Universidade Estadual de Maringá
Centro de Ciências Exatas
Departamento de Física

Estudo da caracterização de materiais por espectroscopia infravermelha por transformada de fourier e análise de componentes principais

Acadêmico: André Farinha Bósio

Orientador: Marcelo Sandrini

Maringá, October 16, 2018

Universidade Estadual de Maringá
Centro de Ciências Exatas
Departamento de Física

Estudo da caracterização de materiais por espectroscopia infravermelha por transformada de fourier e análise de componentes principais

Trabalho de conclusão de curso apresentada ao Departamento de Física da Universidade Estadual de Maringá, como parte dos requisitos para obtenção do título de Bacharel em Física

Acadêmico: André Farinha Bósio

Orientador: Marcelo Sandrini

Maringá, October 16, 2018

Resumo

Neste trabalho foi realizado uma breve revisão bibliográfica à respeito da espectroscopia no infravermelho por transformada de Fourier(FTIR) e análise de componentes principais(PCA). Com os conhecimentos então estabelecidos foi realizado o PCA dos espectros de absorbância de diferentes soluções de água e etanol, sendo observado que a PC1 representava aproximadamente 99% da variância total, e que a parte negativa e positiva do *loading* correspondente revelou que *scores* mais positivos correspondem à soluções mais concentradas do álcool. Foi visto ainda que a aplicando o PCA ao espectro de absorbância de três sucos de frutas cítricas diferentes foi possível realizar uma diferenciação por agrupamentos com respeito à PC1 e PC2, que representaram 91.45% e 7.39% da variância total.

Palavras chave: Espectroscopia, FTIR, PCA

Abstract

In this work, we have a brief literature review concerning the Fourier transform infrared spectroscopy (FTIR) and principal component analysis (PCA). With the knowledge established, the PCA of the absorbance spectra of different solutions of water and ethanol was done, with PC1 representing approximately 99% of the total variance, and that the negative and positive part of the corresponding loading revealed that more positive scores correspond to more concentrated solutions. It was also observed that by applying PCA to the absorbance spectra of three different types of citric fruits it was possible to differentiate them by clusters, with respect to PC1 and PC2, which represented 91.45% and 7.39% of the total variance.

Key words: Espectroscopy, FTIR, PCA

Sumário

1 Introdução

Dentre os diversos ramos da ciência, a espectroscopia ótica é uma das áreas que tem destaque no estudo de materiais. Essa área refere-se a um conjunto de métodos experimentais nos quais é observada interação entre radiação eletromagnética e a matéria. Dentre essas interações podemos citar a absorção, transmissão ou reflexão em diferentes comprimentos de onda, fornecendo em cada caso um espectro característico para o material analisado, o que possibilita a determinação da composição e a estrutura dos tipos de ligações que compõem o material de estudo. No caso da espectroscopia na região do infravermelho (IR - do inglês *infra red*), tal método é amplamente aplicado em estudos moleculares devido a absorção da luz nesta região estar relacionada às vibrações das ligações químicas [1]. O espectro do infravermelho pode ser divido em três regiões: infravermelho médio, infravermelho próximo e distante. Em função do tipo de amostra estudada nesse trabalho, foi utilizada a região do infravermelho médio, que é definida entre 4000 e 400 cm^{-1}. Este intervalo ainda pode ser sub-divido em quatro regiões, sendo definidas da seguinte forma: intervalo entre 4000 e 2500 cm^{-1}, onde são detectados os estiramentos entre os átomos envolvendo ligações do tipo H-X; intervalo entre 2500 e 2000 cm^{-1}, região das ligações triplas; intervalo entre 2000 e 1500 cm^{-1}, região das ligações duplas; intervalo entre 1500 e 600 cm^{-1}, responsável pela região chamada de impressão digital [2].

As características mais atrativas da espectroscopia IR estão relacionadas ao fato da maioria das técnicas serem de natureza direta, não destrutivas e de exigirem pouco ou nenhum tipo de pré-tratamento da amostra. Esses fatores, conferem inúmeras aplicações a espectroscopia no IR nas mais diversas áreas, como por exemplo, nas áreas agrícola, alimentícia, ambiental, farmacêutica, biomédica, nas industrias textil, de polimeros, de óleo e gás, entre outras. [2–4]

Apesar das vantagens provindas das técnicas de espectroscopia IR citadas anteriormente, ainda assim, algumas problemas ou dificuldades podem ocorrer quando utilizadas. Por exemplo, uma análise quantitativa de espectros de IR muitas vezes requer o tratamento de centenas de absorções (picos ou bandas) registradas em um único espectro, a depender da precisão do equipamento. Outro problema surge quanto a diferença entre os espectros dos materiais estudados for muito sutil, com diferenças quase residuais por exemplo. Com isso, muitas vezes se faz necessária a utilização de procedimentos matemáticos e estatísticos para extrair informação relevantes dos dados.

Um procedimento muito aplicado nesses casos é o de análise de componentes principais. Trata-se de uma análise multivariada capaz de estudar a relação de interdependência, por meio de matrizes de covariância e/ou cor-

relação de um conjunto de variáveis de um sistema. Esse tipo de tratamento é útil para a investigação de um grande número de dados, possibilitando diferenciar os materiais estudados, além de, em alguns casos, decompor os espectros dos materiais responsáveis pela diferenciação.

A aplicação deste método em medidas de espectroscopia não é algo novo, encontra-se na literatura trabalhos envolvendo o tema desde a década de 50 [5], e ainda hoje é uma ferramenta poderosa como método exploratório, usado em pesquisas industriais na identificação de adulterações de produtos [6], caracterização de materiais quanto a origem [7], assim como análises qualitativas mais robustas como, além da diferenciação de dois materiais, qual é o composto que causa essa diferenciação [8].

1.1 Objetivos

Os objetivos de presente trabalho são:

- Compreender o funcionamento e os fundamentos da espectroscopia no infravermelho por transformada de Fourier (FTIR-*Fourier-transform infrared spectroscopy*);
- Compreender o método de análise de componentes principais (PCA-*Principal Component Analysis*);
- Apresentar a aplicação do método de PCA por meio de um exemplo numérico simples;
- Realização de um estudo quimiométrico qualitativo do espectro de absorção de misturas de água e álcool etílico;
- Realização de um estudo quimiométrico qualitativo simples do espectro de absorção de 3 frutas cítricas;

2 Revisão bibliográfica

2.1 A Transformada de Fourier num contexto experimental

O método da transformada de Fourier é um procedimento matemático no qual uma certa função, ou sinal, do ponto de vista experimental, é decomposto no espectro de frequências que o compõe. Tal procedimento é interessante para processamento de sinais elétrico oscilantes, com aplicação em filtros de ruído e análises de interferometria. As transformadas se dão pela aplicação das equações abaixo [9],

$$F(x) = \int_{-\infty}^{+\infty} F(\omega) e^{2\pi i x \omega} d\omega \tag{1}$$

$$F(\omega) = \int_{-\infty}^{+\infty} F(x) e^{-2\pi i x \omega} dx \tag{2}$$

onde $F(x)$ é a função original, isto é o sinal obtido, e $F(\omega)$ o espectro de frequências que compõem tal sinal.

Uma abordagem interessante sobre visualização da transformada consiste em imaginar um espaço 3D, onde duas dimensões definem um plano constituído pelas amplitudes do sinal em relação ao tempo, e a terceira dimensão definida pelo espectro das frequências que o compõe, como ilustra a figura 1.

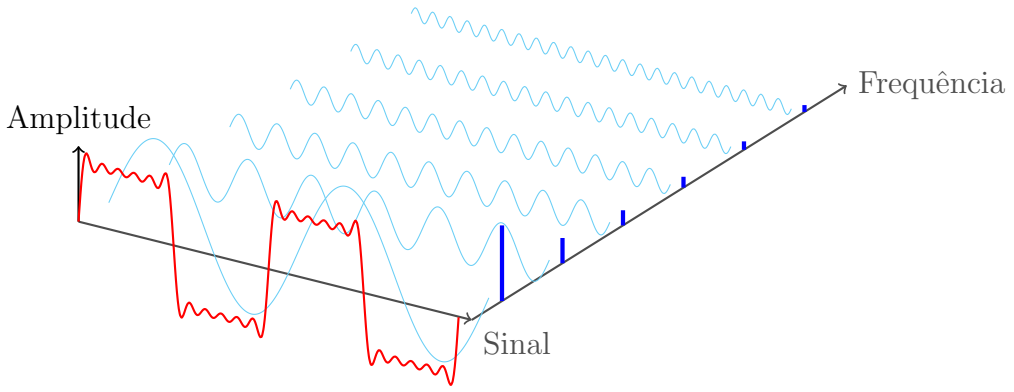

Figura 1: O sinal recebido, em vermelho, composto por uma soma de diferentes senoides, e em azul a transformada do sinal, decompondo as frequências constituintes

2.2 Espectroscopia em infravermelho por transformada de Fourier

A luz infravermelha é caracterizada por comprimentos de onda entre 700 nm e 0,25 cm, de maneira que nessa faixa a absorção está relacionada às vibrações moleculares [1]. A espectroscopia no infravermelho por transformada de Fourier (FTIR) é um método experimental que fornece o espectro de absorção e/ou transmitância do material analisado. Trata-se de uma técnica bastante sensível para a identificação de compostos orgânicos em uma ampla gama de aplicações, sejam eles sólidos, líquidos, pós ou géis [10].

O FTIR utiliza o fundamento da interferometria - um fenômeno ondulatório no qual a superposição de duas ou mais ondas com uma diferença de fase, devido à uma diferença de caminho ótico, gera um padrão de intensidades característico. Apesar de existirem diversos tipos de interferômetros, a maioria tem o princípio de funcionamento similares ao modelo desenvolvido por Michelson [1]. A montagem do interferômetro de Michelson consiste de uma fonte de luz monocromática, um espelho fixo, um espelho móvel, um divisor de feixe (*beamsplitter*) e um anteparo ou sensor arranjados como mostra a figura 2.

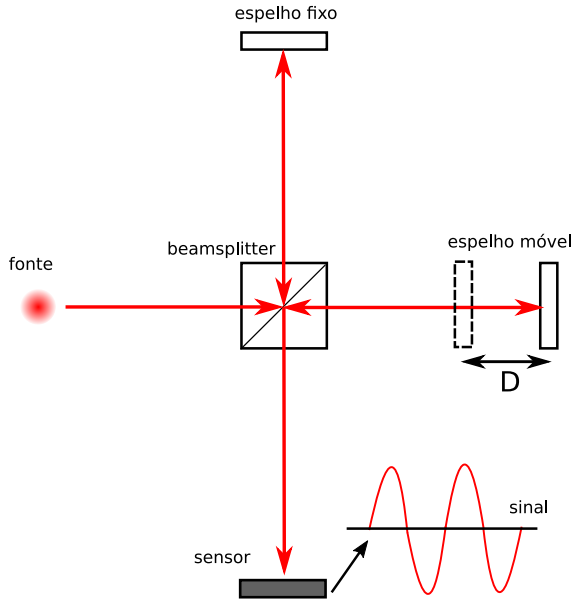

Figura 2: Montagem do interferômetro de Michelson.

Nesse interferômetro, o feixe proveniente da fonte é dividido em dois ao atravessar o *beamsplitter*. Um dos feixes é incidido no espelho móvel enquanto o outro feixe é incidido no espelho fixo. Depois de serem refletidos pelos

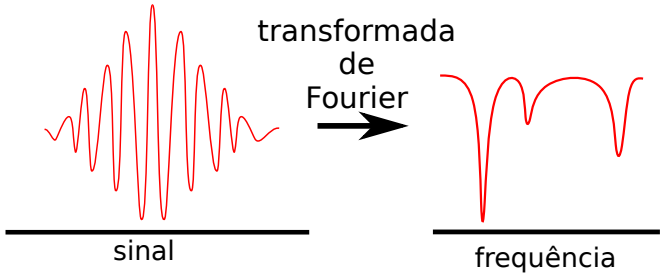

Figura 3: Ideia gráfica da transformada do interferograma proveniente do FTIR

espelhos, os feixes recombinam-se novamente e são direcionados para o sensor onde é formado um padrão de interferência. Esse padrão é gerado em função da diferença entre as distâncias percorridas pelos feixes ($\delta = d_1 - d_2$). Quando os feixes se recombinam no sensor e ambos estão em fase, então a interferência é dita construtiva e a amplitude da onda tem sinal máximo, isso ocorre quando $\delta = n\lambda$, em que $n = 1, 2, 3, ...$ (número inteiro). Por outro lado quando os feixes estão em fases opostas, então a interferência é dita destrutiva e a amplitude da onda final será igual a zero, ocorre quando $\delta = (n - 1/2)\lambda$. Com isso, o padrão de interferência obtido ao mover o espelho é uma senoide com a frequência do laser utilizado, de tal forma que usando a transformada de Fourier é possível passar do espaço original para o espaço das frequências, retornando assim uma função delta de Dirac.

Caso houvesse dois comprimentos de onda ou mais sendo emitidos pela fonte, o padrão de interferência seria uma composição de senoides de com as frequências individuais emitidas pela fonte. Tal raciocínio pode ser expandida para uma fonte de espectro contínuo, de forma que a transformada do padrão de interferência teria como resultado um espectro de frequências característico do feixe.

O FTIR usa o mesmo principio do intereferômetro de Michelson, com a diferença que a fonte de luz é infravermelha, contínua, e que o feixe de luz passa pela amostra a ser analisada, como mostra a figura 4. Como algumas frequências agora foram absorvidas, o interferograma é diferente do esperado para uma luz de espectro sem a amostra, e aplicando a transformada de Fourier nesse novo sinal é possível observar quais frequências foram absorvidas e quais foram transmitidas, bem como suas intensidades, como mostra a figura 3.

Unindo isso ao fato de que a absorção no espectro IR é relacionado às vibrações moleculares, então, é possível caracterizar quais os tipos de ligações estão presentes na material estudado, abrindo possibilidades para uma investigação estrutural, assim como um método de comparação simples comparando o espectro observado com uma referência de um banco de dados [1].

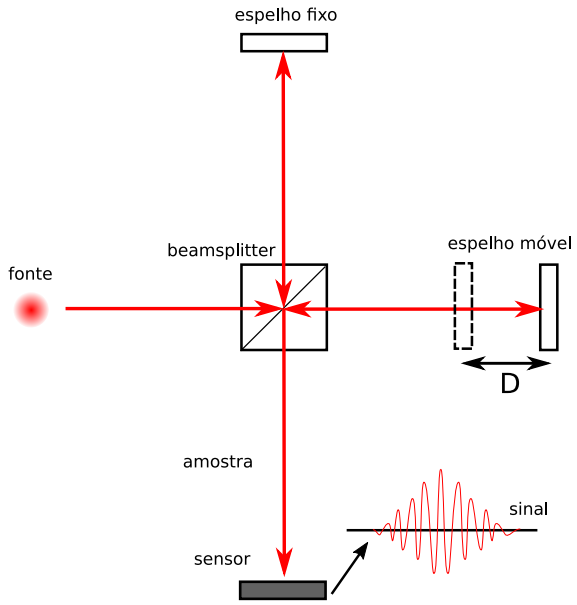

Figura 4: Montagem usual de um FTIR

2.3 Refletância Total Atenuada (ATR)

Outra montagem possível do FTIR é usando a reflexão total atenuada, ou ATR - do inglês *attenued total reflectance*, ilustrado pela figura 5.

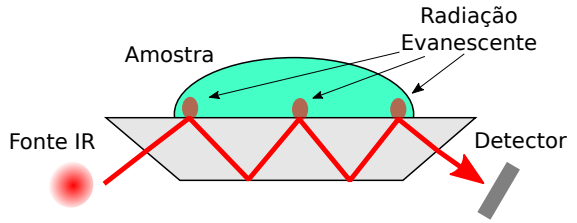

Figura 5: A luz penetra o cristal, sofre sucessivas reflexões totais e, na interface entre o cristal e a amostra a radiação evanescente consegue penetrar uma pequena porção da amostra e ser absorvida

Nesse arranjo a amostra é posicionada sobre um cristal com alto índice de refração, a luz então é incidida em um certo ângulo, de forma que a mesma sofra ao menos uma reflexãs internas totais e, durante esse processo, uma pequena porção da luz penetra na amostra pelo fenômeno da radiação evanescente, de forma que a radiação extrapolada interage com a amostra e é durante esse processo que se da a absorção da luz [1].

Uma das grandes vantagens dessa configuração FTIR-ATR, e motivo pelo qual essa técnica foi usada no presente trabalho, é a facilidade da realização da medida, sem necessidade de pré-tratamento das amostras, ela não exige

que as amostras sejam acomodadas em cubetas, no caso de líquidos, ou que sejam pastilhadas, para o caso de amostras sólidas.

2.4 Análise de componentes principais

O método de análise de componentes principais (PCA - do inglês *Principal component analysis*) é um processo estatístico de análise multivariada, que visa reduzir a dimensionalidade de uma grande base de dados interrelacionados. Tal processo é feito por uma transformação que gera os componentes principais, ou PC's, de maneira que estes novos dados não são correlacionados e os primeiros estão atribuídos a maior variância da informação [11]. Pensando num caso bidimensional, ou seja duas propriedades (x e y) de n amostras, as PC's representam um novo sistema de coordenadas, o qual se alinha melhor com os dados de forma a expressar a maior variância do sistema [11], como ilustra a figura 6.

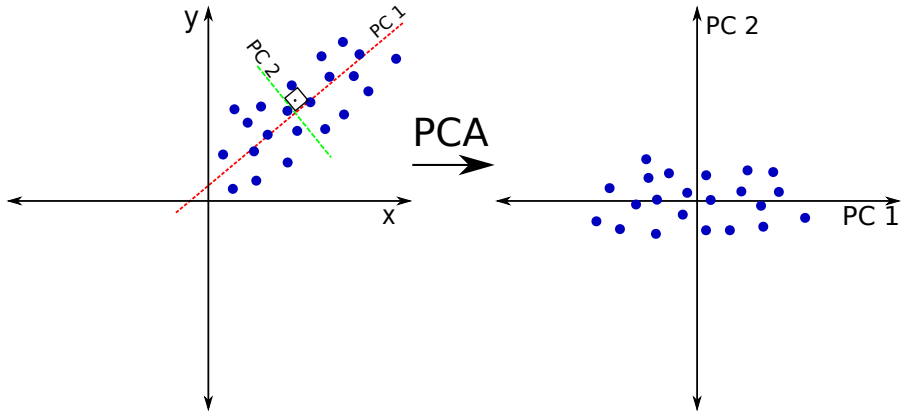

Figura 6: Rotação de um conjunto de amostras bidimensionais, utilizando o PCA

Antes de partirmos para o processo do PCA em si, é importante definir duas coisas, covariância e correlação entre dois tipos de dados X e Y. A covariância é definida pela equação3

$$\sigma_{xy} = \sum_{i}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n} \tag{3}$$

onde $\bar{X}$ e $\bar{Y}$ são as médias, X_i e Y_i as i-ésimas medidas e n o número total de medidas. Tal propriedade é um fator da relação entre o movimento da média entre as variáveis. Se a covariância é positiva os dados tem um comportamento parecido, o maior valor da primeira variável corresponde ao maior valor da segunda por exemplo. Já uma covariância negativa indica uma relação inversa. Por fim, se tal valor for nulo indica uma possível ortogonalidade entre as variáveis, entretanto não é uma conclusão definitiva [12].

É importante notar que a covariância depende da escala das variáveis, ou seja, se por exemplo, duas medidas relacionadas a comprimento foram feitas, uma em centímetros e outra em metros, apenas o valor bruto será levado em conta. Para solucionar esse problema em casos indesejados, temos a correlação, definida pela equação 4, que é a covariância normalizada pelo produto dos desvio padrão das variáveis. Eliminando a influência da escala das observações.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4}$$

Tendo essas duas definições em mãos podemos gerar as chamadas matrizes covariância(C), ou correlação(S), sendo elas simétricas, como mostrado abaixo, exemplificando para três dimensões observadas (*x, y, z*). É importante ainda ressaltar que, devido a simetria, elas possuem tantos autovalores e autovetores quanto suas ordens, e estes são ortogonais entre sí [13].

$$C = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} S = \begin{pmatrix} \rho_{xx} & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & \rho_{yy} & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & \rho_{zz} \end{pmatrix} \tag{5}$$

Passando agora para o processo em si. Vamos supor um grupo de n amostras com p propriedades medidas, formando os vetores $\vec{y_1}, \vec{y_2}, ...\vec{y_n}$, nos quais cada linha é o valor relacionado, de forma que os elementos y_{ij} representam o valor da medida associada a p-ésima dimensão da amostra i.

$$\vec{y_1} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1p} \end{pmatrix} \vec{y_2} = \begin{pmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2p} \end{pmatrix} \vec{y_n} = \begin{pmatrix} y_{n1} \\ y_{n2} \\ \vdots \\ y_{np} \end{pmatrix} \tag{6}$$

Do pondo de vista geométrico, tais conjuntos de dados formam um tipo de elipsoide p-dimensional. O método de PCA, busca então uma nova base, na qual os componentes principais se alinhem com os semi-eixos da dita elipsoide, com seu centro na média das dimensões observadas, de maneira que a PC_1 corresponde ao maior semieixo, a PC_2 ao segundo maior e assim por diante até a PC_p. Para que a rotação ocorra, a primeira coisa a se fazer é a centralização dos dados originais em suas médias, dessa forma o dito elipsoide estará centrado na origem, resultando nos novos vetores:

$$\vec{y_1} = \begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ \vdots - \vdots \\ y_{1p} - \bar{y}_n \end{pmatrix} \vec{y_2} = \begin{pmatrix} y_{21} - \bar{y}_2 \\ y_{22} - \bar{y}_2 \\ \vdots - \vdots \\ y_{2p} - \bar{y}_2 \end{pmatrix} \vec{y_n} = \begin{pmatrix} y_{n1} - \bar{y}_1 \\ y_{n2} - \bar{y}_2 \\ \vdots - \vdots \\ y_{np} - \bar{y}_p \end{pmatrix} \tag{7}$$

A rotação do vetor $\vec{y_i}$, que define as propriedades de uma amostra, pode ser feita usando uma matriz ortogonal, ou seja, uma matriz V quadrada que obedece a propriedade $V^tV = VV^t = I$, onde I é a matriz identidade de mesma ordem de V [13], gerando assim um novo vetor rotacionado $\vec{z_i} = V\vec{y_i}$.

Tal rotação deve acontecer de forma que o novo arranjo de dados não estejam correlacionados e que os mesmos apresentem a maior variância possível, ou seja, a matriz covariância do espaço das PC's (C_v) deve ser diagonal, indicando que todos elementos referentes à covariância sejam nulos. Em razão disso, a matriz rotação V tem que diagonalizar a matriz covariância original(C), isto é, a matriz V em questão pode ser construída utilizando os autovetores ($\vec{v_i}$) da matriz original como linhas, organizando-os em ordem decrescente com relação aos seus respectivos autovalores ($\lambda_1 > \lambda_2 > \lambda_3 > \lambda_p$).

$$\vec{v_1} = \begin{pmatrix} v_{11} \\ v_{12} \\ \vdots \\ v_{1p} \end{pmatrix} \vec{v_2} = \begin{pmatrix} v_{21} \\ v_{22} \\ \vdots \\ v_{2p} \end{pmatrix} \vec{v_p} = \begin{pmatrix} v_{p1} \\ v_{p2} \\ \vdots \\ v_{pp} \end{pmatrix} \tag{8}$$

$$V = \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1p} \\ v_{21} & v_{22} & \dots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \dots & v_{pp} \end{pmatrix} \quad C_z = VCV^t = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_p \end{pmatrix} \tag{9}$$

Com a nova base do espaço das PC's temos os *scores*, que representam as projeções dos dados originais sobre o os novos eixos dos PC's. A posição z no espaço das PC's com relação a PCj de uma amostra i é dada por

$$z_{ij} = \vec{v_j}^t y_i = \begin{pmatrix} v_{j1} v_{j2} \dots v_{jp} \end{pmatrix} \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix} = v_{j1} y_{i1} + v_{j2} y_{i2} + \dots + v_{jp} y_{ip} \qquad (10)$$

Conclui-se que as PC's correspondem aos autovetores da matriz covariância original [12] .

Retornando agora para os variâncias das PC's, ou ainda, os autovalores já que os elementos da diagonal principal são as variâncias, é possível ver que devido à organização decrescente temos as primeiras componentes explicando a maior variância dos dados, de forma que é possível calcular proporção da variância(P), equação 11, total dos dados usando k componentes principais [11], se P for grande usando um k pequeno, 3 por exemplo, é possível em princípio usar somente as k componentes, devida a uma alta proporção de variância

$$P = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \qquad (11)$$

Dessa maneira, é possível interpretar os dados utilizando suas componentes principais, ao invés dos dados brutos, visto que no processo as informações redundantes ou pouco relevantes são removidas, restando usualmente apenas 3 ou 4 componentes para serem analisadas. Essas componentes em geral correspondem a uma grande percentagem da variância total, de modo que as demais poderiam ser descartadas em primeira instância. Entretanto, tal procedimento requer uma análise mais profunda, dependendo do contexto do que esta sendo estudado.

Outro fator importante são os *loadings* dos PC's, que representam os coeficientes da combinação linear dos dados originais, isto é, os elementos do autovetor correspondente à cada PC, nele podemos encontrar seu significado, embora este esteja atrelado ao contexto do estudo. Um significado amplamente usado é com relação a análise multivariada de espectros, onde os *loadings* mostram informações dos espectros que diferenciam os materiais analisados.

2.5 Quimiometria

Uma aplicação direta da análise de componentes principais diz respeito às técnicas de quimiometrias, ou seja, um estudo estatístico de um conjunto de dados químicos. Um objeto extremamente usado para tal estudo diz

respeito à espectroscopia IR. Aplicando o processo de PCA's aos espectros adquiridos e, interpretando que cada comprimento de onda corresponde a uma dimensão, é possível diferenciar vários tipos de materiais com espectros semelhantes, tipos de óleos [6], cafés [8], diferentes tipos de frutas [14], e ainda, em alguns casos, identificar quais compostos isolado que causam essa diferença [8].

Para mostrar a aplicação algébrica do método, vamos usar um suposto conjunto com 11 amostras: 5 de um tipo A conhecido, 5 de um tipo B, também conhecido, e uma amostra U desconhecida, a qual queremos identificar. A partir dos espectros de absorção separa-se as intensidades avaliadas em 4 comprimentos de onda específicos: $1200cm^{-1}$,$1100cm^{-1}$,$1000cm^{-1}$ e $900cm^{-1}$, conforme apresentado na tabela 1.

	Intensidades (u.a.)			
Amostra	1200 cm^{-1}	1100 cm^{-1}	1000 cm^{-1}	900 cm^{-1}
A1	10.1	15.4	80.2	70.4
A2	9.8	14.3	80.1	69.1
A3	14.4	16.2	78.4	72.1
A4	11.5	14.9	77.3	70.8
A5	10.1	15.1	81.7	71.2
B1	70.1	12.3	10.8	10.1
B2	71	11.8	11.1	12
B3	72	11.7	12	15.8
B4	73.9	11.5	12.1	12
B5	72.1	10.2	11.9	11.7
U1	10.7	13.4	79.4	71.5

Tabela 1: Intensidades das bandas de absorção em $1200cm^{-1}$,$1100cm^{-1}$,$1000cm^{-1}$ e $900cm^{-1}$ para cada uma das amostras.

Primeiramente, centralizamos os dados em suas médias resultando num novo arranjo, tabela 2.

Amostra	$1200cm^{-1}$	$1100cm^{-1}$	$1000cm^{-1}$	$900cm^{-1}$
A1	-28.6	2.054	31.563	26.154
A2	-28.9	0.954	31.463	24.854
A3	-24.3	2.854	29.763	27.854
A4	-27.2	1.554	28.663	26.554
A5	-28.6	1.754	33.063	26.954
B1	31.4	-1.045	-37.836	-34.145
B2	32.3	-1.545	-37.536	-32.245
B3	33.3	-1.645	-36.636	-28.445
B4	35.2	-1.845	-36.536	-32.245
B5	33.4	-3.145	-36.736	-32.545
U1	-28	0.0545	30.763	27.254

Tabela 2: Valores das intensidades centralizados com média zero

Além disso, devido as diferenças das intensidades da tabela 1, assim como, a variação em números absolutos na tabela 2 é relevante, então para a análise usaremos a matriz covariância M:

$$M = \begin{pmatrix} 1007.820 & -55.726 & -1125.548 & -968.248 \\ -55.726 & 3.826 & 62.492 & 54.124 \\ -1125.548 & 62.492 & 1260.056 & 1084.560 \\ -968.248 & 54.124 & 1084.560 & 936.598 \end{pmatrix} \tag{12}$$

Nela, a diagonal principal representa as variâncias totais de cada comprimento de onda, e os elementos não diagonais representam a covariância. A matriz M ainda possui 4 autovalores distintos, assim como 4 autovetores, ordenados abaixo em ordem decrescente

$$\lambda_1 = 3203.405 \quad \lambda_2 = 3.357 \quad \lambda_3 = 0.864 \quad \lambda_4 = 0.673 \tag{13}$$

$$v_1 = \begin{pmatrix} 0.560 \\ -0.031 \\ -0.627 \\ -0.540 \end{pmatrix} v_2 = \begin{pmatrix} 0.652 \\ 0.114 \\ -0.066 \\ 0.746 \end{pmatrix} v_3 = \begin{pmatrix} -0.506 \\ 0.241 \\ -0.755 \\ 0.338 \end{pmatrix} v_4 = \begin{pmatrix} -0.067 \\ -0.963 \\ -0.176 \\ 0.191 \end{pmatrix} \tag{14}$$

Com isso, podemos ver a posição das amostras projetados nas componentes principais, com resultados apresentados na tabela 3. Nesse caso temos que a projeção da n-ésima amostra na PC_k (Z_{n1},Z_{n2},Z_{n3} e Z_{n4}) é dada pelas equações abaixo, onde I_x corresponde a absorbância centralizada da enésima amostra no comprimento de onda x.

$$
\begin{aligned}
Z_{n1} &= 0.560I_{1200} - 0.031I_{1100} - 0.627I_{1000} - 0.540I_{900} \\
Z_{n2} &= 0.652I_{1200} + 0.114I_{1100} - 0.066I_{1000} + 0.746I_{900} \\
Z_{n3} &= -0.506I_{1200} + 0.241I_{1100} - 0.755I_{1000} + 0.338I_{900} \\
Z_{n4} &= -0.067I_{1200} - 0.963I_{1100} - 0.176I_{1000} + 0.191I_{900}
\end{aligned} \tag{15}
$$

Amostra	PC1	PC2	PC3	PC4
A1	-50.011	0.970	0.035	0.640
A2	-49.380	2.256	0.513	-0.208
A3	-47.416	-3.313	0.084	1.056
A4	-47.608	-0.375	-1.462	-0.336
A5	-51.374	0.506	0.971	0.463
B1	59.7988	2.639	-0.899	0.937
B2	59.104	0.711	-0.739	0.206
B3	57.051	-2.706	-0.813	-0.390
B4	60.111	-1.078	1.556	0.289
B5	59.431	0.455	0.909	-1.061
U1	-49.705	-0.065	-0.154	-1.597

Tabela 3: As amostras com suas posições na nova base que constitui os PC's, isto é os scores

A figura 7 apresenta a variância associada a cada PC. Vemos que a primeira componente principal já representa mais de 99% da variância total, indicando que somente esta já seria suficiente para explicar a variância do sistema, mas usualmente pelo menos duas PC's são usadas. Dessa forma podemos ver na figura 8 a separação dos dois materiais pelos seus *scores* na PC1. Além disso, vemos que a partir desse método é possível identificarmos a qual grupo de amostras pertence a amostra desconhecida U, neste caso ao conjunto A.

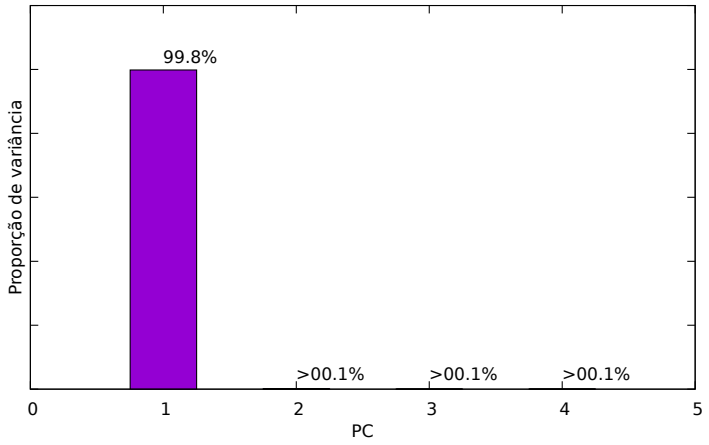

Figura 7: As proporções de variância, nela vemos que somente a PC1 se apresenta relevante

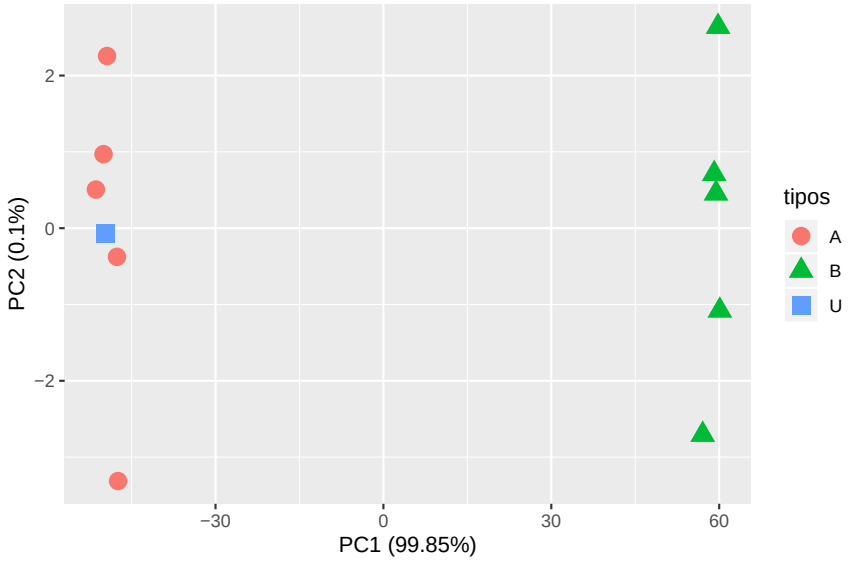

Figura 8: Pontuações das PCs obtidas no exemplo apresentado. No gráfico estão representados PC2×PC1.

Outro fator a se notar, é com relação aos *loadings*, que representam os pesos da combinação linear dos espectros originais, que no caso quimiométrico, representam as contribuições espectrais pela diferenciação. Na figura 9 podemos comparar o *loading* correspondente a primeira PC em comparação ao espectro médio das amostras do tipo A e B. Nela temos uma covariância positiva entre o loading e o espectro das amostras do tipo B($\sigma_{1-B} = 8.84$), observado ainda pela semelhança em $1200cm^{-1}$, enquanto ocorre uma covariância negativa com o espectro A ($\sigma_{1-A} = -18.03$), ou ainda, acompanha

o espectro invertido. Disso tiramos que conforme há um *score* mais positivo temos uma melhor covariância com amostras do tipo B, enquanto que *scores* negativos correspondem a um espectro mais relacionado às amostras do tipo A, como é observado na figura 8. Embora neste exemplo pôde ser demonstrado algumas análises com relação aos PC's na sessão 4.1 será melhor visto a importância dos loadings e na sessão 4.2 a questão dos agrupamentos com relação as *scores*.

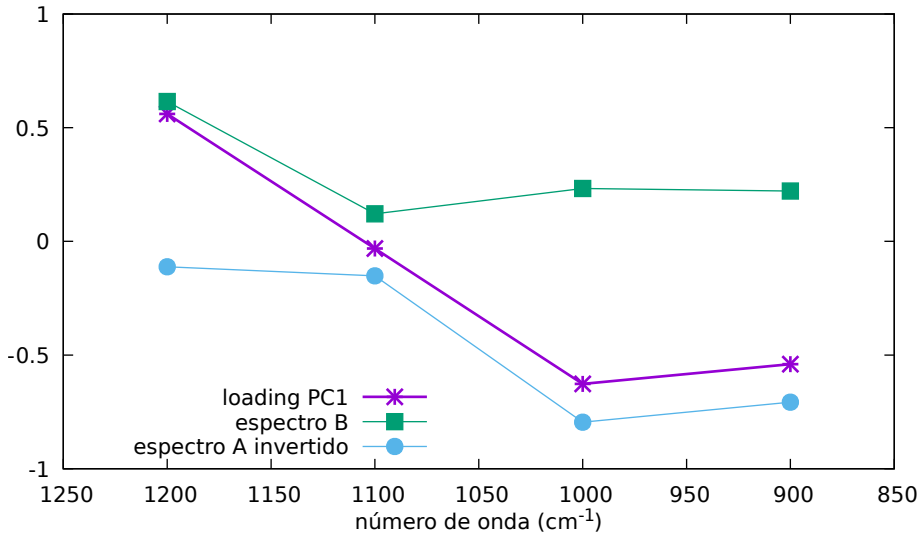

Figura 9: Comparação entre o loading e os espectros médios das amostras A e B

Dessa forma, o método do PCA, como citado anteriormente, é uma excelente ferramenta para análise multivariada, voltada para redução dimensional. Eliminando assim componentes que apresentam pouca representatividade da variância bem como uma breve ideia de como a aplicação desse método pode ser empregada. Neste caso em medidas espectroscopicas para diferenciação de materiais por agrupamento, ou *clusters*, assim como uma observação na razão pela qual eles se separam, notado pela comparação entre os espectros e os loadings.

3 Resultados e Discussões

Após revisitado os conceitos a respeito de espectroscopia FTIR e aplicação de PCA's, vamos agora para duas medidas reais, mostrando o procedimento de diferenciação por cluster. A primeira com o intuito de demonstrar a relevância dos *loadings* no processo de análise, e a segunda como um exemplo de diferenciação por agrupamentos aplicado à um material mais complexo.

Uma pequena quantidade das amostras foi posta sobre o cristal do ATR, após isso foi posta uma ventosa sobre a região da análise e fixada usando o mecanismo do aparelho. A câmara foi fechada, retirada a atmosfera e tomada a medida.

Todos os espectros foram coletados utilizando um FTIR modelo Bruker Vertex 70v em montagem de reflexão total atenuada no range de $4000cm^{-1}$ à $400cm^{-1}$, com 16 scans e uma resolução $4cm^{-1}$.

3.1 Analise quimiométrica qualitativa de misturas água e etanol

Nessa sessão vamos observar a relevância dos *loadings* no estudo do PCA. Como amostras usaremos soluções com diferentes concentrações de etanol e água, visto que são materiais de fácil obtenção e com espectros de FTIR bem conhecidos.

Na figura 10 podemos observar todos os espectros normalizados coletados, de forma que o mais inferior corresponde a água pura, o mais superior ao etanol P.A. Entre estes as soluções com proporções volumétricas de 10.0%, 11.1%, 12.5%, 14.2%, 16.6%, 20.0%, 25.5% , 33.3%, 50.0%.

Apenas observando os espectros é notável que, com o aumento da concentração de álcool a intensidade de seus picos característicos aumentam, particularmente observável nas regiões entre 800 a 1500 cm^{-1} e próximos de 3000 cm^{-1}.

Com o auxílio da linguagem de programção livre R [15], utilizando a estrutura do script do apêndice A, foi realizada a análise de componentes principais das misturas usando a matriz correlação, dando assim mais importância a região onde de fato, há sinal de absorção. Primeiramente vemos que a variância representada pela PC_1 corresponde a 99.2% da variância total, como ilustrado pela figura 11, ou seja, somente ela já seria suficiente para diferenciar as amostras.

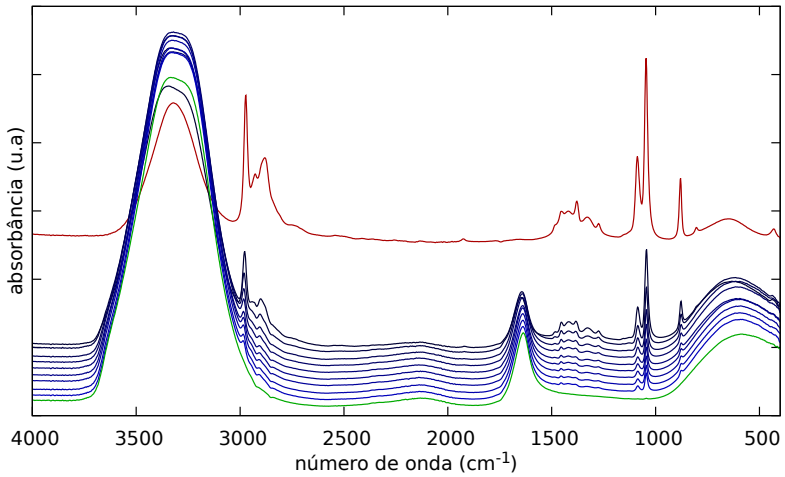

Figura 10: O espectro, ordenado de baixo para cima, de agua pura, as 9 soluções com quantidade de alcool crescente e o etanol puro

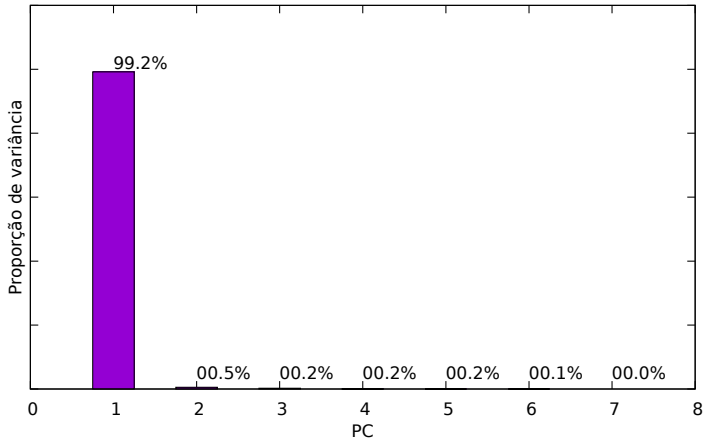

Figura 11: Proporção de variância das PC's, observa-se que somente a primeira, talvez segunda apresentam uma representatividade relevante

Olhando agora para os *scores*, figura 12, vemos que diferentemente do *usual*, não temos a formação de clusters, mas sim um tipo de caminho ou tendência, com relação aos valores para PC1, seguindo uma ordem em função do aumento da concentração de álcool.

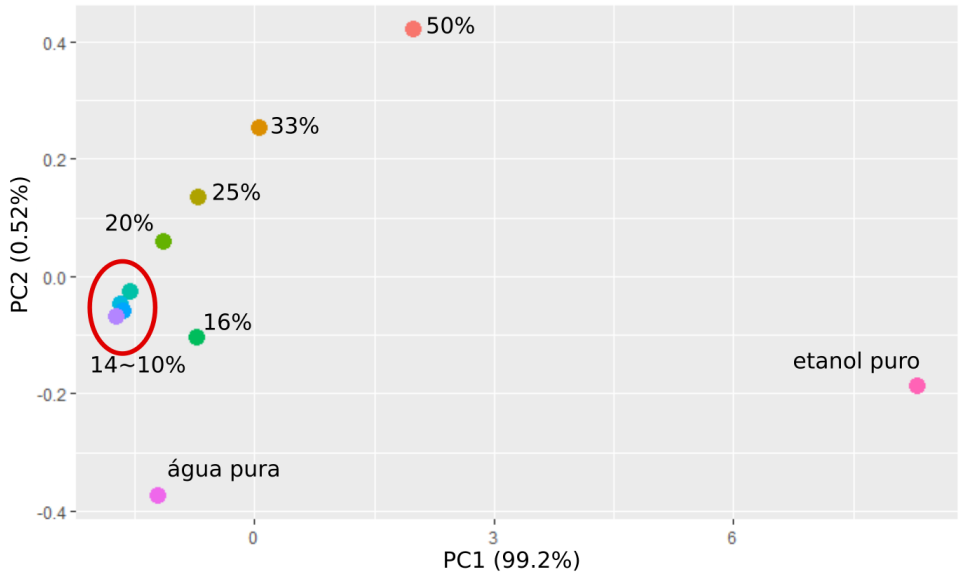

Figura 12: Os scores com relação às PC1 e PC2 para cada solução

Para melhor entendermos o comportamento dos *scores* da PCs, podemos observar os *loadings* com relação a PC_1 e PC_2 comparados com espectros da água e do etanol puros, figura 13. Os *loadings* do PC_1 apresentam formato muito similar ao espectro da água invertido, somado a uma pequena contribuição do espectro do etanol, indicando que na direção negativa do PC_1 temos uma maior relação com o espectro da água, enquanto que valores positivo do PC_1 a correlação é maior com o espectro do etanol, como já observando pelos *scores* na figura 12. Já os *loadings* da PC2, apesar de apresentarem algumas similaridades com do espectro do etanol, não segue um padrão de comportamento bem estabelecido comparando com os espectros das amostras puras, como ocorre com a PC1. Além disso como PC2 corresponde a menos de 1% da variância dos espectros, isso mostra que tal PC não tem grande relevância, ao menos sobre determinação da concentração alcoólica. Com relação à amostra com fração volumétrica de 16% nota-se que ela não segue nenhuma das linearidade das PC_1 ou PC_2, indicando que provavelmente houve algum problema na preparação da amostra, embora este não pode ser encontrado observando o espectro adquirido.

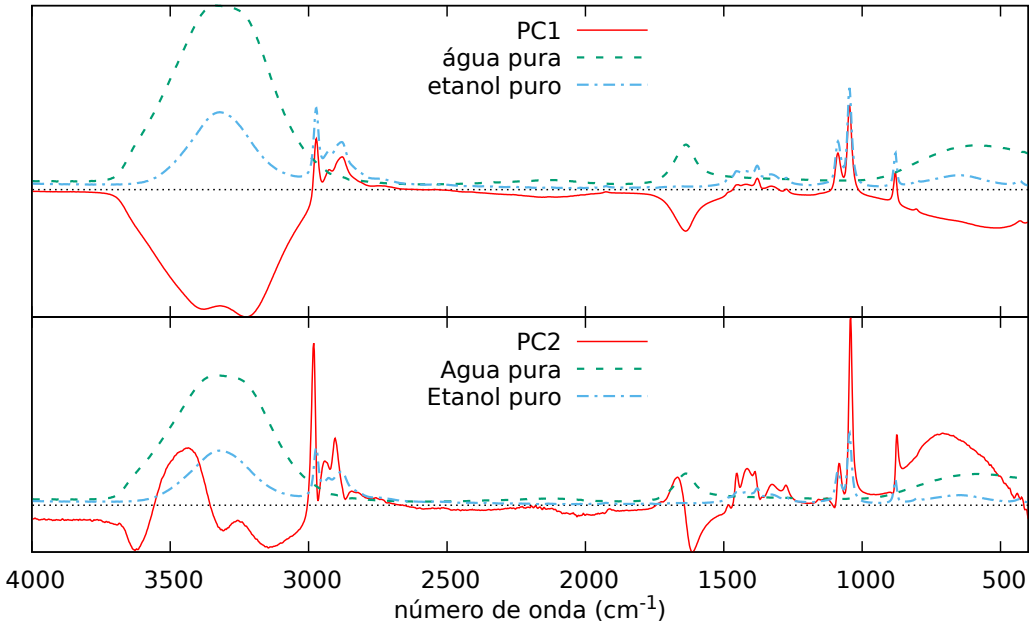

Figura 13: Os loadings referentes às PC's 1 e 2, e os espectros de agua e etanol puro

3.2 Análise quimiométrica qualitativa de sucos de frutas

O material analisado nesta sessão é o suco proveniente de 3 frutas cítricas diferentes, limões, laranjas e tangerinas de maneira que foram utilizadas três de cada espécime. Delas forma coletadas uma pequena quantidade de suco de cada por extração manual. Os espectros obtidos com os parâmetros citados no início da sessão constam na figura 14. como um panorama geral vemos que os espectros são semelhantes, com um grande sinal com relação à região de absorção da água. Entretanto, é possível notar uma diferença na região entre $1800cm^{-1}$ e $800cm^{-1}$, como destacado na figura 15, onde podemos observar que os espectros do limão apresentam um ombro à esquerda do pico em torno de $1600cm^{-1}$, além de variações na região próxima de 1200 cm^{-1}. Já os espectros da laranja e da tangerina apresentam picos na região de entre 1200 cm^{-1} e 1000 cm^{-1}.

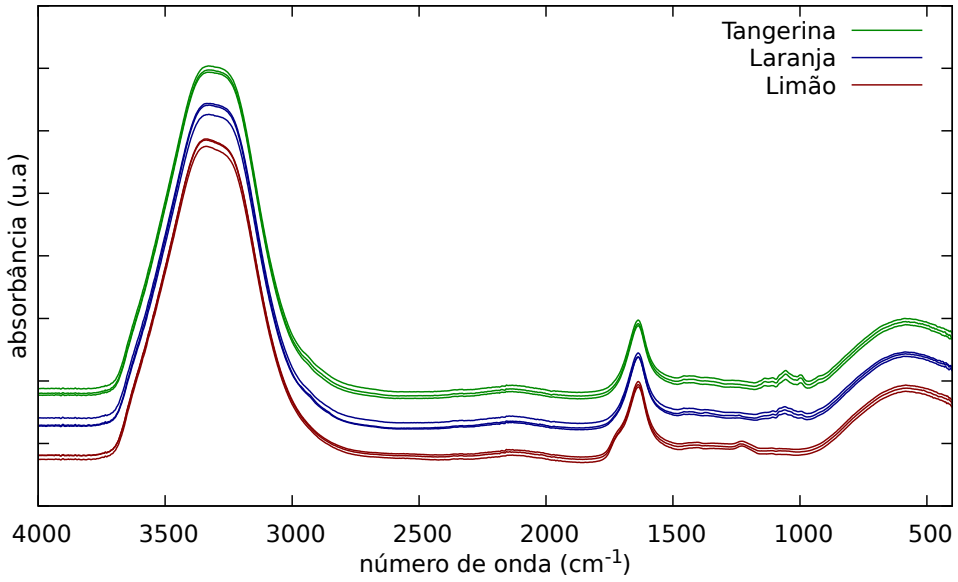

Figura 14: O espectro obtido referente a cada amostra

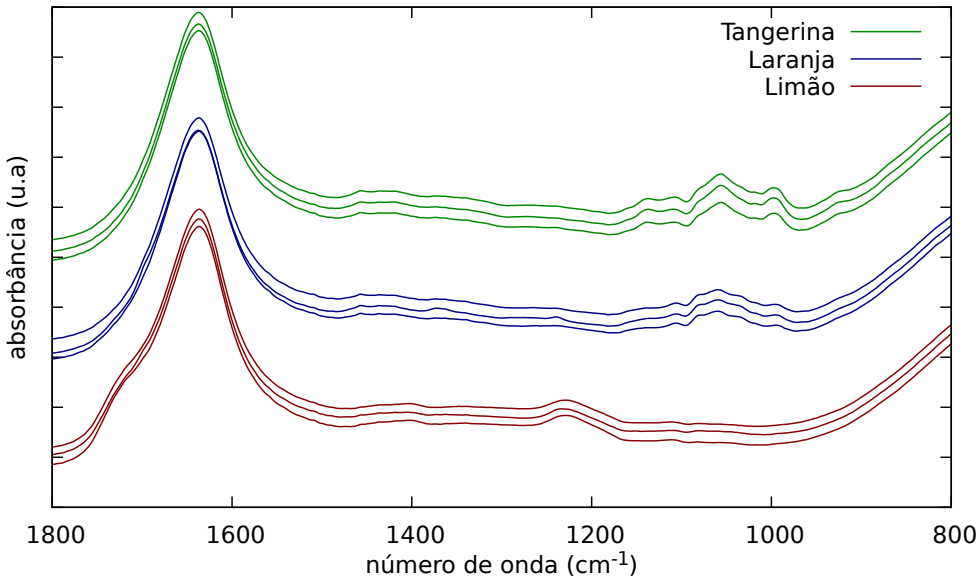

Figura 15: Detalhe da região entre 1800 cm^{-1} e 800 cm^{-1}

Embora o método do PCA seja útil para remover redundância de informações, nesse contexto as regiões em comum, ou desinteressantes, sem picos característicos, com isso, a remoção de tais redundâncias contribui para a análise. Neste caso a região de interesse se situa-se somente entre 1800 cm^{-1} e 800 cm^{-1}, como observado na figura 15. Restringindo então o PCA para esta região onde já temos uma pequena diferença nos ajuda a ter uma melhor interpretação. Com isso em vista, vemos que as duas primeiras componentes principais já representam 98% da variância total, como mostra a figura 16, ou seja, já são suficientes para caracterizar os três tipos de materiais, formando 3 clusters no plot dos *scores*, como visto na figura 17.

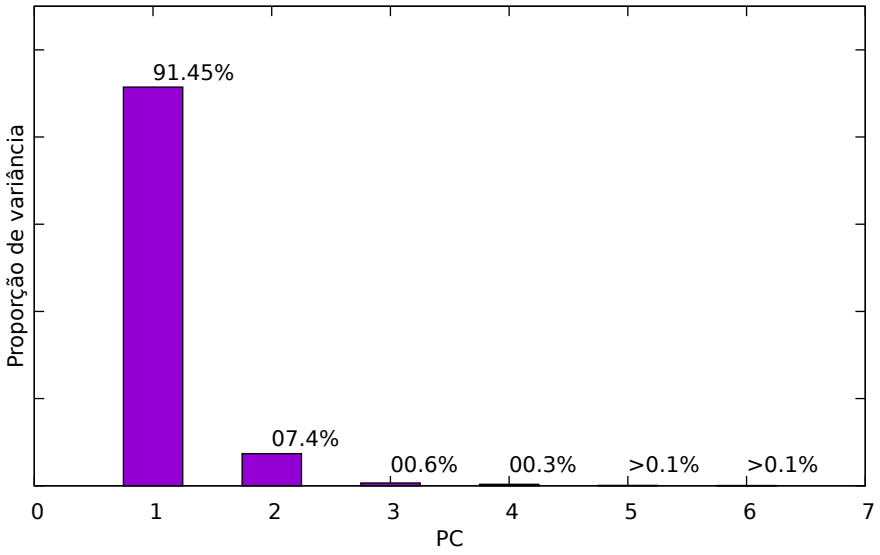

Figura 16: Plot da variância relativa a cada PC.

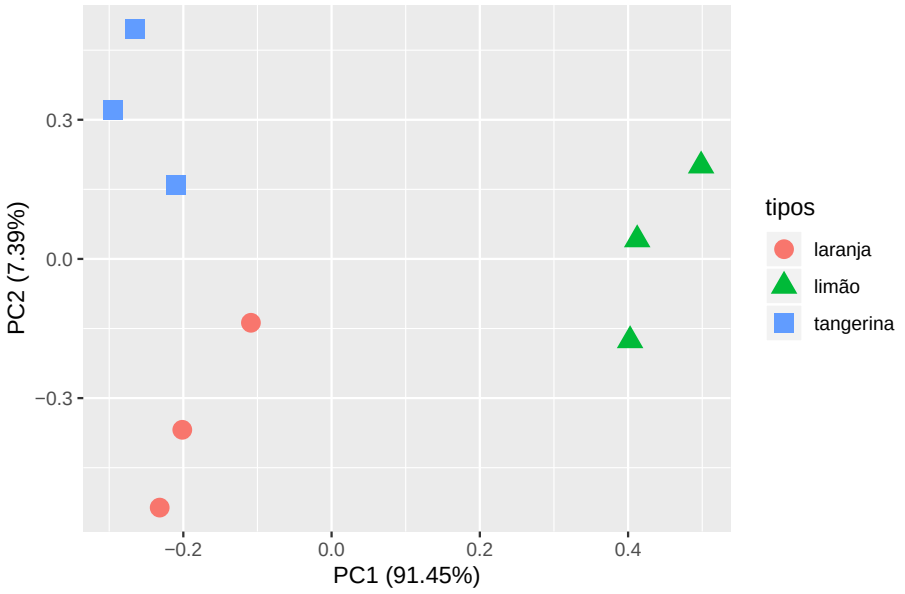

Figura 17: Plot dos scores obtidos dos espectros de absorção das três frutas, Laranja, Limão e tangerina.

Observa-se claramente a formação de três agrupamentos no plot dos *scores* $PC_2 \times PC_1$. Fazendo uma análise qualitativa, vemos que as frutas laranja e tangerina se assemelham mais entre si quanto à PC_1. Observando agora a figura 18, vemos que a parte positiva da PC1 esta ligada à semelhança ao

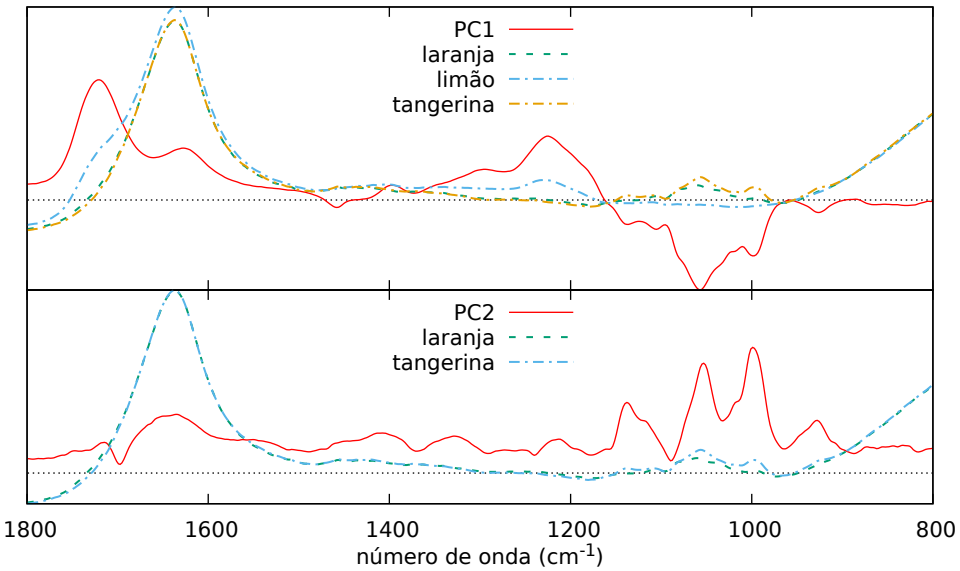

Figura 18: loadings correspondes as PC1 e P2

espectro do limão, em específico ao ombro à esquerda do primeiro pico, e a absorção em 1200 cm^{-1} enquanto que a parte negativa se relaciona com a região em torno de 1100cm^{-1} da laranja e da tangerina. na PC2 por sua vez, vemos que ela é responsável por diferenciar os espectros da laranja e da tangerina pelas intensidades nos sinais observados em torno de 1100cm^{-1}. Tal análise sem mostra coerente com os agrupamentos na figura 17. Entretanto para o estudo detalhado dos compostos que causam esta diferença, é necessário obter amostras puras e comparar com os *loadings*.

4 Conclusão

Tento em vista a teoria discutida e os métodos de análises apresentadas, vemos que o procedimento de análise de componentes principais PCA, em resumo, visa uma transformação dos dados originais para uma nova base constituída dos autovetores da matriz covariância, ou correlação, onde as novas posições no espaço das PC's são combinações lineares das variáveis originais. Isso se mostra extremamente vantajoso para inúmeros estudos de materiais, pois reduz o número de variáveis de um sistema sem perda de informação física ou química. Como já constatado na literatura, trata-se de um excelente método para a caraterização de materiais aliado à técnica de espectroscopia no infravermelho, como exemplificado no caso de uma mistura de duas substâncias simples, etanol e água, onde apenas a primeira componente principal mostrou-se suficiente para a diferenciação(99% da variância total), quanto ao teor alcoólico da mistura. Ainda neste caso vimos que os *loadings* fornecem informações de extrema relevância quando comparados com os espectros de cada componente puro, identificando a fontes de diferenciação nos *score*.

Vimos que, novamente, o método do PCA é útil para a diferenciação de materiais um pouco mais complexos e com alterações mais suaves no espectro de absorção, como no exemplo com 3 tipos de sucos de frutas cítricas. Neste caso as PC1 e PC2, representam 91.45% e 7.39% da variância total, respectivamente, resultando em uma representatividade de 98%, sendo o gráfico dos *scores* $PC_2 \times PC_1$ suficiente para a diferenciação dos três sucos. Fica claro que, com relação à PC1 as pontuações da laranja e tangerina, se agrupam, como esperado devido à sua semelhança espectral, enquanto que o limão se diferencia das outras duas.

Ou seja, a união entre o método da espectroscopia no infravermelho com a análise de componentes principais resultam num excelente casamento, podendo ser aplicado para a simples diferenciação de compostos, quanto para observar quais compostos causam essa diferença.

5 Agradecimentos

Agradeço meu orientador Marcelo Sandrini pelos conhecimentos transmitidos e auxílio tanto no laboratório quanto na revisão deste trabalho, também aos meus insubstituíveis amigos pela companhia e apoio durante o curso, em especial, Ari-n, Milena, Sussu, Rodrigo, Marcola, Vini e Bruna. Agradeço também meus pais pelo apoio e paciência, assim como meus professores e todos outros envolvidos por despertar minha curiosidade e paixão pela ciência. Agradeço ainda à Universidade Estadual de Maringá e às agências de fomento, Capes, CNPq, fundação aráucaria e a COMCAP pela estrutura disponibilizada e, por fim mas definitivamente não menos importante, ao Vitin pela ajuda na utilização do R e a Tina pela companhia nas noites de escrita e estudo.

Referências

[1] Donald L Pavia, Gary M Lampman, George S Kriz, and James A Vyvyan. *Introduction to spectroscopy*. Cengage Learning, 2008.

[2] E. L. Savi. *Avaliação do estagio de oxidação do biodisel e caracterização de blendas disel/biodisel via métodos espectroscópicos*. Tese de doutorado PFI - Universidade Estadual de Maringa, 2017.

[3] A. C. N. Mulati. *Avaliação físico química de complexos de inclusão de insulina e curcumina em cliclodextrinas: Estudo com as espectrocopias Raman, FTIR e fotoacústica*. Tese de doutorado PFI - Universidade Estadual de Maringa, 2015.

[4] G. M. Morais. *Espectroscopia vibracional para avaliação ex-vivo da dinâmica das ações induzidas por fungos patogênicos em tecidos biológicos*. Tese de doutorado PFI - Universidade Estadual de Maringa, 2016.

[5] John J Heigl, MF Bell, and John U White. Application of infrared spectroscopy to analysis of liquid hydrocarbons. *Analytical chemistry*, 19(5):293–298, 1947.

[6] VG Dourtoglou, Th Dourtoglou, A Antonopoulos, E Stefanou, S Lalas, and C Poulos. Detection of olive oil adulteration using principal component analysis applied on total and regio fa content. *Journal of the American Oil Chemists' Society*, 80(3):203–208, 2003.

[7] Lidija Svečnjak, Nikola Biliskov, Dragan Bubalo, and Domagoj Barišić. Application of infrared spectroscopy in honey analysis. 76:191–195, 10 2011.

[8] Romain Briandet, E Katherine Kemsley, and Reginald H Wilson. Discrimination of arabica and robusta in instant coffee by fourier transform infrared spectroscopy and chemometrics. *Journal of agricultural and food chemistry*, 44(1):170–174, 1996.

[9] John Francis James, Raymond N Enzweiler, Susan McKay, and Wolfgang Christian. A student's guide to fourier transforms with applications in physics and engineering. *Computers in Physics*, 10(1):47–47, 1996.

[10] A. L. Smith. *Applied Infrared Spectroscopy.* John Wiley and Sons: New York, 1979.

[11] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[12] Alvin C Rencher. *Methods of multivariate analysis*, volume 492. John Wiley & Sons, 2003.

[13] José Luiz Boldrini, Sueli IR Costa, VL Figueredo, and Henry G Wetzler. *Álgebra linear.* Harper & Row, 1980.

[14] Loredana F Leopold, Nicolae Leopold, Horst-A Diehl, and Carmen Socaciu. Quantification of carbohydrates in fruit juices using ftir spectroscopy and multivariate analysis. *Journal of Spectroscopy*, 26(2):93–104, 2011.

[15] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015.

6 Apêndice A

Aqui temos o código geral, comentado, utilizado para as análises. De maneira que, para cada uma foram feitos pequenos ajustes como mudança da região do PCA, nome do diretório e rótulos.

```
  # bibliotecas utilizadas
library(dplyr)
library(ggplot2)
library(ggfortify)
  # limpa a area de trabalho
rm(list = ls())
  # define wd como a o diretorio onde tem os dados organizados
wd  <- '/home/farinha96br/Dropbox/UEM/TCC/R caraio/analise_final/frutas/
    spectros/'
  # define o ambiente onde estao os dados
setwd(wd)
  # cria uma lista com os arquivos com a extensao .dpt
file_name <- list.files(pattern = '.dpt')
  # le o primeiro arquivo com os comprimentos de onda e as intensidades da
      primeira amostra
df_all <- read.table(file_name[1])
  # le a segunda coluna dos demais arquivos
  # cada linha e a intensidade de um comprimento de onda
  # # # # #
  # le a segunda linha dos demais arquivos e junta tudo num unico dataframe
for(i in 2:length(file_name)){
  df_aux <- read.table(file_name[i])
  df_all <- cbind(df_all, df_aux[,2])
}
  # elimina o arquivo auxiliar
rm(df_aux,i)
  # nomeia as colunas
colnames(df_all) = c('wavenumber',file_name)
  # cria o objeto df_spec com os comprimentos de onda estipulados, e ajeita
      o formato do df_all
df_spec = t(filter(df_all, wavenumber >= 800 & wavenumber  <= 1800))
df_all <- t(df_all)
  # coloca os df_spec e df_all nos formatos dataframe
df_spec = as.data.frame(df_spec)
df_all <- as.data.frame(df_all)
 # cria uma lista com o rotulo das amostras
label = c('comprimento de onda',rep("laranja", 3),rep("limao",3),rep("
    tangerina",3))
  # junta a lista de rotulos nos objetos
df_all = cbind(label, df_all)
df_spec = cbind(label, df_spec)
  # realiza o PCA em todos comprimentos de onda
PCA = prcomp(df_all[-1,-1])
  # realiza o PCA nos entre os comprimentos de onda especificos
PCA2 = prcomp(df_spec[-1,-1])
  # # # # # #
  # plota os scores usando todos os comprimentos de onda
autoplot(PCA, data = df_all[-1,], colour = "label")
  # plota os scores usando todos os comprimentos de onda especificos
autoplot(PCA2, data = df_spec[-1,], colour = "label")
  # cria os objetos com as rotations 1 e 2 de cada PCA
PCA1.PC1 = cbind(t(df_all["wavenumber",-1]),PCA$rotation[,1])
PCA1.PC2 = cbind(t(df_all["wavenumber",-1]),PCA$rotation[,2])
```

```
50 PCA2.PC1 = cbind(t(df_spec["wavenumber",-1]),PCA2$rotation[,1])
51 PCA2.PC2 = cbind(t(df_spec["wavenumber",-1]),PCA2$rotation[,2])
52   # exporta os rotations 1 e 2 do PCA e PCA2
53 write.table(PCA1.PC1,file = 'PCA1-PC1.dat', row.names = FALSE, col.names =
       FALSE)
54 write.table(PCA1.PC2,file = 'PCA1-PC2.dat', row.names = FALSE, col.names =
       FALSE)
55 write.table(PCA2.PC1,file = 'PCA2-PC1.dat', row.names = FALSE, col.names =
       FALSE)
56 write.table(PCA2.PC2,file = 'PCA2-PC2.dat', row.names = FALSE, col.names =
       FALSE)
```