

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

André Marcelo de Souza e Ronei Jesus Poppi*

Departamento de Química Analítica, Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13084-971 Campinas – SP, Brasil

Recebido em 15/2/11; aceito em 6/6/11; publicado na web em 22/7/11

TEACHING EXPERIMENT OF CHEMOMETRICS FOR EXPLORATORY ANALYSIS OF EDIBLE VEGETABLE OILS BY MID INFRARED SPECTROSCOPY AND PRINCIPAL COMPONENT ANALYSIS: A TUTORIAL, PART I. This manuscript aims to show the basic concepts and practical application of Principal Component Analysis (PCA) as a tutorial, using Matlab or Octave computing environment for beginners, undergraduate and graduate students. As a practical example it is shown the exploratory analysis of edible vegetable oils by mid infrared spectroscopy.

Keywords: principal component analysis; infrared spectroscopy; edible oils.

INTRODUÇÃO

A Quimiometria envolve a aplicação de métodos matemáticos, estatísticos e computacionais para investigar, interpretar, classificar e fazer previsão de conjuntos de dados de interesse químico, sendo atualmente considerada como uma disciplina da Química e inserida na grade curricular de diversos cursos de graduação e de pós-graduação em universidades brasileiras.¹

Dentre as diversas subáreas da Quimiometria pode-se destacar o planejamento de experimentos, o reconhecimento de padrões e a calibração multivariada. Na área de planejamento de experimentos busca-se encontrar quais as variáveis que mais afetam um determinado processo, assim como a interação entre elas. No reconhecimento de padrões, a partir de uma vasta gama de informações (medidas químicas ou espectrais, por exemplo) sobre uma série de objetos, pretende-se encontrar agrupamentos de amostras (objetos) que são similares entre si e, assim, detectar tendências nos dados. Na calibração multivariada, busca-se estabelecer um modelo que relacione uma série de medidas (químicas ou espectrais) realizadas em amostras com uma determinada propriedade (concentração, por exemplo). Na literatura podem ser encontrados diversos experimentos didáticos em calibração multivariada,²⁻⁴ reconhecimento de padrões^{5,6} e planejamento de experimentos.⁷

A análise de componentes principais (*Principal Component Analysis*, PCA) é um dos métodos mais importantes utilizados na Quimiometria e é a base para diversos métodos de reconhecimento de padrões, classificação e calibração multivariada. Normalmente, a PCA é utilizada com o objetivo de visualizar a estrutura dos dados, encontrar similaridades entre amostras, detectar amostras anômalas (*outliers*) e reduzir a dimensionalidade do conjunto de dados.⁸ A análise exploratória através da PCA é largamente empregada em Quimiometria⁹⁻¹¹ tendo cada vez mais novos usuários do meio acadêmico (alunos de graduação, pós-graduação e pesquisadores) e da indústria interessados na sua utilização. Contudo, esses novos usuários estão sujeitos a limitações iniciais no estudo da Quimiometria devido à falta

de conhecimento teórico prévio e habilidade de operação de softwares e ambientes computacionais indispensáveis a sua aplicação. Em geral, os artigos científicos e os livros da área apresentam a teoria e aplicação da PCA, entretanto, sua abordagem na utilização dos ambientes computacionais empregados é muito pouco didática. Por outro lado, a importância do entendimento dos procedimentos realizados pelos softwares é fundamental para avaliação dos resultados obtidos, bem como para o questionamento da maneira pelo qual tais softwares os realizam. Diversos softwares quimiométricos e ambientes computacionais estão disponíveis no mercado, dentre eles pode-se destacar o Matlab, Minitab, Pirouette, SIMCA-P+ e o Unscrambler, além de ambientes baseados em software livre como o Octave e o R.

A Quimiometria é dinâmica e constantemente novos algoritmos surgem ou são modificados, por isso, ambientes computacionais são largamente empregados por pesquisadores especializados devido à possibilidade de programação de rotinas dentro destes ambientes, como no caso do Matlab ou Octave. Ambos são ambientes computacionais interativos de alto desempenho voltado para o cálculo numérico, cujo elemento básico de informação é uma matriz, sendo largamente empregado no desenvolvimento de métodos quimiométricos e será objetivo de estudo deste tutorial.

O uso de espectroscopia no infravermelho médio com transformada de Fourier (FT-IR) combinado com análise quimiométrica é constantemente explorado em estudos de reconhecimento de padrões de óleos vegetais comestíveis,¹² na análise de parâmetros físico-químicos de qualidade¹³ e no estudo de autenticidade¹⁴ e adulteração desses óleos.^{15,16} Diversos cursos de pós-graduação e graduação em Química espalhados pelo Brasil, e até mesmo alguns cursos técnicos, possuem em seus currículos a utilização da técnica de espectroscopia no infravermelho médio com transformada de Fourier (*Fourier Transform Infrared*, FTIR). A FTIR é uma técnica rápida, requer o mínimo necessário de preparo de amostras e sua instrumentação é facilmente encontrada nos laboratórios. Esta técnica permite a análise qualitativa de compostos orgânicos porque os modos característicos de vibração de cada grupo provocam o aparecimento de bandas no espectro infravermelho em frequências específicas, que também são influenciadas pela presença de grupos funcionais próximos (acopla-

*e-mail: ronei@iqm.unicamp.br

mentos).¹⁷ Sendo assim, um espectro de infravermelho geralmente contém mais informação do que apenas os valores de posição ou de absorção de alguns picos, atuando como uma impressão digital de uma dada amostra quando utilizado integralmente. Além disso, a espectroscopia FTIR é uma excelente ferramenta para análise quantitativa porque as intensidades de absorção das bandas no espectro são proporcionais à concentração.¹⁸ Em trabalho anterior, Rusak *et. al.*⁶ apresentaram uma proposta de experimento didático para Quimiometria utilizando óleos vegetais comestíveis, PCA e FTIR, entretanto, não foram explorados conceitos sobre os cálculos envolvidos na aplicação da PCA e utilização de softwares quimiométricos. Assim, este trabalho pretende ser um tutorial, com o objetivo de enfatizar conceitos básicos de Quimiometria, tratamento prático do conjunto de dados no ambiente computacional Matlab, explorando os comandos básicos para a aplicação e entendimento da PCA por iniciantes, devido à escassez deste tipo de abordagem na literatura. O experimento descrito neste trabalho pode ser facilmente realizado em cursos de graduação, visto que os equipamentos, materiais e softwares necessários são disponíveis em grande número de instituições no país.

Conforme mencionado anteriormente, métodos de análise exploratória são utilizados para extrair informação e detectar tendências nos dados, baseados nas medidas multivariadas das amostras. De maneira geral, eles podem ser classificados como métodos supervisionados como, por exemplo, *Linear Discriminant Analysis* (LDA),¹⁹ *K-Nearest Neighbor* (KNN),¹⁹ *Partial Least Square Discriminant Analysis* (PLS-DA)¹⁹ e *Soft-Independent Modeling of Class Analogy* (SIMCA),¹⁹ ou não supervisionados como, por exemplo, *Principal Component Analysis* (PCA) e *Hierarchical Cluster Analysis* (HCA)¹⁹ e permitem a interpretação multivariada de conjuntos de dados complexos por meio de gráficos bi- ou tridimensionais. Nos métodos supervisionados é necessário que exista alguma informação inicial sobre a identidade das amostras para a formação das classes e o objetivo é desenvolver um modelo baseado nas informações contidas nas amostras. Por outro lado, nos métodos não supervisionados, a separação de classes acontece sem a necessidade de informações iniciais sobre a natureza das amostras e o objetivo é identificar agrupamentos naturais entre as amostras.¹⁹

A PCA é um método que permite a redução da dimensionalidade através da representação do conjunto de dados em um novo sistema de eixos, denominados componentes principais (PC), permitindo a visualização da natureza multivariada dos dados em poucas dimensões. No espaço original, as amostras são pontos localizados em um espaço n-dimensional, sendo n igual ao número de variáveis. Com a redução de dimensionalidade proporcionada pela PCA, as amostras passam a ser pontos localizados em espaços de dimensões reduzidas definidos pelas PCs, por exemplo, bi- ou tridimensionais. Matematicamente, na PCA, a matriz \mathbf{X} é decomposta em um produto de duas matrizes, denominadas escores (\mathbf{T}) e pesos (\mathbf{P}), mais uma matriz de erros (\mathbf{E})²⁰, como mostrado na Equação 1:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (1)$$

Os escores representam as coordenadas das amostras no sistema de eixos formados pelos componentes principais. Cada componente principal é constituído pela combinação linear das variáveis originais e os coeficientes da combinação são denominados pesos. Matematicamente, os pesos são os cossenos dos ângulos entre as variáveis originais e os componentes principais (PC), representando, portanto, o quanto cada variável original contribui para uma determinada PC. A primeira componente principal (PC1) é traçada no sentido da maior variação no conjunto de dados; a segunda (PC2) é traçada ortogonalmente à primeira, com o intuito de descrever a maior porcentagem da variação não explicada pela PC1 e assim por diante; enquanto os

escores representam as relações de similaridade entre as amostras. A avaliação dos pesos permite entender quais variáveis mais contribuem para os agrupamentos observados no gráfico dos escores. Através da análise conjunta do gráfico de escores e pesos, é possível verificar quais variáveis são responsáveis pelas diferenças observadas entre as amostras. O número de componentes principais a ser utilizado no modelo PCA é determinado pela porcentagem de variância explicada. Assim, seleciona-se um número de componentes de tal maneira que a maior porcentagem da variação presente no conjunto de dados originais seja capturada.²⁰ Diversos algoritmos estão disponíveis para a realização da PCA e quatro deles aparecem frequentemente na literatura: *Nonlinear Iterative Partial Least Squares* (NIPALS),²¹ *Singular Value Decomposition* (SVD),¹⁹ os quais utilizam a matriz de dados \mathbf{X} , *POWER* e *Eigenvalue Decomposition* (EVD)¹⁹ que trabalham com a matriz de produto cruzado $\mathbf{X}'\mathbf{X}$. SVD e EVD extraem os componentes principais simultaneamente, enquanto NIPALS e POWER calculam as PC sequencialmente.

O SVD (Equação 2) é baseado no teorema da álgebra linear que afirma que uma matriz \mathbf{X} (mxn), m colunas e n linhas, pode ser transformada em um produto de três matrizes \mathbf{U} , \mathbf{S} , \mathbf{V}^T (\mathbf{T} subscrito significa transposta) e que têm propriedades específicas: as matrizes \mathbf{U} e \mathbf{V} são quadradas e ortonormais, a matriz \mathbf{S} é uma matriz retangular diagonal contendo os valores singulares na diagonal e todos os elementos fora da diagonal iguais a zero.²¹ Nesse caso, os pesos são dados pela matriz \mathbf{V} e os escores por: $\mathbf{T}=\mathbf{US}$.

$$\mathbf{X} = \mathbf{USV}^T \quad (2)$$

EVD é um método similar ao SVD utilizado para calcular autovalores e autovetores de uma matriz. Ambos, SVD e EVD, são métodos relativamente fáceis de utilizar no Matlab, porque neste software há funções internas para a decomposição de valores singulares e para estimar autovetores e autovalores de uma matriz.

O NIPALS é um método comumente utilizado para o cálculo das componentes principais de um conjunto de dados, onde os vetores dos pesos e dos escores são calculados iterativamente, um de cada vez, em ordem decrescente de importância. O processo iterativo, para o primeiro componente principal, é inicializado com uma primeira estimativa de escores, que pode ser a coluna de \mathbf{X} que tem maior variância. Utilizando estes escores, calculam-se os pesos como: $\mathbf{p}=\mathbf{tX}/\mathbf{t}'\mathbf{t}$ e os pesos são normalizados para comprimento igual a 1. Após isso, os escores são calculados como: $\mathbf{t}=\mathbf{Xp}/\mathbf{p}'\mathbf{p}$. Esses valores de escores são comparados com os anteriores e se forem diferentes (dentro de um critério pré-estabelecido), os pesos são novamente calculados como mostrado acima. Esse processo continua até que os escores sejam semelhantes ou um certo número de iterações tenham sido realizadas. Após a convergência, o produto \mathbf{tp}^T é subtraído de \mathbf{X} , obtendo-se o resíduo \mathbf{E} (Equação 3):

$$\mathbf{E} = \mathbf{X} - \mathbf{tp}^T \quad (3)$$

O processo continua para o próximo componente principal, substituindo \mathbf{E} por \mathbf{X} .

Pré-processamento

A etapa de pré-processamento dos dados é fundamental para o sucesso da análise multivariada. Os principais objetivos da aplicação das técnicas de pré-processamento são eliminar informações não relevantes do ponto de vista químico e tornar a matriz de dados melhor condicionada para a análise, possibilitando a subsequente análise exploratória do conjunto de dados com eficiência. Existe uma vasta literatura disponível a respeito dos diversos métodos de

processamento de dados em espectroscopia. Normalizar os espectros, centrar os dados na média, derivar e suavizar utilizando o algoritmo de Savitzky-Golay e aplicar a correção de espalhamento multiplicativo (MSC, *Multiplicative Scatter Correction*) são alguns dos métodos mais aplicados.²²

Normalização é um tipo de pré-processamento que tem como objetivo reduzir a influência de variações indesejadas presentes no conjunto de dados, garantindo que cada observação seja representada de forma adequada e consistente. Neste artigo, foi utilizado um método comum de normalização aplicado a dados espectroscópicos, chamado normalização por área total. O objetivo da normalização de área total é principalmente reduzir o efeito da intensidade total de perfis de resposta, devido a variações na concentração da amostra e do caminho ótico. A normalização para a área total do espectro é realizada pela divisão de cada variável pela soma dos valores absolutos de todas as variáveis para uma dada amostra, conforme a Equação 4.

$$X_{i,norm} = \frac{x_i}{\sum_{j=1}^n |x_{i,j}|} \quad (4)$$

Centrar os dados na média (Equação 5) consiste em calcular a média das intensidades para cada comprimento de onda e subtrair cada uma das intensidades do valor médio. Desta maneira, cada variável passará a ter média zero, ou seja, as coordenadas são movidas para o centro dos dados, permitindo que diferenças nas intensidades relativas das variáveis sejam mais fáceis de perceber.

$$x_{ij}(\text{centrado na média}) = x_{ij}(\text{original}) - \bar{x}_{ij}(\text{média}) \quad (5)$$

Deslocamento e inclinação de linha de base podem ser corrigidos por derivação dos espectros. Os métodos de alisamento são utilizados para reduzir matematicamente o ruído, aumentando com isto a relação sinal/ruído. Nestes métodos, é selecionada uma janela, a qual contém certo número de variáveis. Os pontos na janela são, então, utilizados para determinar o valor no ponto central da janela e, assim, o tamanho da janela influencia diretamente o resultado do alisamento. No método de Savitzky-Golay, um polinômio de ordem baixa é ajustado aos pontos da janela e utilizado para recalculer o ponto central.²²

A correção de espalhamento multiplicativo é um método de transformação utilizado para compensar os efeitos aditivos e/ou multiplicativos em dados espectrais. Este método remove a influência de efeitos físicos nos espectros, tais como o tamanho de partícula, a rugosidade e opacidade, os quais não trazem informações químicas sobre as amostras e introduz variações espectrais, como o deslocamento da linha de base. Para fazer a correção, o método MSC assume que cada espectro é determinado pelas características químicas da amostra somadas às características físicas indesejadas. A Equação 6 descreve o princípio de funcionamento do MSC:

$$x_{ik}(\text{transformado}) = \frac{x_{ik}(\text{original}) - a_i}{b_i} \quad (6)$$

onde: $x_{ik}(\text{original})$ e $x_{ik}(\text{transformado})$ são os valores de absorbância antes e depois de correção com o MSC em k comprimentos de onda; a_i e b_i são constantes estimadas a partir de uma regressão em mínimos quadrados de um espectro individual x_{ik} contra um espectro médio do conjunto de calibração em todos os comprimentos de onda ou em um subconjunto, seguindo a Equação 7:

$$x_{ik} = a_i + b_i \bar{x} + e_{ik} \quad (7)$$

Onde e_{ik} corresponde a todos os outros efeitos nos espectros que não foram modelados.^{22,23}

PARTE EXPERIMENTAL

Obtenção dos espectros

As análises de infravermelho foram realizadas em um espectrômetro ABB Bomem MB series, operando em modo de absorção, equipado com detector de sulfato de triglicina deuterada (DTGS), sendo os espectros obtidos na faixa de 600 a 4000 cm^{-1} com resolução de 4 cm^{-1} e 16 varreduras. Foi utilizada uma cela desmontável para amostras líquidas com janela de KBr. Antes da medida dos espectros da amostras, foi medido o espectro de infravermelho do ar (ausência da janela de KBr) e tomado como medida do branco. Os espectros foram obtidos da seguinte forma: uma gota de óleo foi colocada sobre a superfície da janela de KBr, o óleo residual remanescente foi removido com o auxílio de algodão enrolado em uma pinça. Então, a cela foi montada e posicionada no aparelho para a medida do espectro. Para obtenção de novos espectros a cela foi limpa com diclorometano.

Metodologia

Foram analisados três tipos de óleos vegetais comestíveis comerciais adquiridos no comércio da região (azeite, canola e soja). Para cada tipo de óleo foram realizadas 6 medidas: azeite (a1;a2;a3;a4;a5;a6), canola (c1;c2;c3;a4;a5;a6) e soja (s1;s2;s3;s4;s5;s6). Além disso, foram utilizadas para previsão três amostras desconhecidas, am1, am2 e am3 sendo que, neste caso, 3 medidas foram realizadas para cada amostra, somando 9 medidas (am11;am12;am13,am21,am22, am23,am31,am32,am33), ou seja, 27 espectros foram adquiridos no total. A ordem das amostras na matriz de dados \mathbf{X} foi azeite, canola, soja e amostras desconhecidas (am1,am2 e am3).

A PCA foi executada em Matlab 7.0 e foi utilizado um *toolbox* gratuito,²³ desenvolvido por Wen Wu and Sijmen de Jong, do FABI – *Vrije Universiteit Brussel*. Esta rotina permite o cálculo do PCA de uma matriz de dados, onde cada amostra é colocada numa linha e fornece os scores e pesos, assim como a porcentagem de variância descrita em cada componente principal e os scores de uma nova matriz teste.

RESULTADOS E DISCUSSÃO

O Matlab possui versões diferentes para diferentes ambientes operacionais e este tutorial está baseado em versões mais recentes para o ambiente Windows XP/Vista/7. As operações descritas a seguir também podem ser realizadas em Octave. O Octave é um software livre e a sua utilização é bastante similar à do Matlab. O Octave pode ser encontrado no site http://download.famouswhy.com/octave/free_download.html. Neste site está disponível a versão 3.2.4. Após a instalação do Octave, pode-se instalar o GUIOctave 1.0.14 (disponível em <http://www.soft82.com/get/download/windows/gui-octave>). Este programa funciona como uma “máscara” para o Octave e permite a utilização de uma interface mais amigável. A sequência típica de passos a ser executada com o objetivo de executar a PCA, a partir dos espectros de FT-IR é:

- (1) Carregar os espectros, um a um, para o ambiente Matlab digitando o comando *load*. Exemplo: Se o nome do arquivo é a1.dat. Então, `>>load a1.dat;`
Qualquer dúvida na execução de algum comando do Matlab pode ser sanada pela utilização do comando *help* como, por exemplo, *help load*. A função *doc* mostra os documentos disponíveis em HTML relativos ao comando *help*. Por exemplo: `>>doc function` Este comando abre a página disponível em HTML relativa ao comando *function*.
- (2) Criar uma matriz contendo todas as amostras.

A matriz **X** é organizada da seguinte maneira: nas linhas são colocadas as amostras e nas colunas as variáveis. No caso dos espectros, nas linhas temos as amostras de óleos e nas colunas os valores de absorvância. Cada espectro exportado do equipamento de infravermelho contém duas colunas: uma com o número de ondas e outra com os dados de absorvância. Por isso, ao montar a matriz é necessário selecionar apenas a coluna que contém os valores de absorvância, sendo, neste caso, a segunda coluna. Os valores entre parênteses representam linha e coluna, respectivamente: (linha, coluna). O sinal de dois pontos (:) representa todos os elementos, neste caso todos os elementos da linha, e o sinal de apóstrofo (') representa transposição. Neste trabalho, foi necessário transpor os espectros porque estes são exportados em colunas, entretanto, na matriz de dados, estes devem estar presentes em linhas.

A matriz **X** contendo as dezoito amostras de óleo pode ser criada da seguinte forma:

```
>>X=[a1(:,2)';a2(:,2)';a3(:,2)';a4(:,2)';a5(:,2)';a6(:,2)';c1(:,2)';
c2(:,2)';c3(:,2)';c4(:,2)';c5(:,2)';c6(:,2)';s1(:,2)';s2(:,2)';s3(:,2)';
s4(:,2)';s5(:,2)';s6(:,2)'];
```

(3) Criar um vetor correspondente aos valores de número de ondas utilizando a primeira coluna: `>>num=a1(:,1);`

(4) Construir e formatar o gráfico contendo todos os espectros presentes na matriz **X** da seguinte forma:

```
>>plot(num,X);
>>grid
```

(5) Realizar a seleção da faixa espectral de 400 a 4.000 cm^{-1} (note que se o espectro obtido já está nesta faixa, não é necessária esta etapa):

```
>>X1=X(:,600:3400);
>>num1=num(600:3400);
>>plot(num1,X1);
>>grid
```

Nesse caso, a utilização de toda essa faixa espectral pode não ser a mais eficiente para os propósitos do trabalho, uma vez que no infravermelho se tem estimativas fundamentais de quais grupos funcionais podem estar presentes na amostra.

(6) Selecionar a faixa de impressão digital de 1.000 a 1.500 cm^{-1} (Figura 1):

```
>>X2=X1(:,400:1000);
>>num2=num1(400:1000);
>>plot(num2,X2);
>>grid
>>ylabel('Absorbância'); xlabel('Número de onda (1/cm)');
```

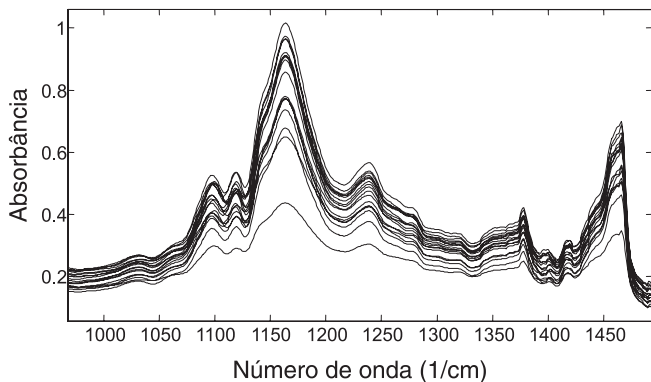


Figura 1. Faixa espectral de impressão digital

Essa região espectral que contém a faixa de impressão digital apresenta alterações significativas na distribuição dos máximos de absorção. Assim, ela é a mais apropriada para realizar a identificação do tipo de óleo baseado em comparações espectrais.

Uma maneira alternativa para visualizar um conjunto de gráficos é organizando todos em sequência através dos *subplots* como, por exemplo:

```
>>subplot(3,1,1),plot(num,x);subplot(3,1,2),plot(num1,x1);
>>subplot(3,1,3),plot(num2,x2);
>>ylabel('Absorbância'); xlabel('Número de onda (1/cm)');
```

Durante a etapa de obtenção dos espectros, que foi realizada ao longo do semestre, observou-se que as pastilhas de KBr foram escurecendo com o tempo. Provavelmente, este fenômeno esteve relacionado com a exposição da pastilha à umidade do ar. Também foi observado que a espessura do filme formado pelo contato da gota de óleo com as duas pastilhas de KBr não era uniforme durante as medidas. Os principais motivos para esta variação foram o fato de que cada aluno preparava pelo menos uma amostra, o que, inevitavelmente, levou a diferenças na espessura do filme; outro motivo para a variação foi que, em muitos casos, a formação de pequenas gotículas de ar presentes no filme era inevitável. Desta forma, os espectros apresentaram diferentes intensidades e deslocamentos de linha de base. Sendo assim, as técnicas de pré-processamento através derivação (primeira e segunda derivadas) e do MSC foram consideradas alternativas viáveis para minimizar os efeitos indesejados nos espectros. A seguir, são apresentados os passos para centrar os dados na média (a partir da matriz **X2**) e para a execução da primeira e segunda derivadas (a partir da matriz **X2** centrada na média) utilizando o algoritmo de Savitzky-Golay e do MSC (a partir da matriz **X2**).

(7) Centrar na média:

```
>>x2m=mean(x2);
>>for i=1:18
>>x2mc(i,:)=x2(i,:)-x2m;
>>end
```

(8) Primeira derivada (Savitzky-Golay)

Como sugestão pode ser utilizada a função "deriv".²³ Para entender o funcionamento desta função e de qualquer outra no Matlab ou no Octave pode ser executada a função *help* como, por exemplo:

```
>>help deriv
```

Desta forma todas as instruções para a aplicação do algoritmo aparecerão na janela do "Command Window".

```
>>[dx1x2nmc] = deriv(x2mc,2,15,2);
>>figure
>>plot(num2,dx1x2nmc)
>>title('Centrado na média + 1a derivada');
>>xlabel('Número de onda (1/cm)')
>>ylabel('1a derivada')
```

(9) Segunda derivada

```
>>[dx2x2nmc] = deriv(x2mc,2,15,2);
>>figure
>>plot(num2,dx2x2nmc) % Figura 4S
>>title('2a derivada')
>>xlabel('Número de onda (1/cm)')
>>ylabel('2a derivada')
```

(10) Correção de espalhamento multiplicativo (MSC)²³

```
>>[Xmsc]=msc(x2mc);
>>plot(Xmsc')
>>grid; ylabel('Absorbância'); xlabel('Número de onda (1/cm)');
title('MSC');
```

A seguir será apresentada uma sequência de passos para a execução da PCA a partir da matriz de **X2** centrada na média e derivada através da primeira derivada, respectivamente.

(11) PCA utilizando o algoritmo KPCA²³

```
>>help kPCA
>>[t,p,percent] = kPCA(dx1x2nmc);
```

(12) Gráfico dos escores (Figura 2)

```

>>plot(t(1:6,1),t(1:6,2),'xg',t(7:12,1),t(7:12,2),'+r',t(13:18,1),t(13:18,2),'*b')
>>grid
>>for i=1:18;
>>text(t(i,1),t(i,2),num2str(i))
>>end
>>xlabel('PC1 (79.8789%)');
>>ylabel('PC2(17.7318%)');
>>title('Centrado na média + 1a derivada + EVD')
>>legend('Azeite','Canola','Soja');

```

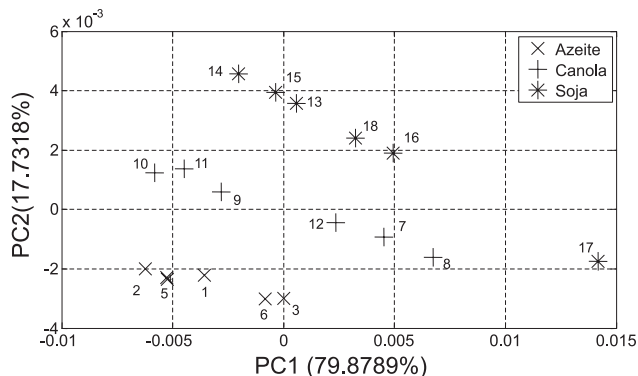


Figura 2. Gráfico de escores da PC1 x PC2 referente às 18 amostras de óleos vegetais comerciais

(13) Gráfico dos pesos (Figura 3)

```

>>plot(num2,p(:,2))
>>xlabel('1/cm');
>>ylabel('PC2(17,7318%)');
>>title('Centrado na média + 1a derivada + EVD')

```

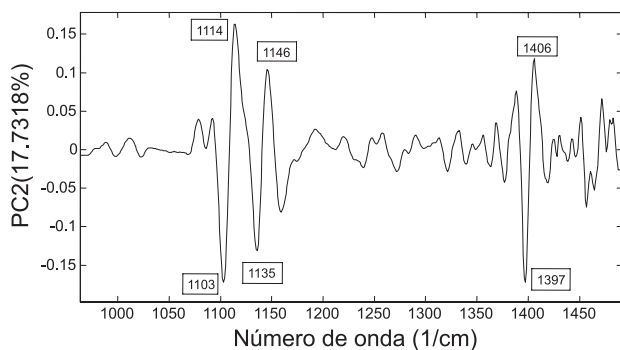


Figura 3. Gráfico dos pesos da PC2

A porcentagem de variância explicada pelas cinco primeiras PC é, respectivamente, 79,88; 17,73; 1,18; 0,60 e 0,40%. O gráfico de escores da PC1 x PC2 (Figura 2) evidenciou a formação de agrupamentos distintos entre as três classes de óleos. Neste gráfico, a PC no. 2 foi responsável pela distinção entre os diferentes tipos de óleos. Na porção superior do gráfico de escores estão as amostras de óleo de soja, na porção inferior estão as de azeite de oliva e, em uma posição intermediária, estão as de óleo de canola. O gráfico dos pesos (Figura 3) destacou que as principais variáveis responsáveis pela formação do agrupamento observado entre as amostras no gráfico de escores são 1103, 1136 e 1397 cm^{-1} , que estão localizadas na porção negativa do gráfico, e as variáveis 1114, 1146 e 1406 cm^{-1} , na porção positiva do gráfico de pesos. Do ponto de vista químico, estas diferenças estão relacionadas com a porcentagem de ácidos graxos saturados e insaturados presentes nestes óleos. O óleo de soja possui em sua composição 50 a 57% de ácido linoleico (C18:2, ou seja, 18 átomos

de carbono e 2 insaturações), 18 a 26% de ácido oleico (C18:1) e 5,5 a 10% de ácido linoleico (C18:3). O azeite de oliva pode possuir em sua composição 64 a 83% de ácido oleico (C18:1), 3,5 a 16% de ácido linoleico (C18:2) e 0,2 a 1,8% de ácido linoleico (C18:3).²¹ A variável correspondente a 1397 cm^{-1} é uma banda de infravermelho típica de deformação angular no plano de ligação CH de grupos cis-olefínicos.²⁴ Portanto, esta banda deve ser responsável por separar o óleo de soja (mais insaturado) dos demais.

As demais bandas são características de óleos vegetais, sendo aquelas ao redor de 1110 cm^{-1} relativas à ligação C-O e em torno de 1400 cm^{-1} a C-H.²⁴

O gráfico dos escores da PC1 x PC3 (Figura 4) mostra a formação de dois agrupamentos distintos, sendo um formado pelas amostras de óleo de canola (parte negativa da PC3) e outro formado pelas amostras de óleo de soja e azeite na parte positiva da PC3. Desta forma, a PC3 é responsável por separar o óleo de canola dos demais. Avaliando o gráfico dos pesos na PC3 (Figura 5) é possível concluir que as principais variáveis responsáveis por esta distinção são 1139, 1463 e 1469 cm^{-1} .

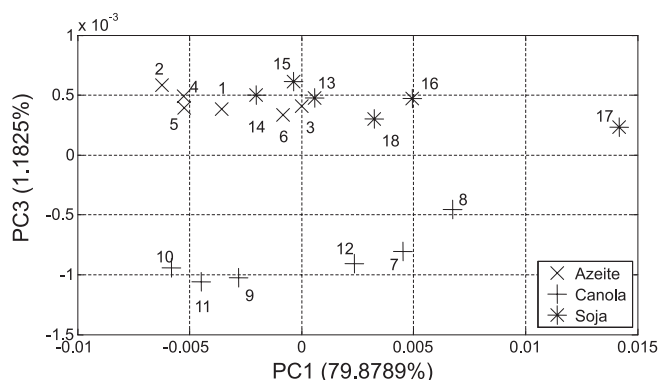


Figura 4. Gráfico de escores da PC1 x PC3

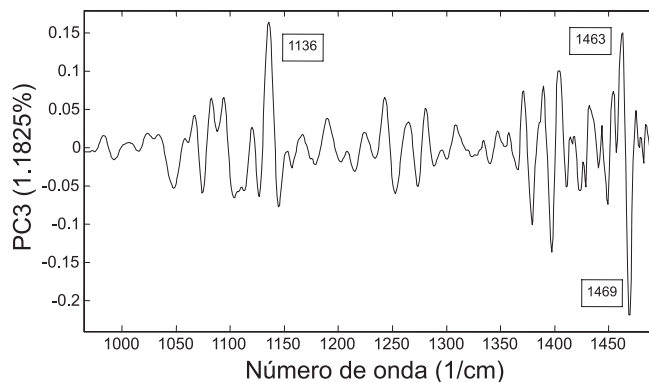


Figura 5. Gráfico dos pesos da PC3

A seguir serão apresentados alguns comandos básicos para a inclusão das 9 amostras desconhecidas (am11, am12, am13, am31, am32, am33, am41, am42, am43) no modelo de PCA. Será necessário criar uma nova matriz de dados (\mathbf{Xc}), derivar e alisar os espectros utilizando a primeira derivada e polinômio de segunda ordem, centrar novamente os dados na média e construir um novo modelo de PCA.

(14) PCA

```

>>[td,pd,percentd] = kpca(xmscmd);

```

(15) Gráfico dos escores (Figura 6)

```

>>plot(t(1:6,1),t(1:6,2),'xg',t(7:12,1),t(7:12,2),'+r',t(13:18,1),t(13:18,2),'*b',t(19:21,1),t(19:21,2),'ok',t(19:21,1),t(19:21,2),'xk',t(22:24,1),t(22:24,2),'om',t(22:24,1),t(22:24,2),'+m',t(25:27,1),t(25:27,2),'oc',t(25:27,1),t(25:27,2),'*c')
grid

```

```

for i=1:27;
text(t(i,1),t(i,2),num2str(i))
end
xlabel('PC1 (74.5796%)');
ylabel('PC2(21.7213%)');
title('Centrado na média + 1a derivada')
legend('Azeite','Canola','Soja');

```

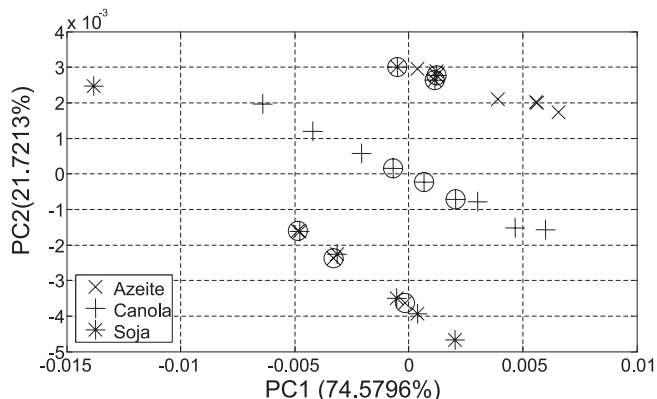


Figura 6. Gráfico de escores da PC1 x PC2 referente às 27 amostras de óleos vegetais comerciais

Finalmente, analisando o gráfico dos escores mostrado na Figura 6 é possível concluir que as amostras 19, 20 e 21 são de óleo de soja, as amostras 22, 23 e 24 são de óleo de canola e, as amostras 25, 26, 27 são de azeite de oliva. O gráfico dos pesos (não mostrado), como esperado, é equivalente ao gráfico mostrado na Figura 3.

Uma alternativa para fazer a PCA dos dados é utilizando uma função interna do Matlab e do Octave que executa o SVD. O Quadro 1 apresenta uma sequência de passos para execução do algoritmo, tanto em Matlab quanto em Octave. Os resultados obtidos utilizando-se o KPCA (EVD) e o SVD são exatamente os mesmos. O Quadro 1 apresenta todos os comando utilizados para executar o SVD a partir dos dados digitados na janela “editor”. Esta é uma alternativa de utilização destes dois softwares que consiste em digitar todos os comandos primeiramente no “editor” e só depois levá-los para o “command window”. O editor é utilizado na construção de algoritmos, portanto, numa etapa inicial, pode ser importante utilizá-lo para executar os comandos apresentados neste tutorial para se familiarizar com esta importante ferramenta do Matlab e do Octave.

Quadro 1. Sequência de comandos executados construção da PCA das 18 amostras de óleo a partir do algoritmo SVD

```

% PCA
[u,s,v]=svd(x);
% Porcentagem de variância explicada
percent = diag(s)/sum(diag(s))
% Gráfico dos escores PC1xPC2
t=u*s(:,1:10);
figure
plot(t(1:6,1),t(1:6,2),'xg',t(7:12,1),t(7:12,2),'+r',t(13:18,1),t(13:18,2),'*b')
grid
for i=1:18;
text(t(i,1),t(i,2),num2str(i))
end
xlabel('PC1');
ylabel('PC2');
% Gráfico dos pesos do 2o. PC
figure
plot(num2,v(:,2))
xlabel('1/cm');
ylabel('PC2(17.7318%)');

```

CONCLUSÃO

Este tutorial pode ser utilizado como um guia de experimento porque possibilita aos alunos e professores manter um histórico de todas as operações executadas. Os autores recomendam para aqueles que não dispõem do Matlab, o uso do software livre Octave, onde todos os comandos executados no Matlab podem ser aplicados diretamente ou adaptados com facilidade.

Foram utilizadas rotinas disponíveis na internet para a execução da PCA e de alguns pré-processamentos como normalização e MSC, entretanto, a partir dos conceitos básicos sobre PCA e a utilização do Matlab para esta finalidade, é fortemente recomendado que o aluno ou pesquisador procure formular suas próprias rotinas de trabalho nestes softwares.

Pretende-se continuar a divulgação de experimentos didáticos de quimiometria, com novos tutoriais baseados na experiência dos autores com a introdução de um módulo de quimiometria na disciplina de Química Analítica Instrumental ministrada na Unicamp para alunos do curso de bacharelado em Química. O próximo tutorial deve versar sobre um experimento para realização de calibração multivariada em análise de fármacos.

MATERIAL SUPLEMENTAR

No material suplementar, disponível em <http://quimicanova.s bq.org.br>, na forma de arquivo PDF, com acesso livre, a Figura 1S apresenta uma cela desmontável para amostras líquidas; a Figura 2S, a janela padrão do Matlab; a Figura 3S, os conjuntos de espectros das 18 amostras de óleo; o Quadro 1S, a sequência de comandos executados para carregar e pré-processar, respectivamente, os 18 espectros das amostras de óleos; o Quadro 2S apresenta a sequência de comandos para a execução da PCA a partir dos dados centrados na média e derivados pela primeira derivada; o Quadro 3S, a sequência de comandos executados para a execução da PCA a partir dos dados centrados na média e derivados através da segunda derivada; o Quadro 4S, a sequência de comandos para a execução da PCA, sendo os espectros pré-processados com o MSC e centrados na média, respectivamente; o Quadro 5S, a sequência de comandos executados para carregar os 27 espectros das amostras de óleos e selecionar as faixas espectrais de interesse e, o Quadro 6S a sequência de comandos executados para executar a PCA a partir dos dados centrados na média e derivados através da primeira derivada. A Tabela 1S apresenta os tipos de óleos vegetais, a quantidade de medidas realizadas para cada tipo de óleo e a ordem das amostras na matriz de dados **X**.

AGRADECIMENTOS

Ao Programa Estágio Docente (PED) da Unicamp, aos Drs. E. H. Novotny e P. H. Março e aos doutorandos M. C. Breikreitz e G. P. Sabin pela revisão do trabalho e constantes contribuições; também ao H. D. Machado pelas fotos e aos órgãos de fomento à pesquisa CAPES, CNPq e FAPESP.

REFERÊNCIAS

- Barros Neto, B.; Scarminio, I. S.; Bruns, R. E.; *Quim. Nova* **2006**, *29*, 1401.
- Houghton, T. P.; Kalivas, J. H.; *J. Chem. Educ.* **2000**, *77*, 1314.
- Wang, L. Q.; Mizaikoff, B.; Kranz, C.; *J. Chem. Educ.* **2009**, *86*, 1322.
- Lima, K. M. G.; Trevisan, M. G.; Poppi, R. J.; de Andrade J. C.; *Quim. Nova* **2008**, *31*, 700.
- Cazar, R. A.; *J. Chem. Educ.* **2003**, *80*, 1026.

6. Rusak, D. A.; Brown, L. M.; Martin, S. D.; *J. Chem. Educ.* **2003**, *80*, 541.
7. Oles, P. J.; *J. Chem. Educ.* **1998**, *75*, 357.
8. Wu, W.; Massart, D. L.; de Jong, S.; *Chemom. Intell. Lab. Syst.* **1997**, *36*, 165.
9. Moita Neto, J. M.; Moita, G. C.; *Quim. Nova* **1998**, *21*, 467.
10. de Sousa, R. A.; Borges Neto, W.; Poppi, R. J.; Baccan, N.; Cadore, S.; *Quim. Nova* **2006**, *29*, 654.
11. da Silva, J. B. P.; Malvestiti, I.; Hallwass, F.; Ramos, M. N.; Leite, L. F. C. da C.; Barreiro, E. J.; *Quim. Nova* **2005**, *28*, 492.
12. Sinelli, N.; Cosio, M. S.; Gigliotti, C.; Casiraghi, E.; *Anal. Chim. Acta* **2007**, *598*, 128.
13. De Luca, M.; Terouzi, W.; Ioele, G.; Kzaiber, F.; Oussama, A.; Oliverio, F.; Tauler, R.; Ragno, G.; *Food Chem.* **2011**, *124*, 1113.
14. Poulli, K. I.; Mousdis, G. A.; Georgious, C. A.; *Food Chem.* **2009**, *117*, 499.
15. Reid, L. M.; O'Donnell, C. P.; Downey, G.; *Trends Food Sci. Technol.* **2006**, *17*, 344.
16. Gurdeniz, G.; Ozen, B.; *Food Chem.* **2009**, *116*, 519.
17. Rohman, A.; Man, Y. B. C.; *Food Res. Int.* **2010**, *43*, 886.
18. Holler, F. J.; Skoog, D. A.; Crouch, S. R.; *Princípios de Análise Instrumental*, 6ª ed., **Bookman: Porto Alegre, 2009**.
19. Brereton, R.; *Chemometrics for Pattern Recognition*, John Wiley & Sons: Chichester, 2007.
20. Wold, S.; Esbensen, K.; Geladi, P.; *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37.
21. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**, *185*, 1.
22. Rinnan, A.; van den Berg, F.; Engelsen, S. B.; *Trends Anal. Chem.* **2009**, *28*, 1201.
23. <http://www.vub.ac.be/fabi/publiek/index.html>, acessada em Julho 2010.
24. Yang, H.; Irudayaraj, J.; *J. Am. Oil Chem. Soc.* **2000**, *77*, 291.