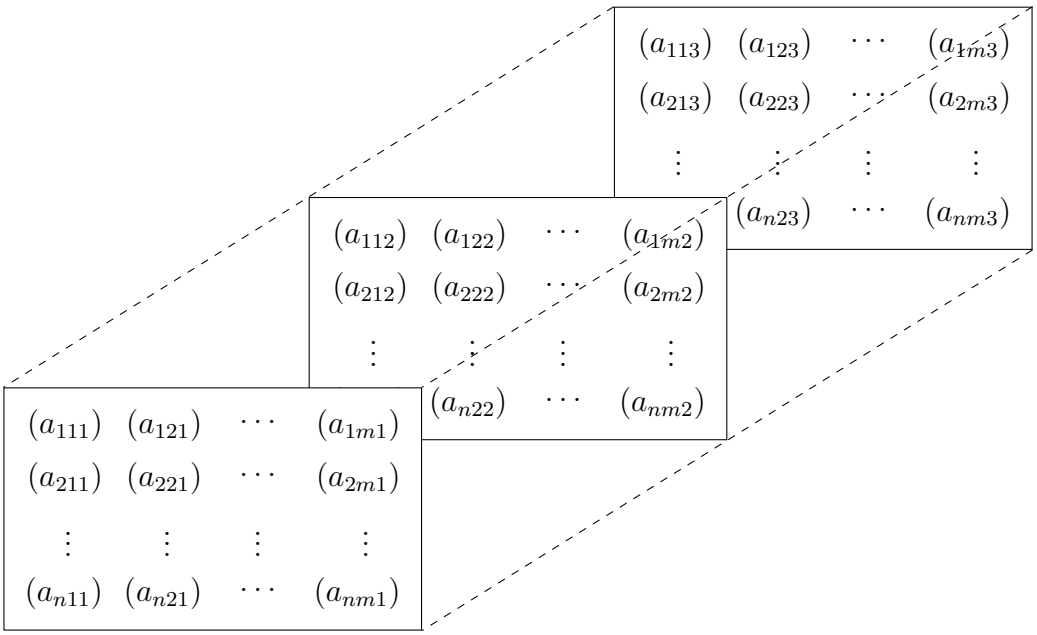

Figure 4.2: 3D matrix containing a hyperspectral image. Each channel represent a particular wavenumber.

pure form are extracted from the library and stored onto a 2D matrix with dimensions $n_{components} \times wavenumber$.

4.2 Evaluation Techniques

In order to evaluate the results of the analysis a multi-step technique was followed. In the case of univariate methods the initial form of inspection of the abundance maps was visual verification by the domain experts. Following which an interactive platform was developed where clicking on a particular position of the abundance map generates the spectrum corresponding to the nearest pixel and plots the same against the pure spectral signatures of the constituent components 4.3.

Since the acquired spectra is a linear combination of the constituents' spectra, one can clearly visualize if the characteristic peak of a pure signature matches that of the acquired spectrum. Since each of the component is mapped to a different color this technique allows one to ensure that the abundance maps are correct. However, clicking at every pixel position and visually verifying of the spectra matches can be very tedious. Thus, we come up with the use of statistics while analyzing multivariate methods.

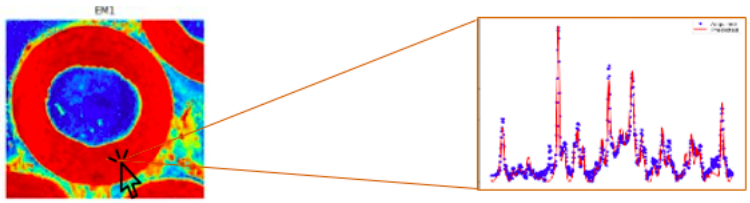

Figure 4.3: Clicking at a particular point on the abundance map generates the spectra corresponding to the pixel at that location against the pure spectral signature of the constituent component of choice

After pixel wise linear regression we use the F-test to ensure that the regression is significant at every position.

- H_0: Regression is insignificant - $\beta_1 = 0$
- H_1: Regression is significant - $\beta_1 \neq 0$

We have n observations (wavenumber and intensity pairs) for each pixel and the number of regression parameters is 1, since we are regressing only with respect to one component. Then

- Sum of Squares for Model: $SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Sum of Squares for Error: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Sum of Squares Total: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Degrees of Freedom for Model: $DFM = 1$
- Degrees of Freedom for Error: $DFE = n - 1$
- Total Degrees. of Freedom: $DFT = n$
- Mean of Squares for Model: $MSM = SSM/DFM$
- Mean of Squares for Error: $MSE = SSE/DFE$
- Mean of Squares Total: $MST = SST/DFT$

The test statistic for our hypothesis is:

$$F = \frac{MSM}{MSE} = \frac{\text{explained variance}}{\text{unexplained variance}}$$

We use the p-value generated by the 'statsmodel' package of python to reject the null hypothesis if the p-value is greater than α and fail to reject if lesser than the same. In our study we have preset the value of α as 0.05, which can be changed by the user.

Following the F-test we use the value of R^2 in order to filter the pixels that have very low R^2 for a particular component. This filtering ensures that the pixel retained has sufficient statistical evidence to be considered as a component. Further this is again visually verified using the technique described above in univariate methods.

4.3 Analysis of Results

4.3.1 Univariate Intensity Map

The intensity corresponding to the peak specified by the user is mapped onto the white light image of the sample to generate a spatial distribution of the component over the ROI. Figure 4.4 shows the intensity map of a component on a portion of the middle layer of a tablet.

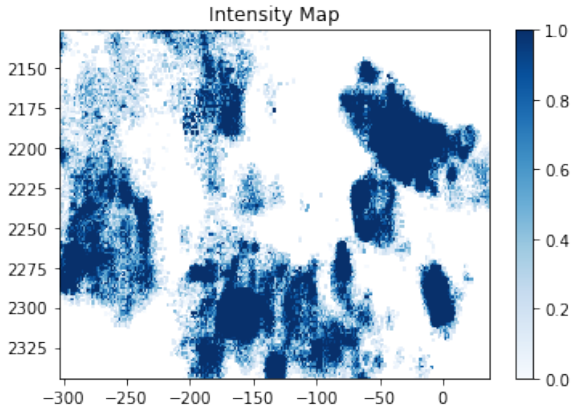

Figure 4.4: Map of one components on the middle layer of the tablet. The variations in the gradient of the colors in blue indicate the variations in the intensities of the peak.

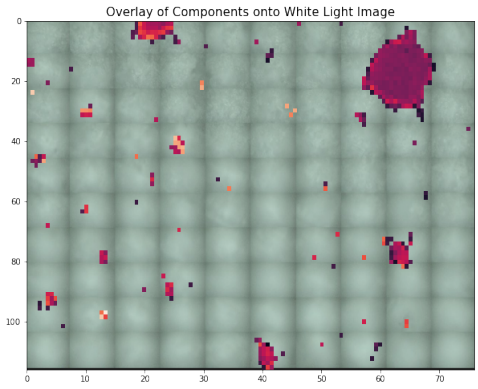

Figure 4.5: Pixel wise Linear Regression based abundance map of component A

4.3.2 Multivariate Intensity Map

The intensity maps generated by multivariate methods are significantly better than that of the univariate methods due to the robustness against noisy spectra. The major drawback of univariate methods of prioritizing the characteristic region of the spectra is avoided in the multivariate methods. Another problem lies in the filtering of the pixels in order to get rid of those that do not contain a particular component. In case of univariate methods the filtering is required to be done based on the concentration value lying between 0-1, which becomes very subjective in nature. However, in case of multivariate techniques the filtering is performed using the value of R^2. This enables the user to have a form of interpretation while filtering the pixels in the abundance maps. Although, the R^2 based filter could be used in case of pixel wise regression this is not applicable in case of matrix factorization. Notice that in Figure 4.5 the actual component is easily identified and filtered out while in case of Figure 4.6 it becomes difficult to filter the component since some other positions seem to have larger area under the characteristic peak than the actual component. This problem arises when the characteristic peak range of components overlap or are disguised within a noisy region as shown in Figure 4.7

4.3.3 Quantification

In order to quantify the constituents of a sample the majority of literature focus on using a calibration data which is further trained to learn a linear regression model and then used to predict the quantities for varying concentration inputs of the sample. However, in this work we try to address the problem of extracting the unknown

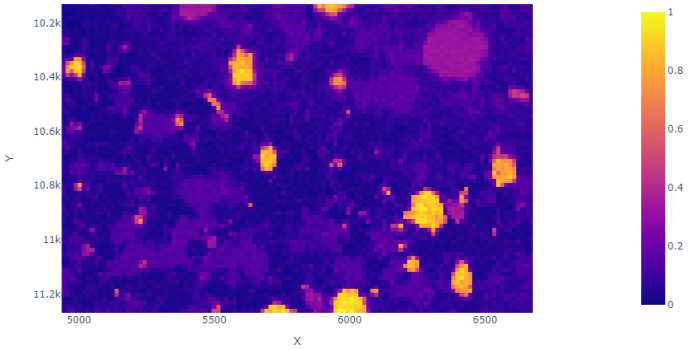

Figure 4.6: Pixel Area based abundance map of component A

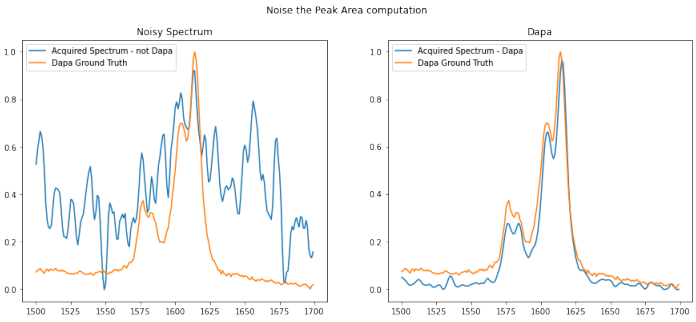

Figure 4.7: Noisy Spectra leads to larger peak area although the spectra is not that of the component under study

quantity of components from a sample directly using the Raman spectroscopic data. The quantitative Raman measurements utilize the following relationship between signal, S_γ, at a given wavenumber, γ, and the concentration of the sample, C -

$$S_\gamma = K\sigma_\gamma\gamma_L(\gamma_L - \gamma_\beta)^3 P_0 C \tag{4.1}$$

- K: Constant that depends on laser beam diameter, collection optics, sample volume and temperature.
- σ_γ: Raman cross-section of the particular vibrational mode.
- γ_L: Laser wavenumber
- γ_β: Wavenumber of the vibrational mode.
- P_0: Laser Power

From the above equation it is apparent that peak signal is directly proportional to concentration. On the basis of this relationship we try to quantify the components. Using the univariate and multivariate methodologies described earlier we generate an abundance map corresponding to each constituent depicting the intensity at every pixel. Now, since the intensity is directly proportional to concentration according to equation 4.1, the intensity map can also be looked upon as the concentration map to some extent. Although we get the concentration map, the actual percentage of a constituent still remains unknown. In order to deal with this problem, we use the idea of relative concentration given in equation 4.2.

$$\begin{aligned}&\textbf{Relative proportion of component i}\\ &= \frac{\text{Sum of pixel values of the concentration map of } i^{th} \text{ component}}{\sum_j (\text{Sum of pixel values of the concentration map of } i^{th} \text{ component})}\end{aligned} \tag{4.2}$$

Chapter 5

Discussions

The aim of this work is to build a quantitative tablet characterization platform that can utilize the information generated from Raman spectroscopy and accurately quantify the components in a sample. A number of permutations of the pre-processing steps followed by univariate or multivariate methods can be used to generate results with high accuracy. We performed all the necessary pre-processing steps and pixel based linear regression on a number of products and quantified the components which were then validated against the actual values tabulated in Table 5.1

The quantification of the components with concentrations close or less than 1 % are difficult to predict accurately primarily due to the instrumentation. The laser width, step size and various such factors affect the proper capturing of data, thereby leading to an average data acquisition. On the contrary, for components with significantly larger quantities were predicted accurately with ¡ 4 % error.

Table 5.1: Validation of quantities of components (%)

Product	Components	Predicted Quantity (%)	Ground Truth (%)
Product-I	Component-1	99.62	99.66
	Component-2	0.11	0.34
Product-II	Component-1	4.39	4.42
	Component-2	69.47	69.00
	Component-3	17.80	17.25
	Component-4	1.67	1.25
Product-III	Component-1	63.03	61.88
	Component-2	24.16	22.27
	Component-3	5.1	3.0

In case of layered tablets we provide the added specification of quantifying the components within each layer separately. This has been accomplished by usage of masking the particular layer of interest and then following the procedure as described elsewhere [Chapter 3].

The quantities predicted are with respect to the total weight of the product. Thus, given the innovator tablet and the constituent components with the respective pure spectral signatures we can extract the Raman spectroscopic data from a cross section of the tablet and quantify each of components.

However, the study is primarily based on the assumption that the product is homogeneous throughout, thereby data extracted from a single cross section is representative of the whole tablet. Also all the multivariate models are made based on the assumption that the acquired Raman spectra is a linear combination of the pure spectral signatures of the constituent components.

Chapter 6

Conclusions

Raman spectroscopy is a powerful and well-established tool for quantitative analysis. The high information content from Raman spectra often allow simple analytical models to be accurate and robust. Spectral pre-processing, noise analysis, and multivariate methods extend the quantitative capabilities to more complex samples. This work has proven that even in the absence of calibration data simple linear models are capable of quantifying and generating spatial distribution of constituents of an analyte. The agile reverse engineering methodology allows fast and accurate breakdown of an unknown sample thereby benefiting the pharmaceutical industry of easier production and development. Although the study is based on data generated from Raman spectra of a drug, it can be further extended to other forms of absorption or emission spectra in different domains. However, the combination of both multivariate and univariate methods may lead to better results. The recent availability of rugged, turnkey analyzers together with flexible, non-contact sampling benefits make the future for quantitative Raman analysis look very promising indeed.

Bibliography

[1] José Manuel Amigo, Jordi Cruz, Manel Bautista, Santiago Maspoch, Jordi Coello, and Marcelo Blanco. Study of pharmaceutical samples by nir chemical-image and multivariate analysis. *TrAC Trends in Analytical Chemistry*, 27(8):696–713, 2008.

[2] Jeremy J. Andrew and Thomas M. Hancewicz. Rapid analysis of raman image data using two-way multivariate curve resolution. *Applied Spectroscopy*, 52(6):797–807, 1998.

[3] Y. Anzai. *Pattern Recognition and Machine Learning*. Academic Press, 1 edition, 1992.

[4] Thomas Bocklitz, Angela Walter, Katharina Hartmann, Petra Rösch, and Jürgen Popp. How to pre-process raman spectra for reliable and stable models? *Analytica chimica acta*, 704(1-2):47–56, 2011.

[5] Jörg Breitenbach, Wolfgang Schrof, and Jörg Neumann. Confocal raman-spectroscopy: analytical approach to solid dispersions and mapping of drugs. *Pharmaceutical research*, 16(7):1109–1113, 1999.

[6] K Brudzewski, A Kesik, K Kołodziejczyk, U Zborowska, and J Ulaczyk. Gasoline quality prediction using gas chromatography and ftir spectroscopy: An artificial intelligence approach. *Fuel*, 85(4):67–76, 2006.

[7] Anna de Juan, Romà Tauler, Raylene Dyson, Claudia Marcolli, Marianne Rault, and Marcel Maeder. Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis. *TrAC Trends in Analytical Chemistry*, 23(1):70–79, 2004.

[8] Ivana Durickovic. Using raman spectroscopy for characterization of aqueous media and quantification of species in aqueous solution. In *Applications of Molecular Spectroscopy to Current Research in the Chemical and Biological Sciences*. University of Pittsburgh at Greensburg, United States of America, 2016.

[9] D. Eastwood, R. L. Lidberg, and K. J. Siddiqui. *The Role of Luminescence and Spectral Pattern Recognition in Environmental Programs*. Springer US, 1991.

[10] KH Esbensen, D Guyot, F Westad, and LP Houmoller. Multivariate data analysis-in practice: An introduction to multivariate data analysis and experimental design.,(camo: Oslo, norway). 2006.

[11] Patrick Chen E Neil Lewis Richard V Vivilecchia Eunah Lee, Wei X Huang. High-throughput analysis of pharmaceutical tablet content uniformity by near-infrared chemical imaging. *SPECTROSCOPY-SPRINGFIELD THEN EUGENE THEN DULUTH-*, 21(11):24, 2006.

[12] WP Findlay and DE Bugay. Utilization of fourier transform-raman spectroscopy for the study of pharmaceutical crystal forms. *Journal of pharmaceutical and biomedical analysis*, 16(6):921–930, 1998.

[13] Giancarlo Fini. Applications of raman spectroscopy to pharmacy. *Journal Of Raman Spectroscopy*, 35(5):335–337, 2004.

[14] Mark S Friedrichs. A model-free algorithm for the removal of baseline artifacts. *Journal of Biomolecular NMR*, 5(2):147–153, 1995.

[15] Paul Geladi. Chemometrics in spectroscopy. part 1. classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 58(5):767–782, 2003.

[16] Paul Geladi, Britta Sethson, Josefina Nyström, Tom Lillhonga, Torbjörn Lestander, and Jim Burger. Chemometrics in spectroscopy: part 2. examples. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 59(9):1347–1357, 2004.

[17] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, and S.E.J. Bell. Recent applications of chemical imaging to pharmaceutical process monitoring and quality control. *European Journal of Pharmaceutics and Biopharmaceutics*, 69(1):10–22, 2008.

[18] Anna de Juan, Sara Piqueras, Marcel Maeder, Thomas Hancewicz, Ludovic Duponchel, and Romà Tauler. Infrared and raman spectroscopic imaging. In *Chemometric Tools for Image Analysis*, pages 57–110, 2014.

[19] Kaho Kwok and Lynne S Taylor. Analysis of the packaging enclosing a counterfeit pharmaceutical tablet using raman microscopy and two-dimensional correlation spectroscopy. *Vibrational Spectroscopy*, 61:176–182, 2012.

[20] Frans W Langkilde, Jonas Sjöblom, Lija Tekenbergs-Hjelte, and Jonni Mrak. Quantitative ft-raman analysis of two crystal forms of a pharmaceutical compound. *Journal of pharmaceutical and biomedical analysis*, 15(6):687–696, 1997.

[21] F.W. Langkilde and A. Svantesson. Identification of celluloses with fourier-transform (ft) mid-infrared, ft-raman and near-infrared spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, 13(4):409–414, 1995.

[22] William H Lawton and Edward A Sylvestre. Self modeling curve resolution. *Technometrics*, 13(3):617–633, 1971.

[23] Sheng Li and Liankui Dai. An improved algorithm to remove cosmic spikes in raman spectra for online monitoring. *Applied spectroscopy*, 65(11):1300–1306, 2011.

[24] Chad A. Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied Spectroscopy*, 57(11):1363–1367, 2003.

[25] M. J. Pelletier. Quantitative analysis using raman spectrometry. *Applied Spectroscopy*, 57(1):20A–42A, 2003.

[26] M. J. Pelletier. Quantitative analysis of main components of natural gas based on raman spectroscopy. *Chinese Journal of Analytical Chemistry*, 47(1):67–76, 2019.

[27] Hajnalka Pataki Zsuzsanna Eke Attila Farkas Geert Verreck Éva Kiss Pál Fekete Tamás Vigh István Wagner Zsombor K Nagy György Marosi Péter Lajos Sóti, Katalin Bocz. Comparison of spray drying, electroblowing and electrospinning for preparation of eudragit e and itraconazole solid dispersions. *International journal of pharmaceutics*, 494(1):23–30, 2015.

[28] C. V. Raman and K. S. Krishnan. A new type of secondary radiation. *Nature*, 121(3048):501–502, 1928.

[29] Christiane Rodriguez. Analysis of pharmaceutical materials through the use of ft-raman microprobe imaging. In Michael D. Morris, editor, *Biomedical Applications of Raman Spectroscopy*, pages 37–43, 1999.

[30] A. de Juan S. C. Rutan and R. Tauler. Introduction to multivariate curve resolution. In *Comprehensive Chemometrics*, pages 249–259, 2009.

[31] S. Sasic. An in-depth analysis of raman and near-in- frared chemical images of common pharmaceutical tablets. *Applied Spectroscopy*, 61(3):239–250, 2007.

[32] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[33] Darren A. Whitaker and Kevin Hayes. A simple algorithm for despiking raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 179:82–84, 2018.

[34] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.

[35] Jianhua Zhao, Harvey Lui, David I McLean, and Haishan Zeng. Automated autofluorescence background subtraction algorithm for biomedical raman spectroscopy. *Applied Spectroscopy*, 61(11):1225–1232, 2007.