



UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA
GRADUAÇÃO EM FÍSICA

PEDRO HENRIQUE MOREIRA LIMA

ESTUDO DOS MÉTODOS DE ANÁLISE MULTIVARIADA EM
ESPECTROSCOPIA VIBRACIONAL RAMAN

FORTALEZA

2016

PEDRO HENRIQUE MOREIRA LIMA

ESTUDO DOS MÉTODOS DE ANÁLISE MULTIVARIADA EM ESPECTROSCOPIA
VIBRACIONAL RAMAN

Monografia de Bacharelado apresentada à
Coordenação da Graduação do Curso de
Física, da Universidade Federal do Ceará,
como requisito parcial para a obtenção do
Título de Bacharel em Física.

Orientador: Prof. Dr. Alejandro Pedro
Ayala.

FORTALEZA
2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Curso de Física

-
- L71e Lima, Pedro Henrique Moreira
Estudo dos métodos de análise multivariada em espectroscopia vibracional Raman / Pedro Henrique Moreira Lima. – 2016.
38 f. : il. algumas color.
- Monografia (Graduação em Física) – Universidade Federal do Ceará, Centro de Ciências, Departamento de Física, Curso de Bacharelado em Física, Fortaleza, 2016.
Orientação: Prof. Dr. Alejandro Pedro Ayala.
Inclui bibliografia.
1. Espectroscopia Raman. 2. *Loadings* (espectros). 3. Análise multivariada. I. Ayala, Alejandro Pedro. II. Título.

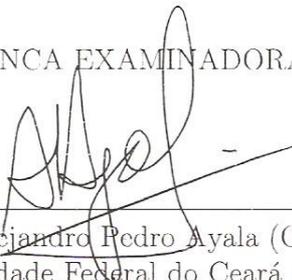
PEDRO HENRIQUE MOREIRA LIMA

ESTUDO DOS MÉTODOS DE ANÁLISE MULTIVARIADA EM ESPECTROSCOPIA
VIBRACIONAL RAMAN

Monografia de Bacharelado apresentada à
Coordenação da Graduação do Curso de
Física, da Universidade Federal do Ceará,
como requisito parcial para a obtenção do
Título de Bacharel em Física.

Aprovada em 04/02/2016.

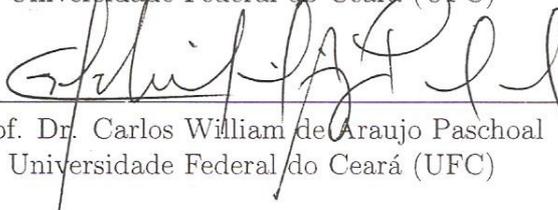
BANCA EXAMINADORA



Prof. Dr. Alejandro Pedro Ayala (Orientador)
Universidade Federal do Ceará (UFC)



Prof. Dr. Alexandre Rocha Paschoal
Universidade Federal do Ceará (UFC)



Prof. Dr. Carlos William de Araujo Paschoal
Universidade Federal do Ceará (UFC)

*Aos Meus Pais
e
amigos*

AGRADECIMENTOS

Gostaria de primeiramente agradecer a Deus pela perseverança, pela força dada e também por nunca me deixar esmorecer durante os tempos difíceis.

Agradeço também a minha família em especial minha irmã, Fernanda, pelo imenso apoio, minha mãe, Ana Cristina, meu pai, Alberto, pelo apoio e o investimento.

Ao meu orientador, Ayala, pelos conhecimentos, projetos e todo tempo dedicado a este trabalho. E também aos meus amigos, Fábio e Bruno, e as amigas do grupo de espectroscopia por todo apoio.

As amizades feitas ao longo do curso e que tornaram esse período da minha vida divertido também, as amigas: Laura Barth, Sofia, Bianca e Ana Lúcia; aos amigos: Augusto, Wagner, Victor, Matheus Falcão, Daniel, Jonathan, Emanuel, Michel, Pablo, Raul, Nathan, João Paulo, Wendel, Rondinelly e Josa. A todos obrigado pelas risadas e companheirismo.

RESUMO

O avanço do estudo de materiais na escala atômica tornou possível entender certos fenômenos que ocorrem nessa região e como eles influenciam o mundo macroscópico. Para alcançar tal entendimento se usa técnicas que facilitem o estudo e ajudem no aprofundamento desse conhecimento. Podem-se destacar três métodos de espectroscopias bem difundidas para essa finalidade: espectroscopia de infravermelho, espectroscopia de fotoluminescência e espectroscopia de espalhamento Raman. Contudo o foco do trabalho reside sobre a espectroscopia de espalhamento Raman. Para procedimentos experimentais é comum ter-se uma grande quantidade de dados a respeito da amostra, consequentemente, precisamos de técnicas que ajudem a interpretar esses dados. O equipamento utilizado (LabRam) nos fornece quatro técnicas: *Principal Component Analysis* (PCA), *Multivariate Curve Resolution* (MCR), *Hierarchical Clustering Analysis* (HCA) e *Divisive Clustering Analysis* (DCA), onde o objetivo é entender e explicar os dois primeiros métodos aplicados a espectroscopia Raman. Como durante as medidas geram-se muitos resultados (espectros), dependendo do que estamos procurando podemos usar o PCA ou MCR para fazer o tratamento desses espectros. O tratamento consiste, basicamente, em álgebra com matrizes, as quais são determinadas por meio dos espectros. E como resultado encontra-se alguns espectros específicos denominados de *loadings*, também, a partir do tratamento empregado determinamos os *scores* que estão relacionados com os *loadings*. Tomando como exemplo o MCR, por meio dos *loadings* que são obtidos desse método determinamos quais os componentes que formam a região que delimitamos da amostra que estudamos.

Palavras-chave: Espectroscopia. MCR. *Loadings*. Espectros.PCA

ABSTRACT

The study of advance materials at the atomic scale has made it possible to understand certain phenomena that occur in this region and how they influence the macroscopic world. To achieve this understanding using techniques that facilitate and help in deepening this knowledge. One can highlight three very spectroscopy methods disseminated for this purpose: infrared spectroscopy, luminescence spectroscopy and Raman spectroscopy. However the focus of the work lies on the Raman scattering spectroscopy. For experimental procedures it is common to have a large amount of data about the sample, therefore, we need techniques to help interpret these data. The equipment (LabRam) gives us four techniques Principal Component Analysis (PCA) Multivariate Curve Resolution (MCR), Hierarchical Clustering Analysis (HCA) and Divisive Clustering Analysis (DCA). Where the objective is to undertand and explain the first two methods applied to Raman spectroscopy. As for the measures are generated too many results (spectra), depending on what we are looking for we can use the PCA or MCR to treat these spectra. The treatment consists basically in algebra with matrices, which are determinated by the spectra. And as a result is termed some specific loadings spectra also from the treatment used to determine the scores are associated with loadings. Taking as an example the MCR by means of loadings which are obtained that method we determine wich components make up the region to delimit the sample we studied.

Keywords: *Spectroscopy. MCR. Loadings. Spectrum. PCA.*

LISTA DE FIGURAS

Figura 1 – Tipos de espalhamento da luz	14
Figura 2 – Microscopia Raman: Lente objetiva do microscópio usada para focalizar o laser sobre a amostra e coletar a radiação espalhada.	16
Figura 3 – Gráfico de análise para uma região qualquer. O gráfico exhibe claramente a localização das classes para uma área analisada.	19
Figura 4 – Tabelas de exemplificação. Na primeira tabela temos a localização das classes e na segunda o quanto de área em percentagem cada uma preenche.	19
Figura 5 – Dendograma de classificação. O eixo horizontal indica a distância	20
Figura 6 – Foto de uma cápsula de remédio com uma região.	24
Figura 7 – Mapeamento da região delimitada na figura 6 de uma cápsula de remédio (fármaco).	25
Figura 8 – Componentes gerados pela técnica de PCA na amostra exibida na figura 7.	26
Figura 9 – Carbonatos com uma região delimitada obtida através do microscópio.	28
Figura 10 – Mapeamento obtido pela Resolução de Curva Multivariada em carbonatos, onde cada cor representa um espectro específico.	29
Figura 11 – Loadings gerados pela técnica de MCR na amostra exibida na figura 9.	30
Figura 12 – Espectro Raman do Metropolol e PC 2.	31
Figura 13 – Espectro Raman do Metropolol e loading 1.	32
Figura 14 – Espectro Raman da Etilcelulose.	33
Figura 15 – Loading 5.	33
Figura 16 – Espectro Raman do HPC.	34
Figura 17 – Componente Principal 1.	34
Figura 18 – Mapeamento da figura 6 por meio do MCR.	35

SUMÁRIO

1	INTRODUÇÃO	10
2	ESPECTROSCOPIA RAMAN	12
2.1	Fundamentos Teóricos	12
3	ANÁLISE MULTIVARIADA - MVA	17
3.1	Fundamentos: Decomposição e Agrupamento	17
4	DECOMPOSIÇÃO - PCA	21
4.1	Pré-processamento	21
4.2	Análise do Componente Principal	22
5	DECOMPOSIÇÃO - MCR	27
6	DISCUSSÃO E RESULTADOS	31
6.1	Componentes Principais e Loadings (Fármaco)	31
7	CONCLUSÃO	36
	REFERÊNCIAS	37

1 INTRODUÇÃO

A luz consiste, em todas as suas formas, de radiação eletromagnética, onde estão inseridos também os raios gama, raios X, ultravioleta, luz visível, infravermelho, microondas e até mesmo ondas de rádio e de televisão. Todas possuem uma velocidade constante específica e podemos caracterizá-las por meio do comprimento de onda (λ) e/ou a frequência (ν), sendo estes geralmente as variáveis mais importantes para diferenciação entre as ondas eletromagnéticas citadas inicialmente.

O avanço do estudo de materiais na escala atômica tornou possível entender certos fenômenos que ocorrem nessa região e como eles influenciam no mundo macroscópico. Dessa forma, se faz necessário o uso de técnicas que facilitem esse estudo e nos ajudem a aprofundar esses conhecimentos. Podemos destacar três técnicas espectroscópicas bem difundidas que são espectroscopia de infravermelho, espectroscopia Raman e espectroscopia de fotoluminescência. Contudo, o foco do trabalho será no Raman, onde se uma onda eletromagnética interage com a superfície do meio parte dela é refletida e outra é transmitida para o interior do material, a qual uma fração dessa parte é absorvida na forma de calor e o restante é retransmitido como uma luz espalhada, que emerge do material. Essa luz apresenta uma parcela composta por frequências diferentes da incidente, por isso o nome espalhamento Raman [1].

Para procedimentos experimentais é comum ter-se uma grande quantidade de dados a respeito da amostra, conseqüentemente, precisamos de técnicas que ajudem a filtrar esses dados facilitando a escolha daqueles que, realmente, são importantes. Por isso necessitamos de técnicas como *Principal Component Analysis* e *Multivariate Curve Resolution*.

Inicialmente, este trabalho consiste em estabelecer as bases para o entendimento, um pouco aprofundado, de como foi a descoberta, o que são e como são determinados os espectros obtidos da espectroscopia Raman.

Com esse entendimento passamos a explicar onde se encaixa a análise multivariada nesse processo. Também abordamos suas principais características e como ela se segmenta afim de atender objetivos específicos. É nesse capítulo que indicamos qual será o caminho adotado no trabalho, ou seja, por qual segmento da análise multivariada iremos estudar.

No capítulo seguinte, tratamos de apresentar como ocorre o cálculo para a determinação dos componentes principais a partir de uma quantidade exacerbada de matrizes de espectros. Em seguida são exibidos os componentes principais obtidos a partir de

uma região pré-selecionada da amostra usada como exemplo para a técnica de PCA. Contudo, antes dessa parte foi reservado um tópico para discorrer sobre os pré-tratamentos que podem ser usados antes da análise multivariada.

No capítulo do MCR segue-se o mesmo objetivo apresentado no capítulo do PCA, dando, claramente, mais atenção as características que identifica essa técnica.

O último capítulo é explicado que informações podemos retirar dos *loadings* e componentes principais, dando assim uma ideia da importância dessas técnicas. Também ali, conseguimos identificar características semelhantes e aquelas que nos ajudam a diferenciar os resultados do MCR em relação aos do PCA.

2 ESPECTROSCOPIA RAMAN

2.1 Fundamentos Teóricos

O efeito Raman foi descoberto em 1928 por Chandrasekhara Venkata Raman, nascido ao sul da Índia. Depois da descoberta do efeito Compton, em 1923, por A. H. Compton, e com a previsão teórica do efeito Raman por Smekal, também em 1923, Raman utilizando um espectrômetro, a luz do sol como fonte de irradiação e o olho humano como detector observou que incidindo um feixe de luz monocromático (obtido da luz solar) sobre a amostra a radiação mudava de direção [2]. Basicamente, ele imaginava que durante a interação da radiação visível com a matéria seria possível variar a energia presente no fóton incidente [3].

Por meio da análise do espectrógrafo observou-se que haviam outras linhas, além da radiação incidente, que sofriam certas modificações causadas por um espalhamento inelástico, pois ocorria tanto uma mudança na direção quanto uma variação no comprimento de onda da radiação incidente. No caso do espalhamento elástico a diferença está no fato de que ocorre somente a mudança na direção da radiação. Este espalhamento é conhecido como espalhamento Rayleigh e o primeiro como espalhamento Raman.

No modelo clássico da espectroscopia Raman, a atividade está relacionada com a variação do momento de dipolo induzido da molécula provocada pelo campo elétrico da radiação incidente. A presença desse campo provoca na molécula um deslocamento das cargas negativas (nuvem eletrônica) em relação ao núcleo (cargas positivas). Como o centro de ambos não se coincidem tem-se, então, a formação do dipolo induzido mencionado anteriormente [2] e [4].

Como resultado da interação do campo elétrico, da radiação incidente, os elétrons passam a vibrar com sobreposição de frequências a partir dessa radiação, consequentemente, tem-se a variação da polarizabilidade (α) cuja posição está relacionada a um modo de vibração da molécula. Ou pode-se dizer que o vetor do momento de dipolo induzido oscila com a sobreposição de frequências e o mesmo pode ser escrito como:

$$\mathbf{P} = \alpha \mathbf{E} \quad (2.1)$$

onde \mathbf{P} é o vetor momento de dipolo induzido e \mathbf{E} o vetor campo elétrico.

Para pequenos deslocamentos é possível desenvolver o (α) em uma série de

Taylor dependendo de uma coordenada interna (q), a qual representa uma coordenada normal do sistema, da seguinte forma:

$$\alpha = \alpha_0 + \left(\frac{d\alpha}{dq} \right)_0 q + \dots \quad (2.2)$$

Os outros termos, de ordem maiores, podem ser desprezados devido a pequena variação com respeito a q . Como essa coordenada varia periodicamente e o campo elétrico da radiação eletromagnética varia com o tempo podemos representá-los por:

$$q = q_0 \cos(2\pi\nu_v t) \quad (2.3)$$

$$\mathbf{E} = \mathbf{E}_0 \cos(2\pi\nu_{vi} t) \quad (2.4)$$

Com ν_v sendo a frequência de vibração e ν_{vi} a frequência de vibração da radiação incidente. Dessa forma o momento de dipolo passa a ser:

$$\mathbf{P} = \alpha_0 \mathbf{E}_0 \cos(2\pi\nu_{vi} t) + \left(\frac{d\alpha}{dq} \right)_0 q_0 \mathbf{E}_0 \cos(2\pi\nu_{vi} t) \cos(2\pi\nu_v t) \quad (2.5)$$

Por meio da regra trigonométrica:

$$\cos(a)\cos(b) = \frac{1}{2} [\cos(a+b) + \cos(a-b)] \quad (2.6)$$

Então tem-se que:

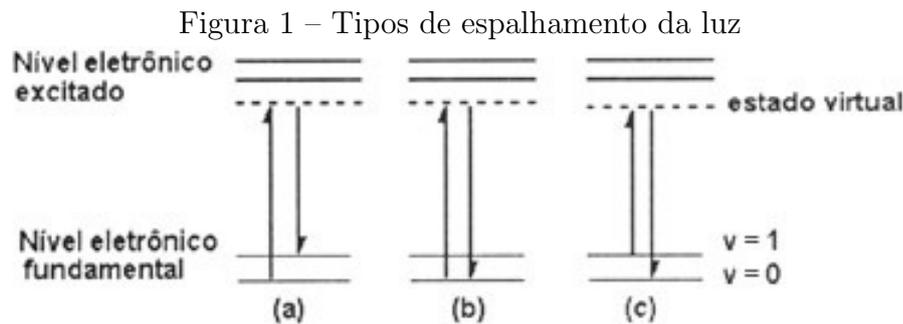
$$\mathbf{P} = \alpha_0 \mathbf{E}_0 \cos(2\pi\nu_{vi} t) + \frac{1}{2} \left(\frac{d\alpha}{dq} \right)_0 q_0 \mathbf{E}_0 [\cos 2\pi(\nu_{vi} + \nu_v)t + \cos 2\pi(\nu_{vi} - \nu_v)t] \quad (2.7)$$

Dessa equação, o primeiro termo indica o espalhamento elástico (espalhamento Rayleigh), onde o mesmo contém apenas a frequência incidente. O segundo termo apresenta as radiações de espalhamento Stokes (frequência $\nu_{vi} - \nu_v$) e espalhamento anti-Stokes (frequência $\nu_{vi} + \nu_v$).

No espalhamento pode-se obter como resultado um fóton de energia superior (anti-Stokes), ou um fóton de energia inferior (Stokes) ou um fóton com o mesmo nível

de energia em relação ao da radiação incidente que interage com a matéria (Rayleigh). Por meio da figura a baixo, no item (b), observa-se que o campo elétrico da radiação eletromagnética leva a molécula ao estado virtual (estado intermediário) em seguida ela retorna para o seu estado vibracional original, exibindo assim o espalhamento elástico.

Pela Figura 1 é possível imaginar que aconteça uma absorção seguida de uma emissão, contudo não é isso que ocorre, pois o estado virtual não pode ser considerado um auto estado.



Fonte: [2]. (a) espalhamento inelástico (região Stokes); (b) espalhamento elástico (Rayleigh); (c) espalhamento inelástico (região anti-Stokes)

Para o caso do espalhamento inelástico pode-se ter, como mostrado pela figura anterior, um fóton de menor energia ou um fóton de maior energia. No item (a) a radiação incidente interage com a molécula no seu estado vibracional original e excita o sistema para um estado virtual, em seguida o mesmo decai para um estado vibracional cuja a energia é superior ao estado original. O fóton espalhado nesse processo apresenta uma energia cuja a diferença em relação ao incidente equivale a excitar a molécula até esse estado final.

No item (c) da figura a radiação encontra a molécula em um estado já excitado e a leva para um estado virtual, depois o sistema decai para o seu estado vibracional fundamental. Nesse processo o fóton espalhado apresenta uma energia superior em relação ao fóton incidente [2].

A população dos estados excitados seguem a lei de distribuição de Boltzman, devido o fóton incidente encontrar a molécula em um estado excitado, e por isso o espalhamento anti-Stokes apresenta intensidade menor do que Stokes.

Um ponto importante, o qual não pode ser esquecido é que o momento de transição induzido com relação aos processos anteriores (espalhamento inelástico na região de Stokes e anti-Stokes) toma a seguinte forma: $\mathbf{P}_{mn} = \mathbf{E} \cdot (\alpha_{ij})_{mn}$. Onde $(\alpha_{ij})_{mn}$ repre-

senta o tensor de polarizabilidade:

$$\alpha = \begin{pmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yz} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{pmatrix} \quad (2.8)$$

Conseqüentemente, a relação dos componentes do momento de dipolo induzido com as componentes do campo elétrico passa a ser descrito como:

$$P_x = \alpha_{xx}E_x + \alpha_{xy}E_y + \alpha_{xz}E_z$$

$$P_y = \alpha_{yx}E_x + \alpha_{yz}E_y + \alpha_{yz}E_z$$

$$P_z = \alpha_{zx}E_x + \alpha_{zy}E_y + \alpha_{zz}E_z$$

Para o espalhamento Raman deve-se considerar dessas equações somente as derivadas dos componentes de α em função do modo vibracional que formam a seguinte relação: $\alpha'_{xy} = \alpha'_{yx}$, $\alpha'_{xz} = \alpha'_{zx}$ e $\alpha'_{yz} = \alpha'_{zy}$ (onde $\alpha'_{ij} = (d\alpha'_{ij}/dq)_0$). Tal relação é conhecida como tensor Raman [4].

A intensidade da radiação espalhada está relacionada com o módulo quadrado do produto entre a polarização da luz incidente (\mathbf{p}_i), o tensor Raman e a polarização da luz espalhada (\mathbf{p}_e):

$$I_S \propto | \mathbf{p}_i \cdot \overline{R} \cdot \mathbf{p}_e |^2 \quad (2.9)$$

Onde I_S representa a intensidade e o \overline{R} o tensor Raman.

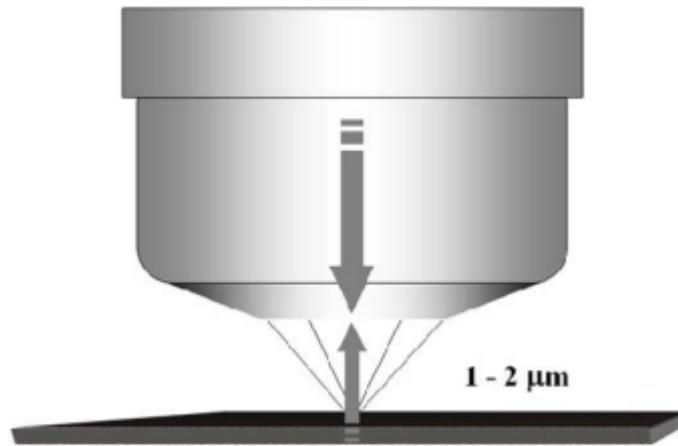
Conhecendo o tensor Raman para uma oscilação específica, se a intensidade da radiação espalhada for diferente de zero para uma combinação de polarizações referente a luz incidente e espalhada, então, o modo vibracional está ao alcance do espalhamento Raman (Raman ativo). Se a intensidade for nula teremos o modo Raman inativo. Como essa atividade Raman é dependente da direção de propagação das oscilações e da geometria de espalhamento, as combinações, juntamente com a geometria são chamadas de regras de seleção.

Com isso e tendo em mãos a radiação espalhada de um material que foi previamente iluminado, podemos obter as frequências dos modos vibracionais da amostra fazendo a diferença entre os espectros do feixe incidente e espalhado. No espectro Raman temos no eixo x (abscissas) a diferença entre os números de onda da radiação incidente e espalhada e no eixo y (ordenadas) as intensidades.

Para o trabalho discorrido aqui utilizamos a microscopia Raman, contudo a única diferença para espectroscopia Raman está na captação da radiação espalhada. Nesse

caso a lente objetiva do microscópio ótico tanto serve para focalizar o feixe incidente quanto para detectar a radiação espalhada [1].

Figura 2 – Microscopia Raman: Lente objetiva do microscópio usada para focalizar o laser sobre a amostra e coletar a radiação espalhada.



Fonte: [5].

3 ANÁLISE MULTIVARIADA - MVA

3.1 Fundamentos: Decomposição e Agrupamento

O procedimento para o estudo de múltiplas variáveis é denominado de *Multivariate Analysis* (MVA) e ele consiste de uma coleção de métodos que podem ser usados quando se precisa realizar várias medições em uma ou mais amostras. Historicamente, as técnicas de análise multivariada eram aplicadas grandemente na área das ciências comportamentais e biológicas, porém o interesse de outras áreas com relação a este campo é notável há muito tempo [6].

Desse procedimento temos duas formas de tratar nossos conjuntos de dados, por meio da Decomposição, onde temos os métodos *Principal Component Analysis* (PCA) e *Multivariate Curve Resolution* (MCR), e por meio do Agrupamento (*Clustering*) onde temos os métodos *Hierarchical Clustering Analysis* (HCA) e *Divisive Clustering Analysis* (DCA). Existem outros métodos, porém o foco desse trabalho será apenas nos dois primeiros (os mais usados PCA e MCR), os quais serão discutidos nos próximos capítulos.

A decomposição consiste, basicamente, em uma técnica de fatorização, pois com o mapeamento feito podemos determinar uma matriz de espectros, a qual fatoramos da seguinte forma:

$$\begin{aligned}
 Espectro_1 &= Score_{e_{1,1}} \cdot Loading_1 + Score_{e_{1,2}} \cdot Loading_2 + LoadingsDescartados \\
 Espectro_2 &= Score_{e_{2,1}} \cdot Loading_1 + Score_{e_{2,2}} \cdot Loading_2 + LoadingsDescartados \\
 &\cdot \\
 &\cdot \\
 &\cdot \\
 Espectro_n &= Score_{e_{n,1}} \cdot Loading_1 + Score_{e_{n,2}} \cdot Loading_2 + LoadingsDescartados
 \end{aligned}$$

São necessários, normalmente, poucos *loadings* ou *Principal Components* (PC), pois com apenas eles já se torna possível ter uma descrição razoável do conjunto de dados. O restante pode ser expresso como resíduos, em outras palavras, como interferência. É nesse caso que se encaixa os *loadings* descartados [7].

Podemos expressar os espectros de uma maneira geral por meio da seguinte função:

$$S_i = \sum_j s_{ij} \cdot L_j + E \quad (3.1)$$

Onde s_{ij} representa os *scores*, L_j os *loadings* e E os resíduos [7].

Com esses *loadings* e *scores* conseguimos reproduzir o gráfico de *loadings* ou PC's, o qual nos ajuda a identificar elementos importantes dos conjuntos de dados. O respectivo gráfico pode fornecer um entendimento da composição química e da distribuição da matriz de espectros.

A maneira como usamos as técnicas de decomposição abordadas (PCA e MCR) são similares. Quando selecionamos o número esperado de materiais químicos presentes na amostra, para facilitar, denominaremos de componentes principais (PC's) no PCA e *loadings* no MCR.

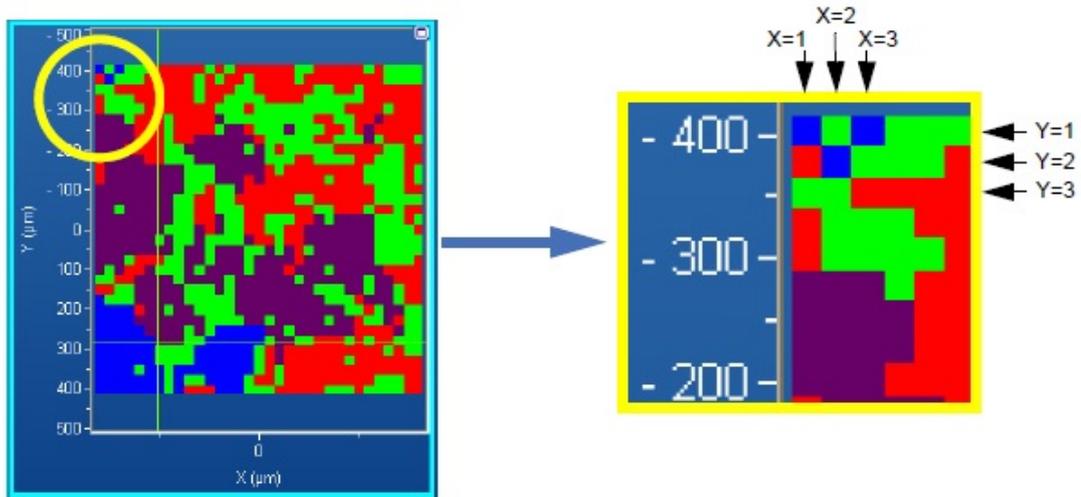
Na análise de *cluster* o que procuramos é um padrão no conjunto de dados, os quais agrupamos em *clusters*. Basicamente, podemos definir o objetivo de tal técnica como sendo o de encontrar um agrupamento mais favorável, no qual os dados presentes são semelhantes, contudo os *clusters* devem ser diferentes uns dos outros. Ou seja, a técnica de agrupamento está relacionada, basicamente, na construção de grupos cujos os elementos que os compõem apresentam similaridades com os elementos do próprio grupo.

O Agrupamento é um método útil para o desenvolvimento de pesquisas em várias áreas, podemos aplicá-lo no campo da biologia, medicina, economia, ciência computacional e etc, porém pela linha de pesquisa que seguimos ele não se torna muito útil, pois a partir das técnicas de Decomposição conseguimos extrair da amostra os dados que necessitamos. Portanto o aprofundamento de tal procedimento não faz parte dos objetivos deste trabalho, logo ele será abordado de uma maneira superficial com relação a espectroscopia vibracional Raman, onde nesta área ele é utilizado para classificar o conjunto de espectros em grupos com propriedades espectrais similares.

O uso do *Hierarchical Clustering Analysis* (HCA) e *Divisive Clustering Analysis* (DCA) são muito similares, precisa-se apenas selecionar a quantidade de classes, basicamente, para as medidas. No entanto cada um dos métodos fornece um gráfico específico como resultado, ajudando a identificar certas características presentes no conjunto de dados [8].

No caso do DCA, para uma definição breve, podemos dizer que ele faz uso de um método de particionamento denominado de k-médio, o qual se baseia nas medidas de variação dentro do *cluster* para formar os grupos homogêneos. O objetivo de tal procedimento, especificamente, está em segmentar os dados de tal forma que a variação presente no aglomerado seja minimizada [9]. Por meio da figura 3 podemos ter uma ideia de como os resultados são apresentados:

Figura 3 – Gráfico de análise para uma região qualquer. O gráfico exibe claramente a localização das classes para uma área analisada.



Fonte: [8].

Uma maneira de exemplificar a localização e a estatística das classes pode ser representada por meio da Figura 4:

Figura 4 – Tabelas de exemplificação. Na primeira tabela temos a localização das classes e na segunda o quanto de área em porcentagem cada uma preenche.

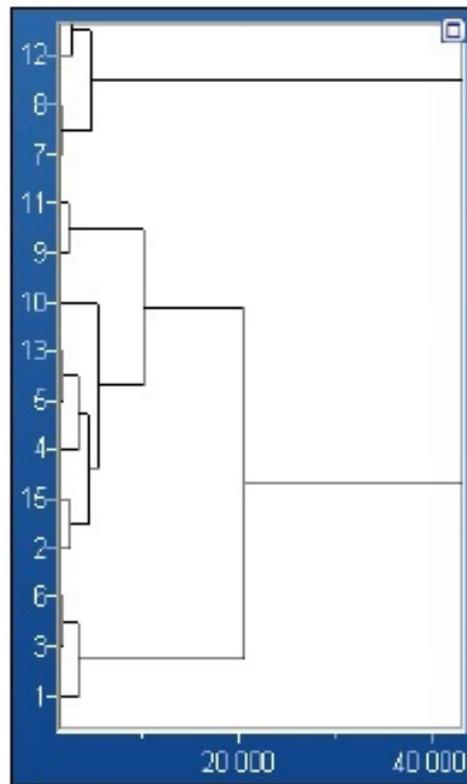
X index	Y index	Class
1	1	Class 3
1	2	Class 1
1	3	Class 2
1	4	Class 1
1	5	Class 1
1	6	Class 4
1	7	Class 4
1	8	Class 4
1	9	Class 4
1	10	Class 1

	Class 1	Class 2	Class 3	Class 4
Number of spectra	352	330	95	312
% of spectra	32%	30%	9%	29%

Fonte: [8].

Para o HCA, inicialmente, cada objeto de estudo é considerado um *cluster* individual. Estes *clusters* são, então, unidos sequencialmente de acordo com a sua semelhança. A princípio os dois *clusters* mais semelhantes (aqueles com menor distância entre si, distância essa determinada pelo método de Ward) são unidos para formar um novo conjunto na parte inicial da hierarquia. Em seguida outro par de *clusters* é unido e ligado ao próximo estágio da hierarquia e assim segue até que ela esteja completa. Pela Figura 5 temos uma exemplificação desse processo.

Figura 5 – Dendrograma de classificação. O eixo horizontal indica a distância



Fonte: [8].

4 DECOMPOSIÇÃO - PCA

4.1 Pré-processamento

Antes de passar para a técnica em si, uma etapa importante na análise é o pré-processamento que fazemos para que possamos encontrar melhores resultados. Os objetivos das técnicas de pré-processamento consistem em remover informações desnecessárias do ponto de vista químico e dessa forma a matriz de dados passa ter melhores condições para a análise.

Existem diversos métodos de pré-processamentos, porém o LabRam nos fornece apenas três: Normalização, Média Central e Auto Escala. Para o pré-processamento multivariado integrado o conjunto de dados ativos são usados para criar uma matriz onde cada espectro representa uma linha.

Na normalização é possível reduzir as variações indesejadas que aparecem no conjunto de dados e assim conseguimos medidas mais adequadas e consistentes. Neste trabalho o método de normalização disponível é denominado de normalização por área total cujo objetivo está em reduzir o efeito da intensidade total dos resultados obtidos causados por variações na concentração da amostra e do caminho ótico. A normalização apresentada pode ser descrita pela:

$$\mathbf{I}_i^{pre-processada} = \mathbf{I}_i / \sum_{k=1}^m \Delta s_k i_{ik} \quad (4.1)$$

Onde \mathbf{I}_i representa um vetor de linha com $i = 1...n$ e n indica o número de espectros. i_k representa o vetor coluna com $k = 1...m$ e m indicando o numero de pontos espectrais. Através da equação 4.1 a normalização para a área ocorre devido a divisão de cada variável pela soma dos valores absolutos de todas as variáveis de uma amostra específica.

Para o caso da média central esta técnica se baseia em calcular a média das intensidades para cada comprimento de onda e subtrair cada uma das intensidades do valor médio. Podemos expressar essa relação por meio da seguinte equação:

$$\mathbf{I}_i^{pre-processada} = [\mathbf{I}_i - \mathbf{I}_m] \quad (4.2)$$

Onde temos $\mathbf{I}_m = \sum_{k=i}^n \mathbf{I}_i / n$, indicando a média dos espectros. Assim as médias de cada uma das nossas variáveis passará a ser zero. Desta maneira as coordenadas podem ser levadas para o centro dos dados, permitindo perceber, facilmente, a diferença de

intensidades com relação as variáveis.

Por fim, na técnica da auto escala aplica-se a auto escala em cada coluna, as quais subtraímos com relação a média e em seguida as dividimos pelo desvio padrão. A equação que representa tal processo pode ser dada por:

$$i_{ik}^{pre-processada} = (i_{ik} - \mu_k) / \sigma_k \quad (4.3)$$

Onde $\mu_k = \sum_{i=1}^n i_{ik} / n$ e $\sigma_k = \sqrt{\sum_{i=1}^n (i_{ik} - \mu_k)^2 / n}$ [10] [8].

4.2 Análise do Componente Principal

Principal Component Analysis, esta é uma das técnicas de processamento de dados que abordaremos sobre uma ótica descritiva de modo geral nesta seção e mais a frente veremos sua aplicação para a espectroscopia Raman. Podemos dizer que ela entra na categoria dos métodos de Decomposição, onde o objetivo é determinar componentes chave para um conjunto de dados que nos ajude a interpretá-los [8]. Não é incomum esta técnica ser empregada em diversas áreas de pesquisas (nanotubos, farmacêutica, etc), porém o foco deste trabalho está no seu uso na espectroscopia vibracional Raman.

Como ja foi mencionado, PCA representa um método de Decomposição, sendo assim, outro fator que não pode ser esquecido é que ele, também, consiste em reduzir o número de variáveis de um conjunto de dados mantendo a máxima variação entre elas e através desse novo conjunto de variáveis identificamos padrões que estavam escondidos e os classificamos de acordo com as informações presente em cada um [11].

Com a definição geral estabelecida vejamos como ocorre o passo a passo de tal técnica de maneira algébrica. Na álgebra linear podemos exibir uma base vetorial ortonormal como:

$$B = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}_{m \times 1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m} = I \quad (4.4)$$

Onde cada linha representa uma base vetorial ortonormal b_i com m componentes e podemos associar os dados obtidos em relação a estas bases.

Considerando X como uma matriz de (m x n) contendo as medidas em Raman

e P a matriz de transformação linear onde:

$$P \cdot X = Y \quad (4.5)$$

Y: é a nova representação dos conjuntos de dados.

Podemos interpretar a equação (4.5) como uma mudança de base, logo é possível afirmar que P representa a matriz de transformação. Geometricamente, seria como uma rotação que novamente transforma X em Y e que as linhas de P (p_1, \dots, p_m) formam um conjunto das novas bases vetoriais para representar as colunas de X.

Assumindo a dependência linear o problema se reduz para uma mudança de base adequada e os vetores linha (p_1, \dots, p_m) nesta transformação passam a ser os principais componentes de X. O primeiro e o segundo PC são, geralmente, escolhidos para representar os eixos do gráfico de espalhamento (*scatter plot*), pois apresentam maior variação [11].

Com a álgebra desse processo entendida, vejamos como é feito tal método pelo equipamento. Com a amostra pronta para ser analisada, primeiramente delimitamos a região que procuramos analisar (como mostra a figura 6). O *software* (LabSpec 6) nos permite escolher quatro métodos de análise *Principal Component Analysis, Multivariate Curve Resolution, Hierarchical Clustering Analysis e Divisive Clustering Analysis*, dos quais, selecionamos o PCA.

O LabRam gera, a partir da região escolhida, uma matriz de espectros feitas em Raman:

$$X = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}_{m \times n} \quad (4.6)$$

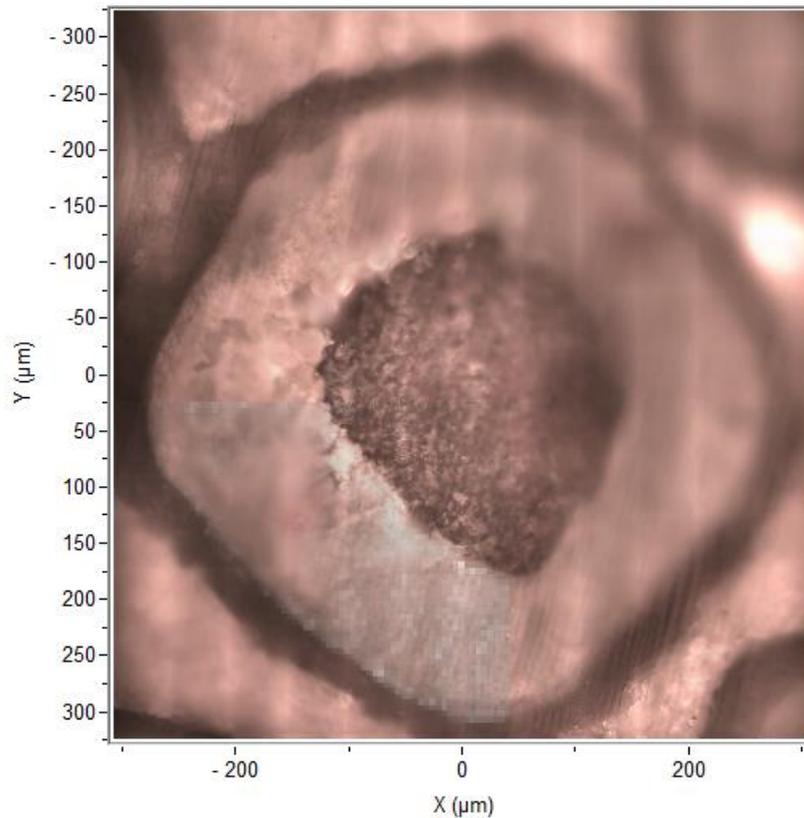
Onde as colunas representam as variáveis (número de onda) e as linhas indicam as amostras ou cada ponto da região em que foi realizada uma medida em Raman.

Em seguida é feito o auto escalonamento dessa matriz por meio do desvio padrão (medida da dispersão da variável com relação a sua média S_k) de cada variável.

$$U = \begin{pmatrix} a_{11}/S_1 & \cdots & a_{1n}/S_n \\ \vdots & \ddots & \vdots \\ a_{m1}/S_1 & \cdots & a_{mn}/S_n \end{pmatrix}_{m \times n} \quad (4.7)$$

Temos cada S_k ($k = 1 \dots n$) representando o desvio padrão das variáveis. Dessa nova matriz

Figura 6 – Foto de uma cápsula de remédio com uma região.



Fonte: LabRam HR.

calcula-se a matriz de covariância:

$$\Lambda = U^T \cdot U / (N - 1) \quad (4.8)$$

N é o número de linhas e Λ é a matriz de covariância.

Determinada a matriz de covariância o próximo passo é calcular a matriz de autovetores (também chamados de PC's) e os autovalores. Abaixo tem-se um exemplo simples de como determinar esses autovalores.

$$\begin{pmatrix} i - \lambda & c \\ b & j - \lambda \end{pmatrix} = 0 \quad (4.9)$$

$$\lambda^2 - \lambda(i + j) + ij - bc = 0 \quad (4.10)$$

Onde esses λ representam os autovalores e a matriz utilizada é a matriz de covariância.

Com os autovalores em mãos é possível encontrar os autovetores para cada um deles. O equipamento realiza uma transformação linear utilizando a matriz de covariância e os seus autovalores. Todo esse procedimento se faz necessário, pois dessa forma reduzimos a quantidade de variáveis.

A matriz inicial (X) é determinada a partir do mapeamento, o qual pode ser visto na figura 7, e a partir de cada pixel do mapa determina-se um espectro, consequentemente, teremos uma quantidade exagerada de espectros, *Principal Components* e *scores* para interpretar, porém com o PCA podemos reduzir a quantidade de PC's significativamente e dessa forma necessitaremos de poucos *scores*.

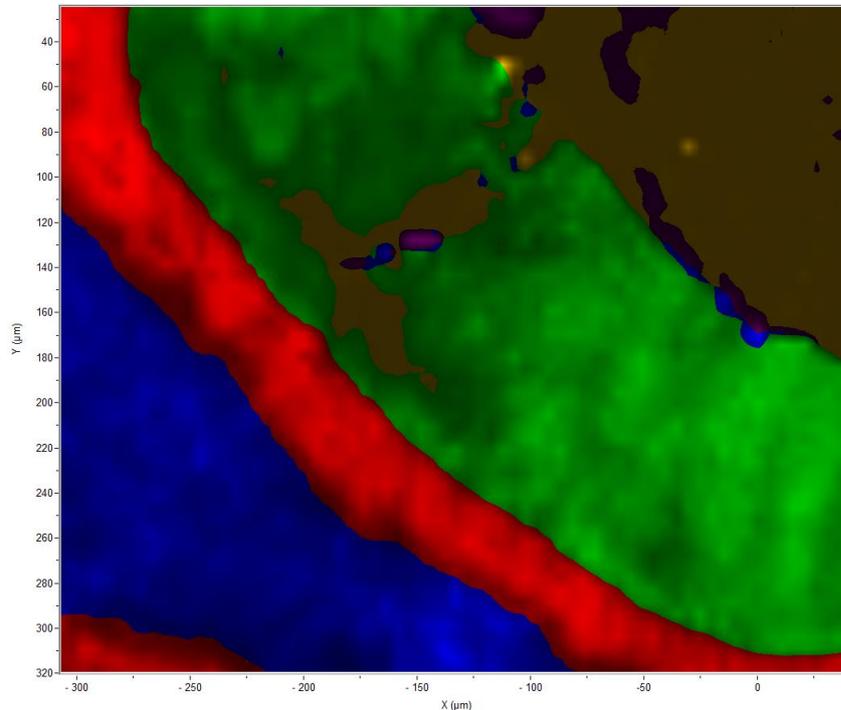
Por meio dos autovetores determinamos os *Principal Components* e com os autovalores temos a dimensão desses PC's. Do primeiro autovalor tem-se o tamanho do PC_1 e do segundo autovalor o tamanho do PC_2 .

A determinação desses scores é importante, pois garante também uma maior variância entre os componentes principais. Para determinar os *scores* usa-se a seguinte equação:

$$[X] \cdot ([P]^t)^{-1} = [T] \quad (4.11)$$

Onde $[X]$ é a matriz ($n \times m$) escalonada com os dados originais, $[P]$ é a matriz ($m \times m$) de autovetores e $[T]$ a matriz ($n \times m$) de escores dos PC's [12].

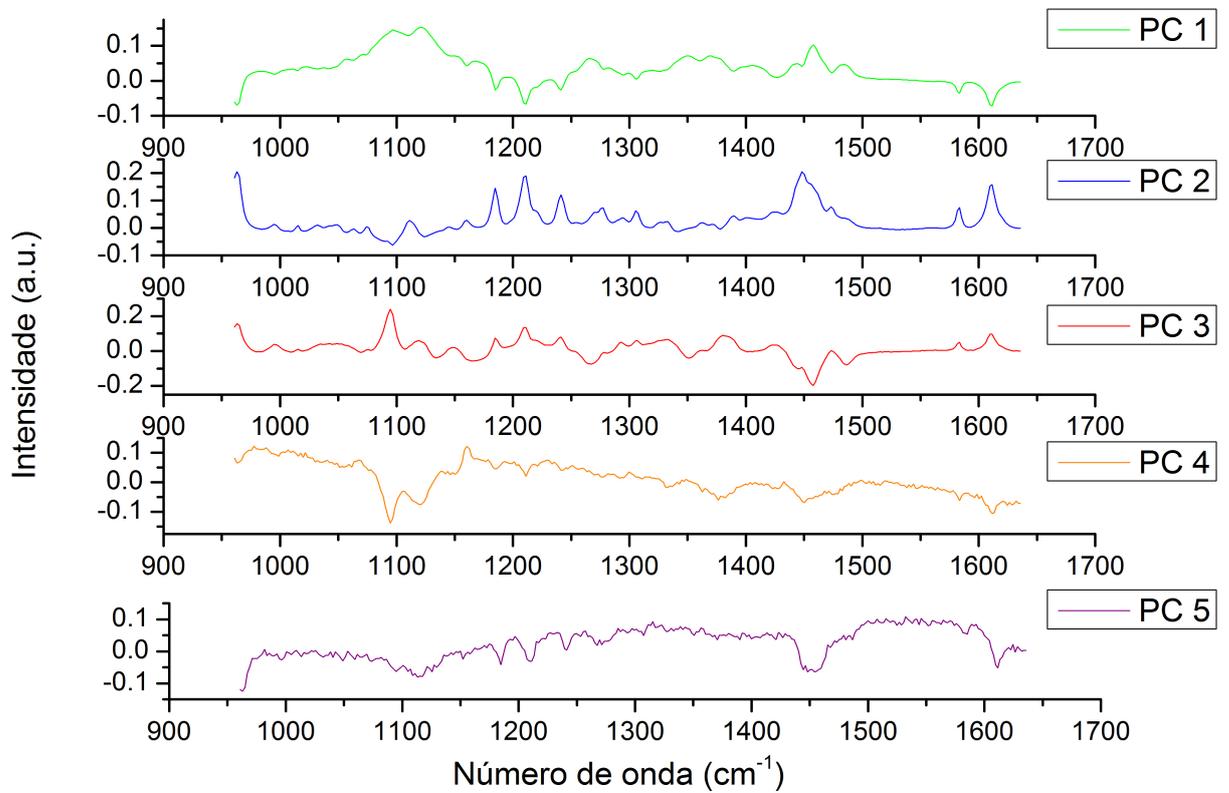
Figura 7 – Mapeamento da região delimitada na figura 6 de uma cápsula de remédio (fármaco).



Fonte: LabRam HR.

A figura que segue abaixo serve para exemplificar os PC's que foram obtidos a partir da figura 7. Tais resultados serão abordados e estudados mais a frente.

Figura 8 – Componentes gerados pela técnica de PCA na amostra exibida na figura 7.



Fonte: LabRam HR.

5 DECOMPOSIÇÃO - MCR

Com uma abordagem semelhante ao do PCA faremos um estudo descritivo do método de *Multivariate Curve Resolution*. Alguns pontos poderão ser omitidos, pois como ele faz parte do processo de Decomposição alguns passos serão idênticos ou semelhantes aos que foram descritos no capítulo anterior.

O MCR é uma ferramenta de pesquisa que nos auxilia a entender o resultado para misturas que apresentam respostas desconhecidas, diferentemente do PCA esse método fornece uma interpretação mais física/química dos resultados, ele é um procedimento mais lento, contudo os *loadings* obtidos são mais quimicamente puros.

A matriz de espectros, determinada a partir do mapeamento é dada por meio da seguinte função:

$$D = C \cdot S^T + E \quad (5.1)$$

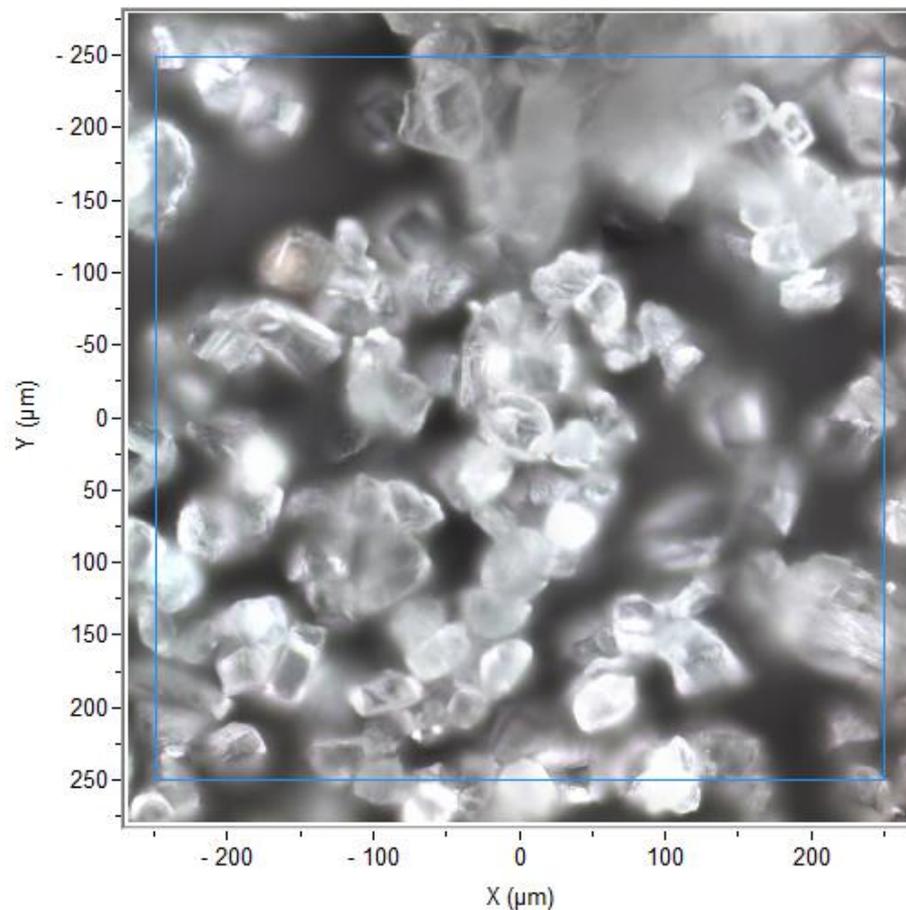
Onde D é a matriz de espectros ($m \times n$), C representa a matriz de concentração, S^T é a transposta da matriz dos compostos puros dos espectros Raman, obtidos da amostra, e E os resíduos. Facilmente, pode-se ver a semelhança dessa função com a equação 3.1 [13] e [14].

No MCR também precisamos determinar o número de componentes ou fatores, assim como no PCA. Para resolver quantos fatores serão necessários o primeiro passo é encontrar n, o número de componentes químicos ou espécies responsáveis pela variância nos dados; passo dois, encontrar os perfis de concentração dessas espécies, ou seja, a matriz C; passo três, determinar o perfil espectral dessas espécies, no caso a matriz S^T . Para isso ser possível assumimos que o posto (*rank*) da matriz de dados é igual ao número de espécies espectroscopicamente ativas, ou seja ele deve ser igual ao número de espécies que produzem sinal analítico presentes na mistura, e isso é possível quando não há outra contribuição como interferência instrumental. Algebricamente, tal número refere-se a linhas ou colunas linearmente independentes. Uma vez que D pode ser decomposto em um produto de duas matrizes, logo o seu posto pode ser determinado da seguinte forma: $\text{posto}(D) \leq \min.[\text{posto}(C), \text{posto}(S^T)]$. Assim se uma das matrizes for de posto completo, basta observar o posto da outra matriz para ter-se uma análise do posto de D [14].

Por meio das figuras 9 e 10, as quais exibem a imagem feita pelo microscópio e pela técnica de MCR, facilmente percebemos uma semelhança com a do PCA. Durante

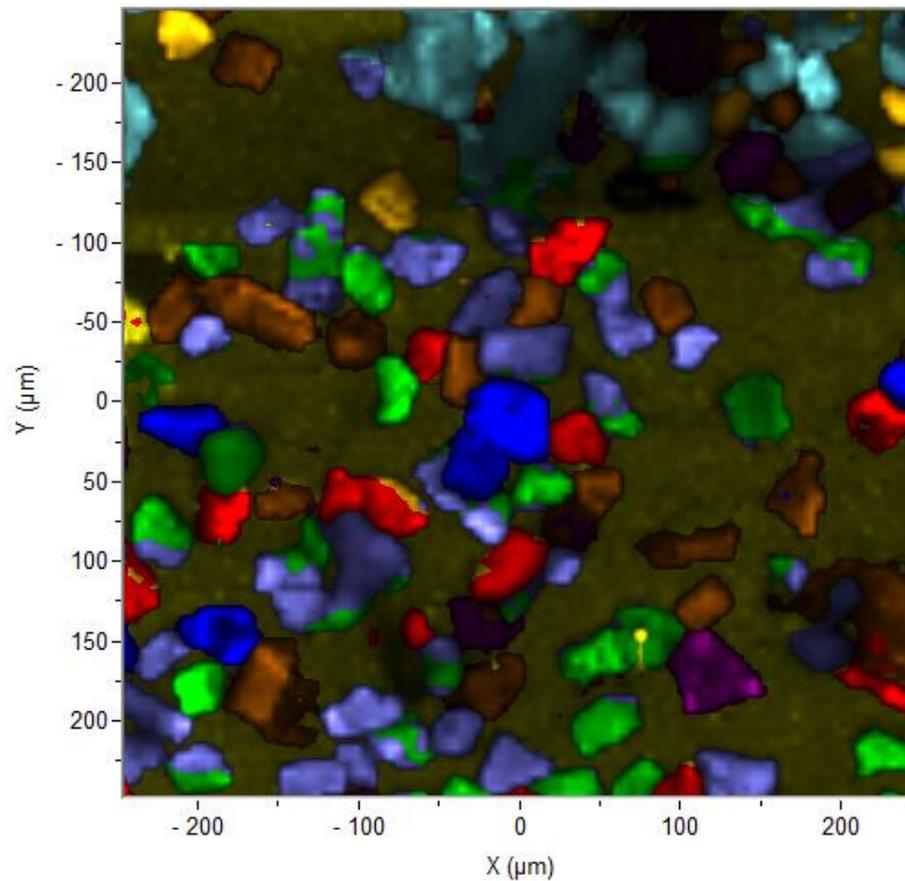
o processo de tal técnica também temos a formação das *scores images* e do gráfico de espalhamento obtido a partir dos *loadings*, contudo diferentemente ao do PCA eles não utilizam os valores negativos presentes. No caso do gráfico de espalhamento dos *loadings* o procedimento para determina-los é semelhante ao descrito anteriormente, a diferença está no tipo da matriz de espectros formada, pois, como foi apresentado, ela possui certas características que as diferenciam da matriz de espectros do PCA e com isso teremos que a maior variância não se inicia mais do primeiro *loading*. E um fator importante com relação as *scores images* é que por meio do *KnowItAll[®] Horiba edition* junto com as bibliotecas espectral da Horiba (a fabricante) podemos identificar o que está presente em cada *loading* de MCR [7].

Figura 9 – Carbonatos com uma região delimitada obtida através do microscópio.



Fonte: LabRam.

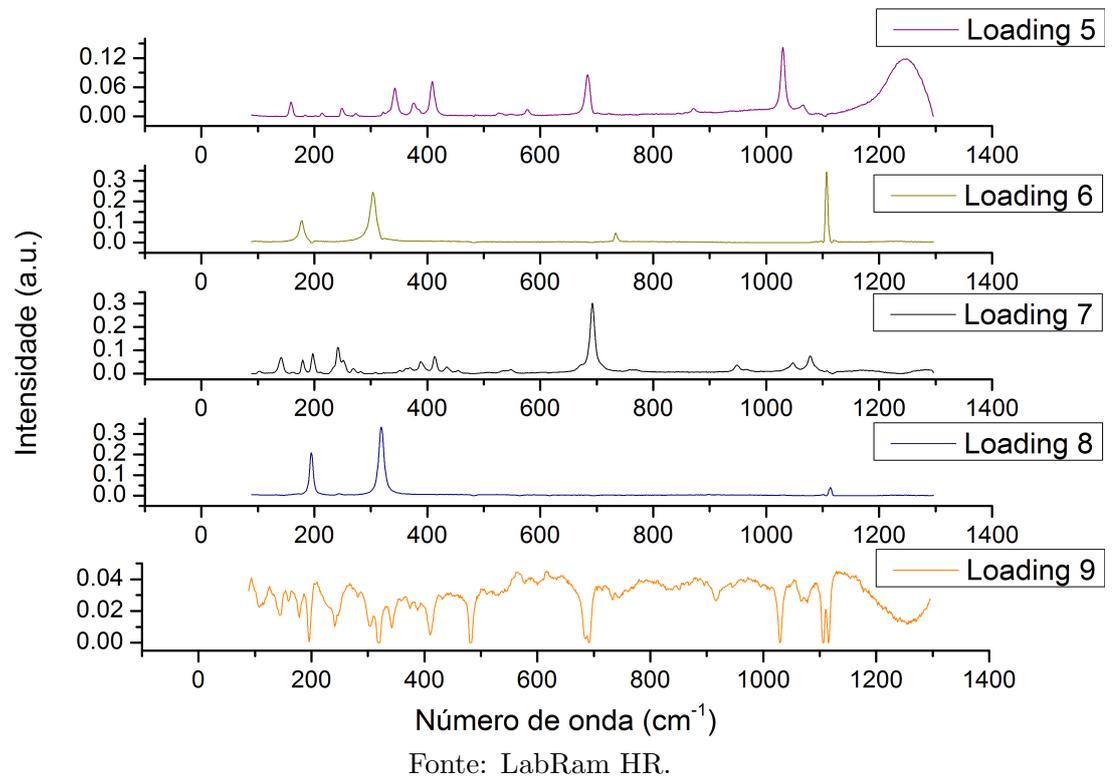
Figura 10 – Mapeamento obtido pela Resolução de Curva Multivariada em carbonatos, onde cada cor representa um espectro específico.



Fonte: LabRam HR.

Semelhante ao PCA, por meio da figura a seguir, a qual exibe os espectros obtidos a partir da análise de MCR na amostra de carbonatos, exemplificamos os tipos de espectros gerados por esse método.

Figura 11 – Loadings gerados pela técnica de MCR na amostra exibida na figura 9.



6 DISCUSSÃO E RESULTADOS

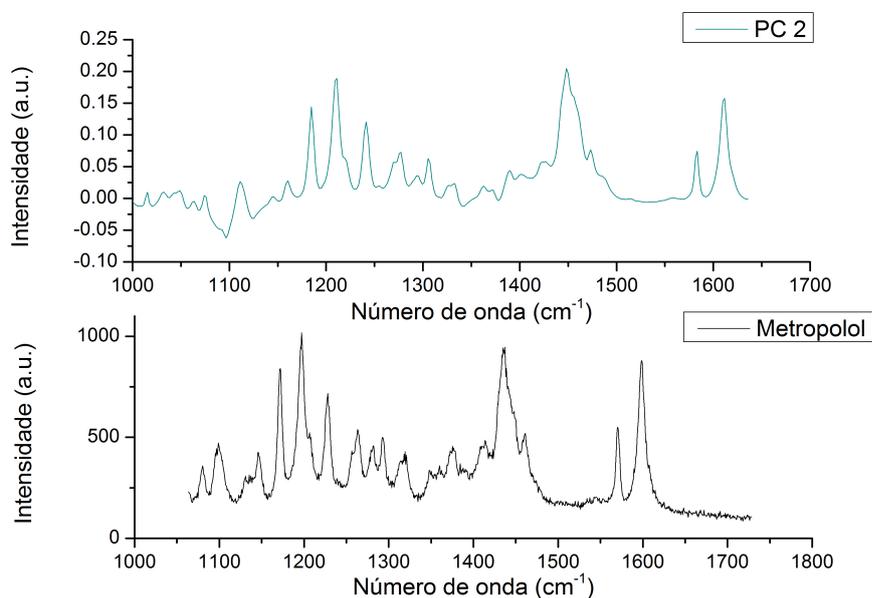
Neste capítulo apresentaremos como os dados foram obtidos, o que podemos retirar dos resultados e por fim os comparamos identificando pontos semelhantes e diferentes.

6.1 Componentes Principais e Loadings (Fármaco)

A Análise do Componente Principal foi usada a fim de fazermos uma análise exploratória da amostra e essa técnica quimiométrica nos permitiu, resumidamente, reduzir a quantidade de dados assim o novo conjunto passa a ser representado com o menor número de novas variáveis, as quais já foram mencionadas anteriormente, são combinações lineares das variáveis originais e que conhecemos por Componentes Principais.

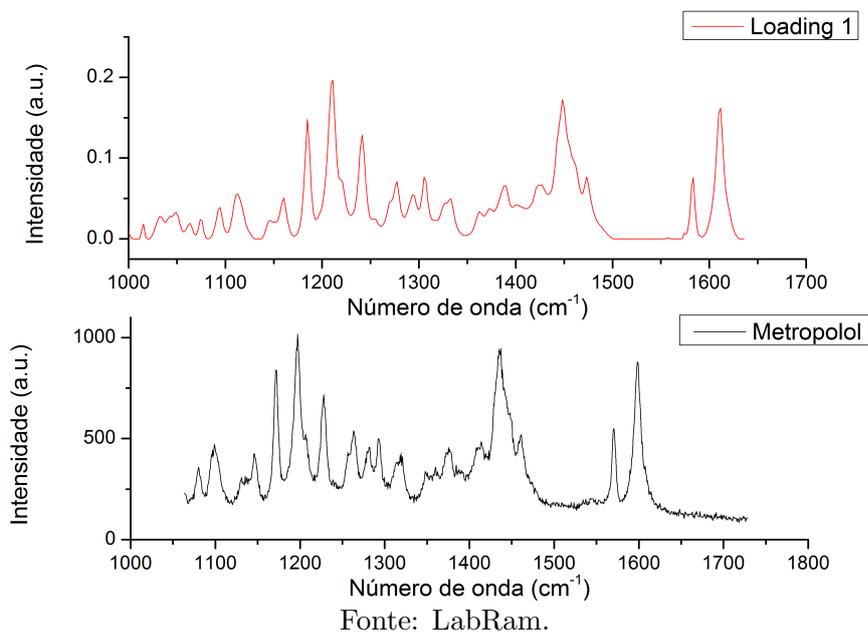
Logo abaixo podemos fazer uma comparação entre a Componente Principal 2, retirada da figura 7 e que representa a segunda maior variância, e o espectro de metropolol.

Figura 12 – Espectro Raman do Metropolol e PC 2.



Fonte: LabRam.

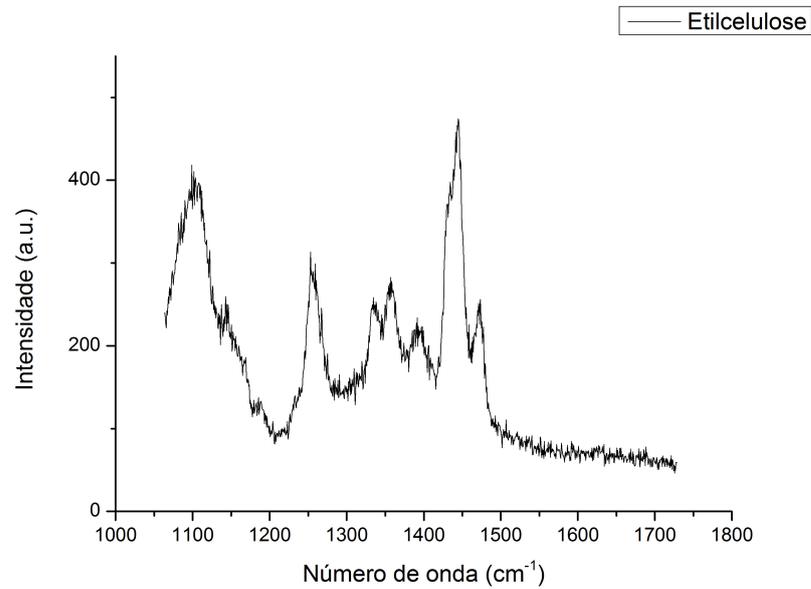
Figura 13 – Espectro Raman do Metropolol e loading 1.



Analisando todos os espectros podemos observar certas semelhanças. Os picos na região entre 1100 cm^{-1} e 1300 cm^{-1} e na região entre 1400 cm^{-1} e 1500 cm^{-1} do metropolol se repetem nos espectros PC 2 e *loading* 1 indicando assim a presença de tal composto na amostra. Para ajudar a enxergar a distribuição do metropolol na figura 7 adotamos a cor azul para o espectro da figura 14 fazendo referência à mesma cor presente no mapeamento. O espectro do metropolol exibido na figura 13 foi obtido por meio também, do LabRam utilizando um feixe com $632,8\text{ nm}$ de comprimento de onda e que repetia 50 vezes (50 acumulações) a medida em cada ponto.

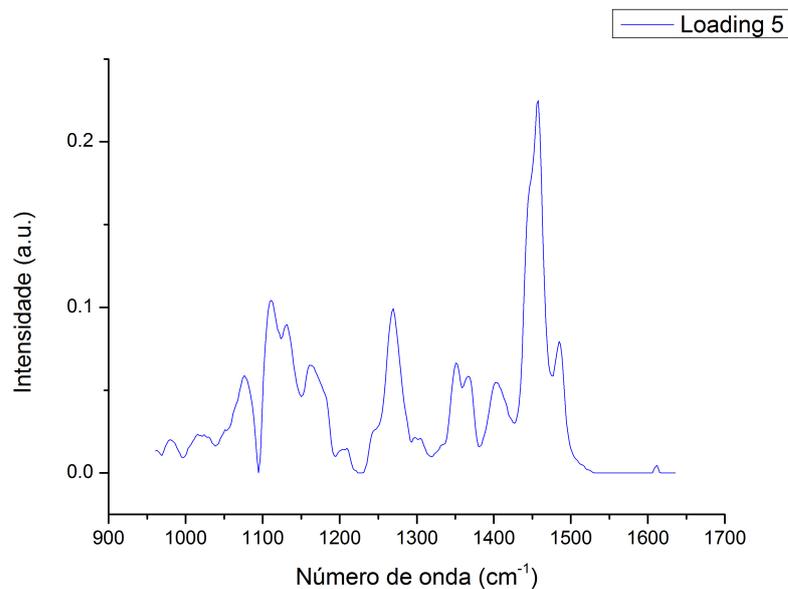
Seguindo a análise dos PC's e dos *loadings* conseguimos identificar outros compostos presentes (Etilcelulose e HPC), porém só foi possível determinar a Etilcelulose por meio do *loading* 5 e o HPC através do PC 1 como pode ser visto por meio das figuras abaixo.

Figura 14 – Espectro Raman da Etilcelulose.



Fonte: LabRam.

Figura 15 – Loading 5.

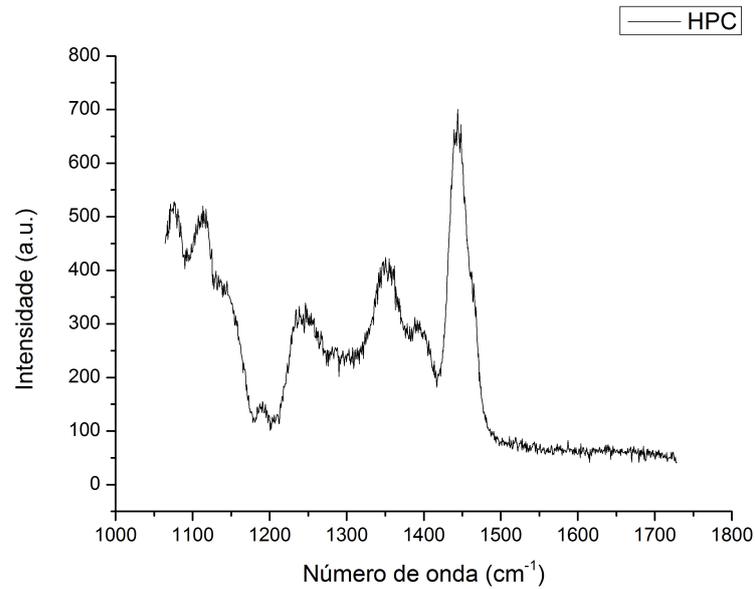


Fonte: LabRam.

Observando os números de onda na região entre 1100 cm^{-1} e 1450 cm^{-1} podemos enxergar os picos que caracterizam tal composto. O mesmo vale para o HPC e o PC 1, contudo a região de números de onda passa a ser de 1100 cm^{-1} a 1400 cm^{-1} .

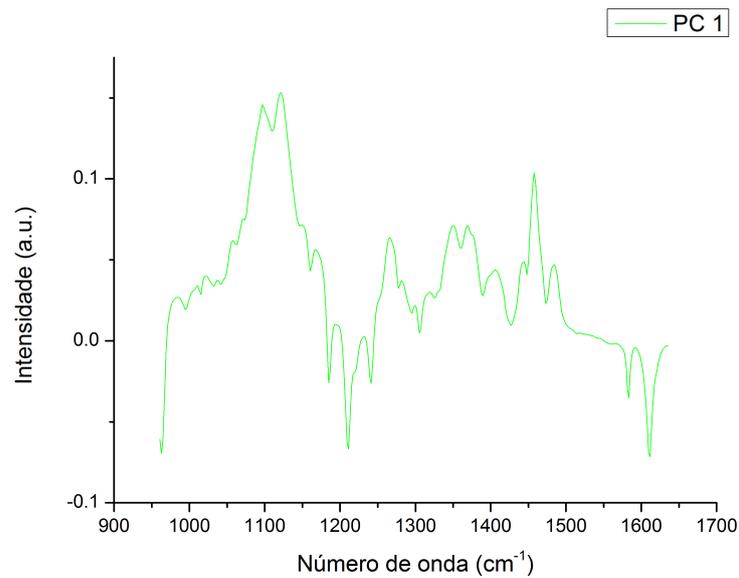
Os espectros tanto da Etilcelulose quanto do HPC obtemos também do LabRam. Para a Etilcelulose usou-se um feixe com comprimento de onda de 785 nm, onde a medida se repetia 50 vezes e para o HPC o comprimento de onda era de 632,8 nm e a medida também se repetia 50 vezes.

Figura 16 – Espectro Raman do HPC.



Fonte: LabRam.

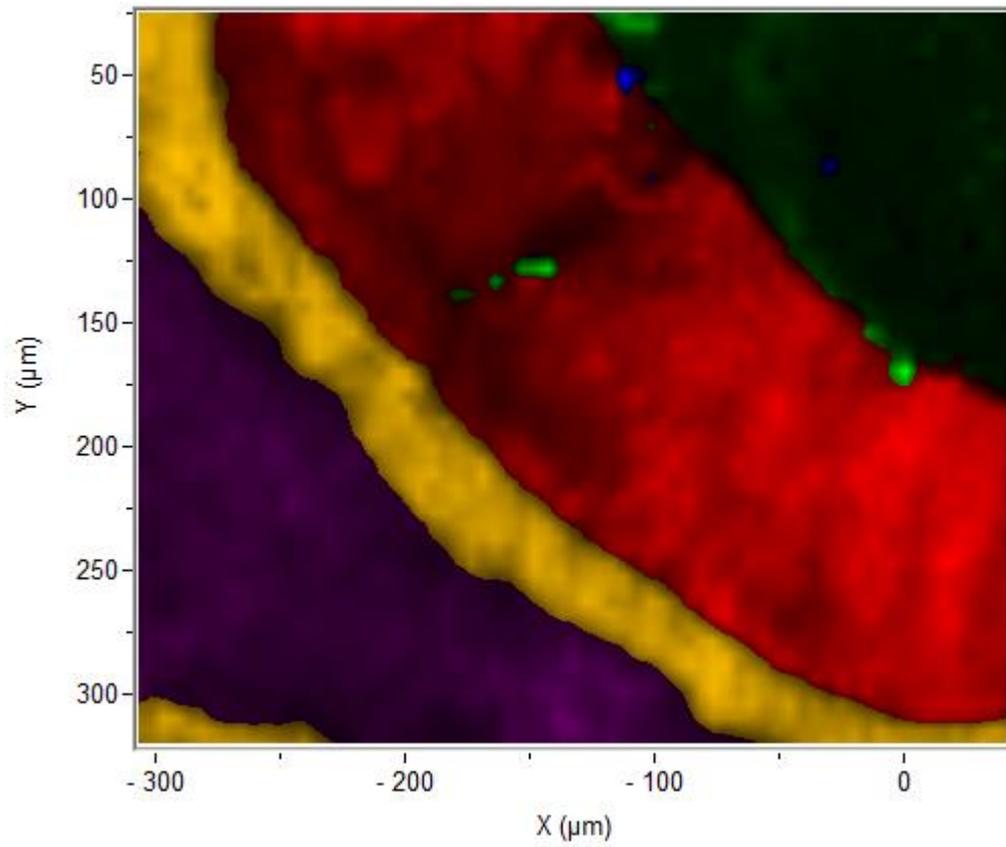
Figura 17 – Componente Principal 1.



Fonte: LabRam.

Observando cada mapeamento apresentado por meio da figura 7 e da figura 18, a seguir, percebemos a semelhança entre elas, onde aqui as usamos para identificar os compostos indicados pelos espectros, no entanto comparando o PC 2 e o *loading* 1, claramente, concluímos que os *loadings* podem apresentar uma resposta mais parecida com o real facilitando na identificação.

Figura 18 – Mapeamento da figura 6 por meio do MCR.



Fonte: LabRam.

7 CONCLUSÃO

Neste trabalho estudamos e descobrimos como são utilizadas as técnicas estatísticas da análise multivariada na espectroscopia vibracional Raman.

Por meio do PCA e MCR conseguimos identificar compostos e como eles estão distribuídos na amostra (fármaco). No entanto, enquanto o MCR nos fornece *loadings* mais parecidos com os espectros das amostras presentes na literatura, ou seja, podemos reconhecê-los como respostas reais instrumentais, já o PCA entrega os resultados bem mais rápidos, mas os PC's não são muito parecidos com os espectros da literatura (as vezes eles não fornecem os PC's corretamente, pois as contribuições mistas reais não possuem a independência estatística ou ortogonal como propriedades naturais).

Esse trabalho se mostrou útil, pois foi possível entender dois métodos importantes na filtração de dados e na classificação dos resultados (espectros). Dessa forma temos conhecimento de mais duas técnicas que poderão contribuir enormemente para futuras pesquisas.

REFERÊNCIAS

- [1] RODRIGUES, Ariano De Giovanni; GALZERANI, José Cláudio. Espectroscopias de infravermelho, Raman e de fotoluminescência: potencialidades e complementaridades. *Rev. Bras. Ensino Fís.*, São Paulo, v. 34, n. 4, p. 1-9, Dec. 2012. Available from http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-11172012000400009&lng=en&nrm=iso. access on 26 Nov. 2015. <http://dx.doi.org/10.1590/S1806-11172012000400009>.
- [2] H. KELAINE CHAVES GOMES, Espectroscopia Raman por Transformada de Fourier e análise de molhabilidade nos filmes finos de carbono amorfo hidrogenado (a-C:H). 2013. Dissertação (Mestrado em Física)-Centro de Ciencias Tecnológicas, Universidade do Estado de Santa Catarina, Joinville, 2013.
- [3] FARIA, D. L. A. de; SANTOS, L. G. C.; GONCALVES, N. S.. Uma Demonstração Sobre o Espalhamento Inelástico de Luz: Repetindo o Experimento de Raman. *Quím. Nova*, São Paulo, v. 20, n. 3, p. 319-323, jun. 1997. Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40421997000300014&lng=pt&nrm=iso acessos em 25. nov. 2015. <http://dx.doi.org/10.1590/S0100-40421997000300014>.
- [4] SALA, Oswaldo. **Fundamentos da Espectroscopia Raman e no Infravermelho**. 2º ed. São Paulo. Unesp. 2008.
- [5] http://crq4.org.br/sms/files/file/Espectroscopia_Raman_4.pdf acesso em 28. jul. 2015.
- [6] Rencher, Alvin C. *Methods of Multivariate Analysis*. 2º ed. Canada. John Wiley & Sons, Inc. 2003
- [7] *Raman Horiba Scientific* LARAT, VICENT. *Multivariate Analysis*. Disponível em <http://www.horiba.com/us/en/scientific/products/raman-spectroscopy/raman-academy/webinars/multivariate-analysis-in-labspec-6-mva-webinar/>. Acesso em: 12 mai. 2015. 2012. 37p.
- [8] *SCIENTIFIC, HORIBA. LabSpec 6: Multivariate Analysis Module*. Horiba Jobin Yvon. Disponível em: http://www.horiba.com/fileadmin/uploads/Scientific/Documents/Raman/SO-TN01_-_LabSpec_6_spectroscopic_software_suite.pdf Acessado em 4 abr. 2015. 2013. 23 p.
- [9] Mooi, Erik e Sarstedt, Marko. *A concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*. Disponível em: http://www.guide-market-research.com/index.php?option=com_content&view=article&id=24&Itemid=38&d4dad6935f632ac35975e3001dc7bbe8=245f7499c16d77fbfa7b0ea81d752136. Acessado em 8. ago. 2015.

- [10] SOUZA, André Marcelo de; POPPI, Ronei Jesus. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I. *Quím. Nova*, São Paulo, v. 35, n. 1, p. 223-229, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422012000100039&lng=en&nrm=iso>. Acessado em 05 Set. 2015. <http://dx.doi.org/10.1590/S0100-40422012000100039>.
- [11] Shlens, Jonathon. *A Tutorial on Principal Component Analysis*. *Computing Research Repository*. abs/1404.1100. 2014. Disponível em: <<http://arxiv.org/pdf/1404.1100>>.
- [12] Disponível em: <http://cosmic.mse.iastate.edu/library/e-notes.html>. Acessado em: 23 abr. 2015.
- [13] MARCO, Paulo Henrique et al. Resolução multivariada de curvas com mínimos quadrados alternantes: descrição, funcionamento e aplicações. *Quím. Nova*, São Paulo, v. 37, n. 9, p. 1525-1532, 2014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422014000900017&lng=en&nrm=iso>. acessado em 15 Mai. 2015. <http://dx.doi.org/10.5935/0100-4042.20140205>.
- [14] GARRIDO, M.; RIUS, F. X.; LARRECHI, M. S. *Multivariate curve resolution-alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes*. *Analytical and bioanalytical chemistry*, v. 390, n. 8, p. 2059-2066, 2008. Disponível em: <http://link.springer.com/article/10.1007/s00216-008-1955-6>. Acessado em: 06 Jun. 2015.