ISSN 2359-4799
Volume 9 / Número 1 / Ano 2023 – p. 01-14
DOI: 10.36524/ric.v9i1.1868

REVISTA ifes ciência

Tutorial para aplicação didática de quimiometria em software gratuito – Parte I: Análise de Componentes Principais em dados de infravermelho médio e propriedades físico-químicas de amostras de petróleo

Tutorial for didactic application of chemometrics in free software – Part I: Principal Component Analysis in petroleum mid-infrared data

[1*]Gabriely Silveira Folli
[2]Pedro Henrique Pereira da Cunha
[3]Mariana Kuster Moro
[4]Paulo Roberto Filgueiras

[1]Laboratório de Quimiometria do Centro de Competência em Química do Petróleo − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo, 29075-910, Brasil. E-mail: gabriely.folli@ufes.br.
[2]Laboratório de Quimiometria do Centro de Competência em Química do Petróleo − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo, 29075-910, Brasil. E-mail: pedro.h.cunha@edu.ufes.br.
[3]Instituto Federal da Bahia (IFBA), campus Porto Seguro, R. José Fontana, 1, Porto Seguro - BA, 45810-000, Brasil. E-mail: marianakustermoro@gmail.com.
[4]Laboratório de Quimiometria do Centro de Competência em Química do Petróleo − NCQP, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo, 29075-910, Brasil. E-mail: paulo.filgueiras@ufes.br.
*Autor de correspondência

Artigo submetido em 30/08/2022, aceito em 02/03/2023 e publicado em 10/03/2023.

Resumo: O desenvolvimento tecnológico aliado ao avanço da instrumentação trouxe novos desafios para os químicos, o de extrair informações de grandes bancos de dados. Neste contexto surgiu a quimiometria, área da química dedicada ao tratamento de dados analíticos de origem multivariada. Por ser uma área relativamente nova, materiais didáticos a respeito do tema são escassos e abordagens de quimiometria ficam restritas a cursos de pós-graduação. Assim, este artigo apresenta um tutorial para aplicação, em sala de aula, em diferentes níveis acadêmicos, de um de seus métodos mais usuais e importantes: a Análise por Componentes Principais (PCA, do inglês *Principal Component Analysis*). Uma análise exploratória de amostras de petróleo por espectroscopia na região do infravermelho médio (MIR) e por propriedades físico-químicas é utilizada para o ensino do método. Apresentamos e explicamos todas as etapas para elaboração de um modelo completo de PCA, desde a instalação do *software* gratuito GNU Octave até a elaboração das figuras finais. São disponibilizados todos os algoritmos desenvolvidos para leitura e tratamento dos dados analíticos. Ao final, é feita uma interpretação acerca dos resultados obtidos, de modo que as discussões e conclusões sejam facilmente compreendidas por todos.

Palavras-chave: quimiometria; PCA; análise exploratória; tutorial; Octave.

Abstract: Technological development allied to the instrumentation advancement brought new challenges for chemists, to extract information from large databases. In this context, chemometrics emerged, an area of chemistry dedicated to the analytical data treatment of multivariate origin. As it is a relatively new area, teaching materials on the subject are scarce and chemometric approaches are restricted to postgraduate courses. Thus, this article presents a tutorial for applying Principal

Component Analysis (PCA), one of the most common and important methods. A tutorial that can be performed at different academic levels and applied in the classroom. A crude oil samples exploratory analysis by mid-infrared region (MIR) spectroscopy and by physicochemical properties were used to teach the method. We present and explain all the steps for the elaboration of a complete PCA model, from the installation of the free software GNU Octave to the elaboration of the final plots. All algorithms developed for reading and processing analytical data are available. Finally, an interpretation is made of the results obtained, so that the discussions and conclusions are easily understood by all.

Keywords: chemometrics; PCA; exploratory analysis; tutorial; Octave.

ISSN 2359-4799
Volume 9 / Número 1 / Ano 2023 – p. 01-14
DOI: 10.36524/ric.v9i1.1868

REVISTA Ifes ciência

1 INTRODUÇÃO

Sistemas químicos reais, muitas vezes, apresentam uma complexidade inerente, sendo necessárias ferramentas avançadas de análise de dados. Com o desenvolvimento tecnológico e o consequente aumento da capacidade de processamento dos computadores, a Quimiometria vem contribuindo cada vez mais, sobretudo na área de Química Analítica, com a elucidação de sistemas químicos complexos, principalmente os sistemas multivariados, que normalmente dispõem de grande volume de dados (ADAMS, 1995). Em vista disso, é crescente a disseminação da Quimiometria em ambientes acadêmicos como laboratórios de pesquisa, laboratórios de ensino e salas de aula, bem como o número de publicações envolvendo o tema (NETO, 2006). Apesar do crescimento, pesquisas apontam que a abordagem do assunto nos centros de ensino ainda é insuficiente, mesmo em cursos superiores de Química (PEREIRA, 2014). Na maioria deles, a Quimiometria não faz parte da ementa do curso, portanto, não é abordada pelos docentes. Quando faz parte, muitas vezes, é apresentada de forma superficial e apenas teórica, sem uma aplicação prática (PEREIRA, 2014). Ademais, os docentes comumente enfrentam dificuldades para encontrar bons materiais didáticos com aplicações.

Dada a escassa difusão da quimiometria entre alunos e professores dos cursos de graduação, sua dificuldade de implementação prática e a grande aplicabilidade de seus métodos na área de Química Analítica, este artigo foi elaborado com o objetivo de orientar, de forma didática e objetiva, docentes e alunos na utilização de alguns métodos quimiométricos por meio de sua aplicação em dados reais. O objetivo principal não é explorar as bases teóricas da Quimiometria, mas sim apresentar um tutorial com aplicabilidade prática, permitindo que docentes e discentes utilizem suas ferramentas com facilidade, como forma de difusão do assunto, principalmente, em ambientes acadêmicos.

O conjunto de algoritmos, ou seja, uma sequência de códigos que seguem uma determinada ordem com objetivo de realizar o processamento de interesse é denominada rotina. Neste artigo descreveremos, detalhadamente, o passo-a-passo para aplicação do método de reconhecimento de padrões mais conhecido, Análise por Componentes Principais (PCA, do inglês *Principal Component Analysis*), em dois conjuntos diferentes de dados reais, provenientes de amostras de petróleo. A descrição abordará desde a instalação do software, obtenção dos pacotes de dados, passando pela construção das rotinas, obtenção dos modelos e gráficos, até a análise e interpretação dos resultados.

A PCA é um método não-supervisionado de análise exploratória com objetivo de reduzir a dimensionalidade dos dados. A redução é feita a partir da condensação das informações importantes em componentes principais (PC, do inglês *Principal Component*). Essa decomposição facilita a interpretação do sistema, permitindo obter resultados e conclusões que não seriam alcançáveis por meio das variáveis originais (WOLD, 1987; JOLLIFFE, 2015). A PCA já é utilizada com grande aplicabilidade em sistemas químicos, permitindo identificar amostras anômalas, relações entre variáveis e

agrupamentos ou relações entre amostras. Neste artigo, a análise de PCA é dada a partir de duas matrizes diferentes: uma contendo dados de espectros de infravermelho na região do médio (característica de dados contínuos) e dados discretos de 10 propriedades físico-químicas. Ambas as matrizes são referentes a amostras de petróleo provenientes de diferentes origens geográficas.

2 REFERENCIAL TEÓRICO

A PCA está alocada em uma subárea da quimiometria denominada de reconhecimento de padrão não supervisionado. Na PCA é realizada a decomposição da matriz de dados $\mathbf{X_{ij}}$ em matriz de scores ($\mathbf{T_{ih}}$) e matriz de *loadings* ($\mathbf{P_{hj}}$) mais uma matriz de resíduo ($\mathbf{E_{ij}}$). O subíndice *i* representa o número de amostras (linhas) e *j* o número de variáveis (colunas). A Equação 1 apresenta o cálculo da PCA (WOLD, 1987).

$$X_{ij} = T_{ih} \cdot P_{hj}^{t} + E_{ij}$$ **Equação 1**

Os vetores ortogonais ($\mathbf{t_n p_n^t}$), também conhecidos por PCs, são definidos pela combinação linear das variáveis originais e ordenadas pela quantidade de informação (variância) presente em cada uma delas, como mostra a **Equação 2**. O subíndice *n* representa o número de componentes principais. A primeira componente principal (PC1) retém a maior fonte de informação, porque ela provém da primeira decomposição da matriz **X**. Já, a PC2 apresenta a segunda maior fonte de informação, sendo ortogonal a PC1, e assim por diante (WOLD, 1987).

$$X_{ij} = t_1 p_1^t + t_2 p_2^t + \cdots + t_n p_n^t + E_{ij}$$

Equação 2

O objetivo da PCA é encontrar relações naturais entre as variáveis e por consequência reduzir a dimensão de sua matriz. Deseja-se reduzir a quantidade de variáveis em um novo conjunto de variáveis não correlacionadas. O resultado pode ser expresso por meio de gráficos de *scores* e *loadings*, cujos eixos são as PCs. A matriz de *scores* contém informações das amostras nas novas coordenadas de componentes principais (ao novo sistema de eixos com menores dimensões), assim, sua projeção pode indicar agrupamento de amostras. Enquanto que os *loadings* indicam a contribuição de cada variável original nas componentes principais (BRERETON, 2003; CHO, 2014; CENTNER, 1998), seu gráfico indica quais variáveis estão mais correlacionadas entre si e quais contribuem mais para um agrupamento de amostras ou separação delas.

3 PROCESSOS METODOLÓGICOS/MATERIAIS E MÉTODOS

Neste trabalho, foram utilizados dados contínuos (espectros no infravermelho) e discretos (propriedades físico-químicas). Estes dados foram aplicados para construção e aplicação dos exemplos do tutorial de PCA desenvolvidos pelos autores. Todas as rotinas, funções e dados de entrada (espectros e propriedades físico-química) estão disponíveis no GitHub dos autores (https://github.com/PHPCunha/IFES-Ciencia).

3.1 MIR

Espectros de infravermelho na região do médio (MIR, do inglês *Mid-infrared spectroscopy*) de 200 amostras de petróleo foram adquiridos. Estes espectros possuem caráter contínuo, ou seja, existem variáveis próximas representando a mesma informação química. Assim, há grande correlação entre estas variáveis (colinearidade). Portanto, os dados de MIR foram aplicados no exemplo utilizando dados contínuos e demonstrados por meio dos gráficos resultantes.

3.2. PROPRIEDADES FÍSICO-QUÍMICAS

Setenta amostras foram utilizadas para construção deste trabalho. Os modelos foram construídos a partir das propriedades

físico-química: grau API, viscosidade, ponto de fluidez (PF), teor de compostos saturados, aromáticos, resinas e asfaltenos, fator de caracterização de *Universal Oil Products Company* (KUOP), teor de enxofre e número de acidez total (NAT). As propriedades físico-químicas possuem caráter discreto. Com isso, foram utilizadas para construção de modelos de PCA para exemplificar sua aplicação e gráficos construídos para dados discretos.

3.3. *SOFTWARES* E ALGORÍTMOS

Os algoritmos foram construídos no software GNU Octave (John W. Eaton, versão 7.1.0, 2022). A versão 7.1.0 do programa pode ser obtida gratuitamente no endereço https://www.gnu.org/software/octave/. As rotinas, funções, espectros, pacotes e propriedades físico-química estão disponíveis no GitHub dos autores (https://github.com/PHPCunha/IFES-Ciencia). É recomendado abrir a rotina nomeada de “01 - Tutorial_PCA” para dar sequência ao tutorial.

4 RESULTADOS E DISCUSSÃO

4.1. OBTENÇÃO E INSTALAÇÃO DO *SOFTWARE* E DOS PACOTES DE DADOS

Primeiramente, é necessário a obtenção do programa GNU Octave pelo endereço https://www.gnu.org/software/octave/. Ao acessar a página, é preciso clicar na aba referente ao sistema operacional do computador, baixar e instalar o programa. Após a instalação, são disponibilizados dois atalhos na área de trabalho. Recomenda-se utilizar a opção “GNU Octave (GUI)” cuja interface é mais organizada e adequada para iniciantes.

Para conduzir as operações matemáticas requeridas pela PCA, é necessário que os pacotes “statistics-1.4.3.tar.gz” e “io-2.6.4.tar.gz” sejam instalados no Octave. Estes pacotes, necessários também para aplicação de diversos outros métodos quimiométricos, podem ser encontrados no GitHub dos autores. O primeiro contém ferramentas estatísticas e o segundo permite a construção e edição de planilhas.

Recomendamos abrir a rotina “00 - Instalação de pacotes”, também disponível no GitHub, para acompanhar o passo-a-passo da instalação dos pacotes.

Após a instalação dos pacotes, o usuário deve abrir o Octave na janela “Editor”, copiar o endereço da pasta onde estão salvos os pacotes e colar conforme a seguir, selecionando o diretório de interesse:

```
>>cd('C:\Users\Exemplo\...\Pacotes
');
```

A seguir, instalar os dois pacotes salvos nesta pasta (“statistics-1.4.3.tar.gz” e “io-2.6.4.tar.gz”):

```
>> pkg install statistics-1.4.3.tar.gz
```

```
>> pkg install io-2.6.4.tar.gz
```

Para executá-los, basta selecionar as linhas dos comandos e apertar a tecla F9, excluindo os “>>”, se não excluí-los resultará em erro. É possível verificar se a instalação ocorreu corretamente por meio do comando “pkg list”, o qual irá listar, na janela de comandos, todos os pacotes instalados. Para concluir, os pacotes devem ser carregados no programa por meio dos comandos:

```
>> pkg load statistics
```

```
>> pkg load io
```

4.2. TRATAMENTO DOS DADOS E EXECUÇÃO DA PCA A PARTIR DOS ESPECTROS DE MIR

A matriz de dados **X**, ou seja, os espectros de infravermelho médio, obtidos na etapa experimental, encontrados no GitHub dos autores, estão armazenados no “MIR_data.mat”. Após o download, para carregá-los no Octave, ainda na janela

"Editor", é preciso direcionar o diretório no endereço onde os dados foram salvos no computador do usuário (conforme realizado para instalar os pacotes no passo 4.1), digitar e executar:

>>cd('C:\Users\Exemplo\...\IFES Ciencia');

>> load('MIR_data.mat');

Todas as funções são acompanhadas por explicações internas. Caso tenha dúvidas do funcionamento de algum comando, pode-se utilizar a função de ajuda a seguir. Nesse exemplo, a função que utilizamos para visualizar foi a função plot. Essa função é dada para visualização de gráficos. Então, ao prosseguir com essa função, o octave irá explicar todas as formas de utilização da função plot e a funcionalidade de cada uma.

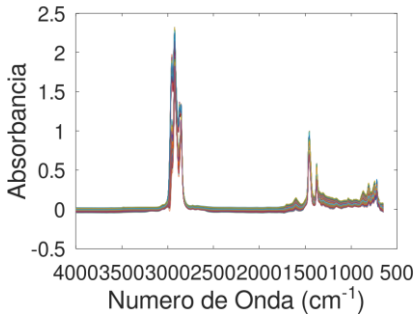
>> help plot;

Para visualizar o espectro é bem simples, pode-se fazer uso do comando a seguir para plotar o gráfico do espectro. A primeira entrada é dada pelo eixo da abscissa e, no nosso exemplo, é respectivo à distribuição dos valores de comprimento de onda, **num**. Já, o eixo Y (não confundir com **vetor y**), é dado pelos valores de absorbância dispostos na nossa matriz de dados **X**.

>> plot(num,MIRdata)

O carregamento dos dados dará origem ao vetor **y** com 200 linhas e à matriz MIRdata com 200 linhas e 3351 colunas. Cada linha da matriz contém os valores de absorbância do espectro de uma amostra, totalizando 200 linhas (ou seja, 200 amostras). Cada coluna representa uma variável e contém os valores espectrais de todas as amostras em um determinado comprimento de onda, totalizando 3351 variáveis, **Figura 1**.

Figura 1. Gráfico dos espectros de MIR brutos.



Fonte: Autor.

O **vetor y** é formado por 4 classes numéricas, variando de 1 a 4, de forma que cada número representa uma origem geográfica diferente onde o petróleo foi encontrado.

Para aplicar a PCA, é preciso fazer o download da pasta "PCA", disponível no GitHub dos autores (https://github.com/PHPCunha/IFES-Ciencia), na qual contém todas as rotinas necessárias. Em seguida, direcionar o diretório do Octave para esta pasta (conforme realizado para instalar os pacotes no passo 4.1):

>>cd('C:\Users\Exemplo\...IFES Ciencia\PCA');

A função pcamodel, antes da execução da PCA, realiza um pré-processamento dos espectros com os métodos escolhidos. O pré-tratamento consiste na correção espectral das amostras para suavizar possíveis variações indesejadas (ruídos ou interferentes). Tem como objetivo evidenciar as variâncias dispostas no grupo amostral e traduzir em menor conjunto de informações mais relevantes (mais importantes). Ou seja, reduzir informações indesejáveis que podem interferir na construção dos modelos, tornando-os mais robustos e livres de informações não desejadas (Ferreira, 2015).

Existem diferentes pré-processamentos disponíveis como o método de correção multiplicativa de sinal (MSC, do inglês multiplicative scatter correction) (Dhanoa, 1994), centralização na média (center) autoescalamento dos

dados (auto), variação de padrão normal (snv, do inglês Standard Normal Variate) (Dhanoa, 1994), derivada (deriv), cada um com uma finalidade específica. A escolha do melhor pré-tratamento é feita ao analisar os resultados obtidos (gráfico de *scores* e *loadings*) explicados à posteriori. Neste caso, foram escolhidos o msc e center por apresentar os melhores resultados (melhores separações em relação à origem geográfica).

```
>> pretrat = {'center'};
```

Também podem ser realizadas combinações entre os pré-tratamentos, separando-os por vírgulas:

```
>> pretrat = {'msc','center'};
```

Após o passo anterior, será criado o modelo a partir do pré-tratamento escolhido com a função "pcamodel". Nota-se que são necessárias quatro informações de entrada (inputs) para prosseguir com a função. A primeira refere-se à matriz de espectros "MIRdata", a segunda ao pré-tratamento escolhido no passo anterior, a terceira ao número de componentes principais (PC) desejado e a última ao vetor de classes (vetor **y**).

```
>>modelo=pcamodel(MIRdata,pretrat,PC,y);
```

O número limite de componentes principais é dado pelo número de variáveis (quantidade de colunas da matriz de dados ("MIRdata") subtraído de 1 unidade. Isso porque o objetivo da PCA é redimensionar e comprimir os dados, por isso a quantidade máxima de componentes principais deve ser inferior ao número de variáveis.

Apesar de estarmos construindo um modelo não supervisionado, o vetor **y** foi incluído na função, entretanto, ele não é utilizado nos cálculos do modelo, mas apenas na construção e interpretação dos gráficos que serão gerados adiante. Com isso, será realizada a escolha do melhor pré-tratamento de acordo com a disposição dos resultados.

Na Janela de Comandos, a função apresenta como resposta uma tabela contendo a porcentagem de variância explicada pelo modelo PCA para cada componente principal, **Tabela 1**. A variância é dada pela quantidade de informação da matriz **X** que foi possível explicar com a PC. De acordo com os resultados na tabela, 84,99% de toda informação contida na matriz **X** foi explicada pela PC1, 14,62% pela PC2, 0,37% pela PC3, 0,02% pela PC4 e 0,00% pela PC5. É possível perceber que a maior concentração de informação explicada (variância) está contida na PC1, isso porque ela é resultado da primeira decomposição da matriz **X**. Já a PC2 é resultado do resíduo da primeira decomposição, enquanto a PC3 é o resíduo da segunda. Dessa forma, a PC1 será sempre a componente que explica a maior parte da informação contida na matriz **X**, a PC2 será a segunda, e assim por diante. Além disso, pode-se identificar, **Tabela 1**, a variância somada pelas PCs (variância total). Verifica-se que a soma das PC1, PC2, PC3 e PC4 explicou 100% da informação disposta na matriz **X**. Ou seja, somente 4 componentes principais foram suficientes para extrair 100% da informação disposta no conjunto de dados.

Tabela 1. Porcentagem de variância explicada pelo modelo PCA a partir da matriz de espectros.

PC	Variância explicada (%)	Variância Total (%)
1	84.99	84,99
2	14,62	99,60
3	0,37	99,98
4	0,02	100,00
5	0,00	100,00

Fonte: Autor.

Para obter os gráficos da PCA (*scores*, pareto e *loadings*) basta executar os comandos a seguir. O dígito 1 indica o

aparecimento do gráfico e o dígito 0 o não aparecimento.

>> options.Score = 1;

>> options.Pareto = 1;

>> options.Loading = 1;

Em seguida, executa-se a função "pcaplot" para plotar os gráficos pretendidos. Nessa função, são necessários quatro inputs para sua execução. O primeiro é o modelo PCA criado, o segundo e o terceiro são as PCs que irão compor os gráficos, no nosso caso, utilizamos a PC1 e PC2; o último input representa as opções de gráficos por meio da variável "*options*", recém-criada. As componentes principais escolhidas darão origem aos gráficos de *scores* e *loadings*.

>> pcaplot(modelo,1,2,options);

É possível obter o gráfico de *scores* com outras componentes principais, conforme mostrado a seguir. No qual foram utilizadas a PC2 e PC3. Por exemplo:

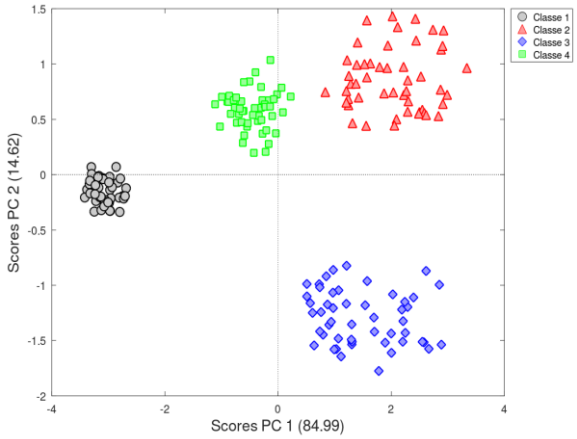
>>pcaplot(modelo,2,3,options)

Utilizando a PC1 e PC2 como inputs, os seguintes gráficos são gerados: gráfico de *scores* da PC1 versus *scores* PC2 (**Figura 2**), espectro original e gráfico de *loadings* da PC1 e PC2 (**Figura 3**) e gráfico de pareto das variâncias das PCs (**Figura 4**).

O gráfico de *scores* apresenta a disposição espacial das amostras em relação aos seus respectivos agrupamentos. Ou seja, amostras que possuem menor variância (variação ou diferença) entre si, estarão mais próximas. Enquanto que amostras que possuem maior variância, estarão mais espaçadas. Cada classe está designada por uma cor e forma. A classe 1 apresenta forma de círculo cinza, a classe 2 apresenta forma de triângulo vermelho, a classe 3, losango azul, e a classe 4, quadrado verde. Nota-se que as amostras estão agrupadas conforme as classes (origem geográfica). Portanto, o objetivo de separar as amostras de acordo com sua localização geográfica foi alcançado, de acordo com o gráfico de *scores* PC1 versus PC2, utilizando o pré-tratamento msc e center.

Observa-se que as classes 2 e 3 possuem maiores semelhanças (estão mais agrupadas) em comparação com as demais classes (1 e 3). Ou seja, elas apresentam menores diferenças entre si. Além disso, por meio da PC2, é possível separar as classes 2 e 4 das classes 1 e 4. Já, a PC1 foi capaz de separar as classes 2 e 3 da maioria das amostras da classe 4 e 1.

Figura 2. Gráfico de *scores* da PC1 versus PC2 para PCA criada a partir da matriz de espectros.

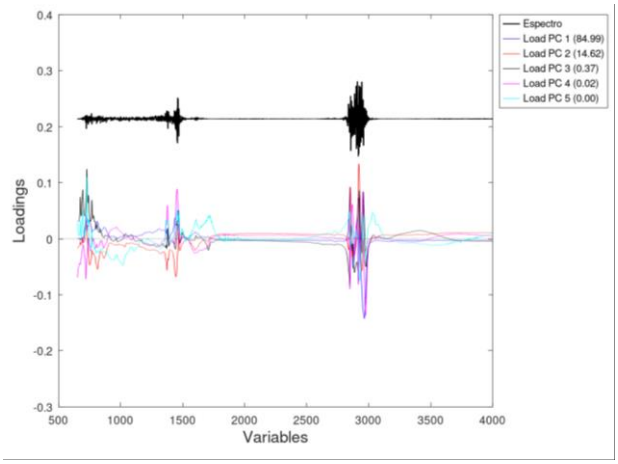


Fonte: Autor.

O gráfico de *loadings*, **Figura 3**, apresenta a importância das variáveis originais na construção da PCA. No exemplo, utilizamos espectros de MIR, logo podemos destacar as regiões espectrais mais importantes para o comportamento das amostras. O primeiro gráfico apresenta o espectro médio original, em que o eixo Y é dado pelos valores de absorbância e o eixo X pelo número de onda. Já o segundo gráfico apresenta no eixo y os *loadings*, que representam a importância das variáveis para a construção do modelo e o eixo X o número de onda. Nota-se que o maior valor absoluto de *loading* traduz na maior importância dessa variável para a

separação ou agrupamento do conjunto amostral, seja ela positiva ou negativa. Com isso, para esse exemplo, o número de onda de 1000-1200 cm^{-1} foi a região espectral de maior importância para as separações ou agrupamentos.

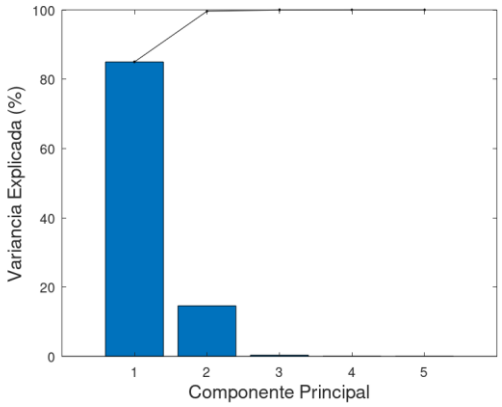
Figura 3. Espectro médio e Gráfico de *loadings* da PCA criada a partir da matriz de espectros.



Fonte: Autor.

O gráfico de Pareto, **Figura 4**, apresenta as mesmas informações dispostas na **Tabela 1**. As barras azuis indicam a variância explicada por cada PC e a linha crescente indica o somatório de variância explicada entre as PCs, ou seja, a variância total. Assim, a PC1 e a PC2, utilizadas para descrever nossos dados, foram capazes de explicar 99,60% de toda informação disposta na matriz ***X***.

Figura 4. Gráfico de Pareto das PCs para PCA criada a partir da matriz de espectros



Fonte: Autor.

4.3. TRATAMENTO DOS DADOS E EXECUÇÃO DA PCA A PARTIR DAS PROPRIEDADES FÍSICO-QUÍMICAS

Para obtenção da PCA a partir das propriedades físico-químicas, é aconselhável fazer a limpeza das variáveis da memória do Octave, bem como da janela de comandos, além de fechar janelas abertas, por meio dos comandos:

```
>> clc;
>> clear all;
>> close all;
```

Fazendo isso, será necessário carregar os pacotes novamente por meio dos comandos:

```
>> pkg load statistics
>> pkg load io
```

As propriedades físico-químicas, obtidas na etapa experimental, encontradas no GitHub dos autores (https://github.com/PHPCunha/IFES-Ciencia), estão armazenadas em um arquivo txt "Petroleum.txt". Após o download, para carregá-los no Octave, na janela "Editor", é preciso direcionar o diretório para o endereço onde os dados foram salvos no computador do usuário (conforme realizado para instalar os pacotes no passo 4.1):

```
>>cd('C:\Users\Exemplo\...\IFES Ciencia');
```

Em seguida, digitar os seguintes comandos:

```
>> Dados = load('Petroleum.txt');
>> X = Dados(:,3:end);
>> y = Dados(:,2);
>> amostra = Dados(:,1);
```

O carregamento dos dados dará origem ao vetor ***y*** com 70 linhas, de forma que cada linha representa uma amostra. Da mesma forma feita no passo anterior (passo 4.2), o vetor y é formado por 4 classes de números de 1 a 4 e cada número representa uma origem geográfica diferente.

O carregamento dos dados também dará origem à matriz **X** com 70 linhas e 10 colunas. Cada linha representa uma amostra e contém valores das propriedades físico-químicas. Ademais, cada coluna representa uma propriedade físico-química. Em resumo, os dados referem-se a 10 propriedades físico-químicas de 70 amostras com classes representando 4 diferentes origens geográficas.

Para aplicar a PCA a estes dados, é preciso direcionar o diretório do Octave para a pasta PCA (conforme realizado para instalar os pacotes no passo 4.1):

>>cd('C:\Users\Exemplo\...\PCA');

Para dados discretos, é necessário realizar um pré-processamento com o método autoescalonamento (auto). Este método foi escolhido porque as variáveis discretas (propriedades físico-químicas), muitas vezes, possuem ordem de grandeza muito diferente umas das outras, fazendo com que as de maior ordem se sobreponham às de menor. O pré-tratamento autoescalamento, além de centrar os dados na média, realiza uma transformação de forma que as variáveis adquiram igual importância nos cálculos do modelo PCA, equilibrando os pesos das informações de cada variável independente. Para isso, pode-se usar o pré-tratamento "auto" ou combiná-lo com qualquer outro pré-tratamento. Nesse exemplo, foi utilizado somente o pré-tratamento "auto" que além de autoescalar os dados, também os centraliza na média. Para executar o pré-processamento:

>> pretrat = {'auto'};

A etapa de criação do modelo dada pela função da PCA apresenta 4 inputs, já descritos anteriormente. Neste exemplo, foram utilizados como input 9 PCs, uma vez que o total de variáveis é 10 e 9 é o limite de PCs possíveis (PClimite = número de variáveis – 1). Por fim, o vetor y foi incluído na função, não para os cálculos do modelo, mas apenas para construção e interpretação dos gráficos que serão gerados adiante.

>>modelo=pcamodel(X,pretrat,9,y);

Na Janela de Comandos, a função apresenta como resposta a **Tabela 2**, contendo a porcentagem de variância explicada pelo modelo PCA para cada componente principal. Nota-se, diferentemente da PCA a partir dos espectros, a variância explicada é mais distribuída entre as PCs, uma vez que suas variáveis são discretas e, portanto, mais independentes umas das outras. No caso de variáveis contínuas, como os espectros, muitas apresentam as mesmas informações químicas e estão altamente correlacionadas. Enquanto que, em variáveis discretas, há menor quantidade de informações compartilhadas.

Tabela 2. Porcentagem de variância explicada pelo modelo PCA a partir da matriz de propriedades físico-químicas.

PC	Variância explicada (%)	Variância Total (%)
1	47,18	47,18
2	29,64	76,82
3	8,23	85,05
4	6,17	91,22
5	4,04	95,26
6	2,32	97,58
7	1,86	99,43
8	0,40	99,84
9	0,16	99,99

Fonte: Autor.

Para obter os gráficos da PCA, deve-se executar primeiramente os seguintes comandos, em que o dígito 1 representa o aparecimento do gráfico e 0 não aparecimento:

>> options.Score = 1;

>> options.Pareto = 1;

>> options.Loading = 1;

Em seguida, executa-se a função "pcaplot" para plotar os gráficos pretendidos. Nessa função são necessários quatro inputs para sua execução. O primeiro é o modelo PCA criado, o segundo e o terceiro as PCs de interesse que nesse exemplo foram a PC1 e PC2 e o quarto e último as opções dos gráficos criadas anteriormente pela variável "options". As componentes principais escolhidas darão origem aos gráficos de *scores* e *loadings*.

```
>> pcaplot(modelo,1,2,options);
```

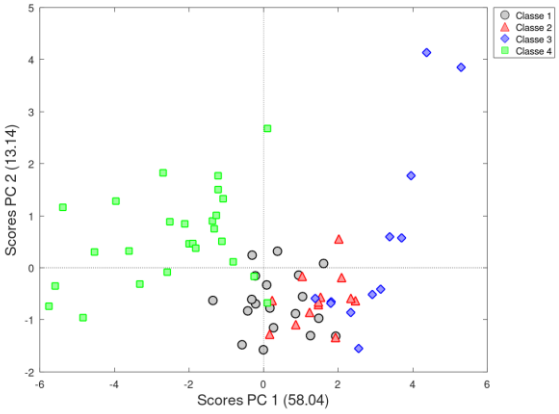
Também utilizamos a combinação da PC1 e PC3, executando a linha a seguir:

```
>>pcaplot(modelo,1,3,options)
```

Utilizando a PC1 e PC2 como inputs, os seguintes gráficos foram gerados: gráfico de *scores* da PC1 *versus* *scores* PC2 (**Figura 5**), gráfico biplot de scores e *loadings* da PC1 *versus* PC2 (**Figura 6**) e gráfico de pareto das variâncias das PCs (**Figura 7**).

O gráfico de *scores* (**Figura 5**) mostra que as amostras pertencentes à classe 4 podem ser separados das outras classes, majoritariamente, pela PC1. Pode-se identificar algumas divisões, tais como a permanência da classe 4 no eixo negativo da PC1 e a classe 3 no eixo positivo da PC1. Além disso, a classe 1 permaneceu, majoritariamente, no eixo negativo da PC2. Todas essas características também são observadas na **Figura 2**. Ou seja, as amostras apresentam características semelhantes mesmo em fontes de informações diferentes. Nota-se que a separação de classes utilizando dados de propriedades físico-químicas PC1 *versus* PC2 não resultou em resultados tão satisfatórios se comparados à PCA utilizando espectros. Isso porque a porcentagem de informação explicada pela PC1 e PC2 das propriedades físico-químicas é inferior (76,82%) se comparada à PC1 e PC2 dos espectros (99,60%).

Figura 5. Gráfico de *scores* da PC1 *versus* PC2 para PCA criada a partir da matriz de propriedades físico-químicas.



Fonte: Autor.

Como existem 10 variáveis estudadas, há a possibilidade em se ter um gráfico biplot de *scores* e *loadings*, observado na **Figura 6**. Para a construção deste gráfico temos que utilizar os seguintes comandos:

```
>> options.Score = 2;
```

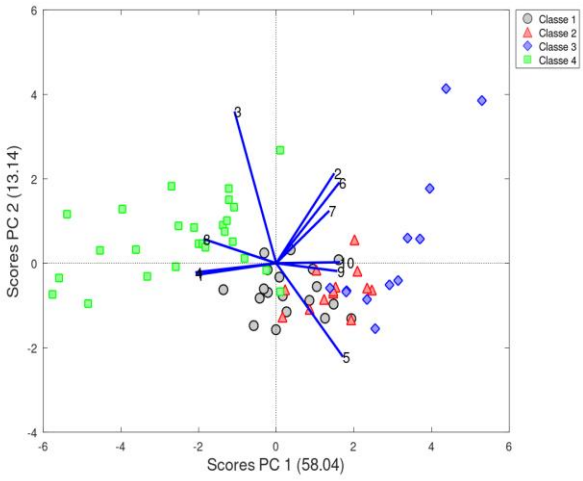
```
>> pcaplot(modelo,1,2,options);
```

Esse gráfico indica a importância de cada variável em relação à disposição especial das amostras. Podemos observar a influência positiva ou negativa da variável em relação ao grupo amostral. Além disso, é possível observar a correlação e independência entre variáveis. Variáveis que são perpendiculares (ângulo de 90º entre si) são ditas independentes, uma vez que o cosseno(90º) entre elas é igual a zero e, portanto, a correlação entre essas variáveis é nula. Já quando as variáveis possuem angulação de 180º entre si, elas são totalmente correlacionadas e inversamente proporcionais, uma vez que cosseno(180º) é igual a -1 e, portanto, sua correlação é máxima e inversa (Ku et al., 1995).

Nota-se que na figura de *biplot* do exemplo, **Figura 6**, a variável 3 (Ponto de Fluidez) influencia negativamente o agrupamento das classes 1 e 2. Ou seja, um menor valor dessa variável ajuda no

agrupamento. Em contrapartida, a variável 5 (Teor de aromáticos) influencia positivamente o agrupamento dessas classes. Com isso, as variáveis 3 (Ponto de Fluidez) e 5 (Teor de aromáticos) são inversamente proporcionais. Já, as variáveis 6 (Teor de Resinas) e 2 (Viscosidade) apresentam baixa correlação entre si, pois são quase perpendiculares. Nota-se que as variáveis mais importantes para o gráfico de PC1 x PC2 foram as variáveis 3 (Ponto de Fluidez) e 5 (Teor de Aromáticos). Isso porque apresentam os maiores vetores, ou seja, maiores importâncias para a atual configuração espacial das amostras.

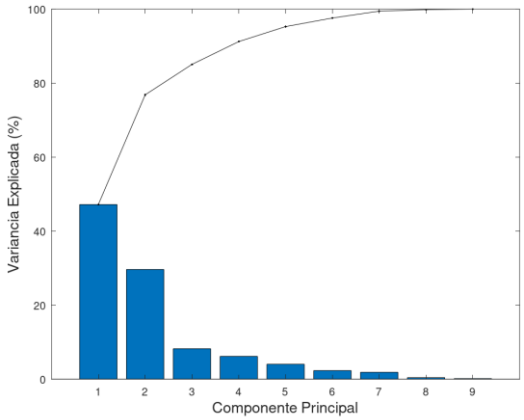
Figura 6. Espectro médio e Gráfico *biplot* de *scores versus loadings* da PCA criada a partir da matriz de propriedades físico-químicas.



Fonte: Autor.

O gráfico de pareto, **Figura 7**, apresenta as informações dispostas na **Tabela 2**. Nota-se que, em comparação com a **Figura 4**, temos maior distribuição da variância entre as componentes principais.

Figura 7. Gráfico de Pareto das PCs para PCA criada a partir da matriz de propriedades físico-químicas.



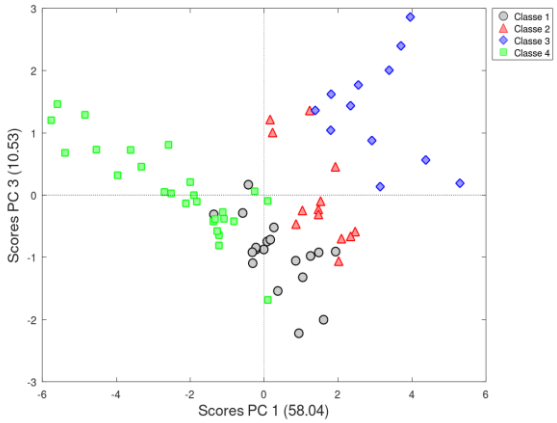
Fonte: Autor.

Ao analisar os resultados obtidos pela PCA utilizando a PC1 *versus* PC2 e a distribuição da variância explicada, também foi utilizado a comparação com outras combinações de componentes principais para elaboração dos gráficos de *loadings* e *scores*. Os melhores resultados foram encontrados pela combinação da PC1 *versus* PC3. Este gráfico foi obtido executando a linha a seguir:

>>pcaplot(modelo,1,3,options)

O gráfico de scores, **Figura 8**, mostra melhor agrupamento das classes de interesse, com menor quantidade de amostras dispersas. Assim como na **Figura 1**, as classes 2 e 3 possuem maiores semelhanças (estão mais agrupadas) em relação às classes 1 e 4. As classes 1 e 4 apresentam maiores semelhanças (maior agrupamento) em relação às classes 2 e 3. Além disso, a classe 3 pode ser separada da classe 1 pela PC3.

Figura 8. Gráfico de *scores* da PC1 *versus* PC3 para PCA criada a partir da matriz de propriedades físico-químicas.



Fonte: Autor.

4.4 EDIÇÃO DE FIGURAS

Todas as legendas e títulos de eixos podem ser editados conforme desejar. Para isso, é recomendado abrir somente a imagem a ser editada e prosseguir com os comandos explicados a seguir. Vamos utilizar a **Figura 8** como exemplo.

Para criar somente o gráfico de *score*s, com PC1 e PC3 você utiliza os seguintes comandos:

```
>> options.Score = 1;
>> options.Pareto = 0;
>> options.Loading = 0;
>> pcaplot(modelo,1,3,options)
```

A função “legend” apresenta os nomes das classes de interesse dadas na legenda do gráfico de *scores*. No nosso caso, temos quatro classes, então, o pcaplot as nomeia automaticamente como “classe 1”, “classe 2”, “classe 3” e “classe 4” que representam poços diferentes de petróleo, logo, podemos modificar para Poço A, Poço B, Poço C e Poço D, usando o seguinte comando:

```
>>legend('Poco A','Poco B','Poco C','Poco D');;
```

O local espacial no qual estará localizada a legenda também pode ser alterado. Algumas opções que podem ser utilizadas são; norte (‘north'), sul ('south'), leste ('east'), oeste ('west'), nordeste ('northeast'), noroeste (‘northwest'), sudeste ('southeast'), sudoeste ('southwest'), norte fora da figura ('northoutside'), sul fora da figura ('southoutside'), leste fora da figura ('eastoutside'), oeste fora da figura, ('westoutside'), nordeste fora da figura ('northeastoutside'), noroeste fora da figura ('northwestoutside'), sudeste ('southeastoutside'), sudoeste fora da figura ('southwestoutside').

O pcaplot foi programado para automaticamente utilizar a localização: ‘northeastoutside', entretanto, para a **Figura 8** podemos utilizar a configuração 'southeast', com o seguinte comando:

```
>>legend('Location','southeast');
```

Para atualizar a fonte dos textos inseridos dentro da figura, basta executar o próximo comando. A fonte dos textos dispostos dentro dos gráficos de exemplo foi o tamanho 20.

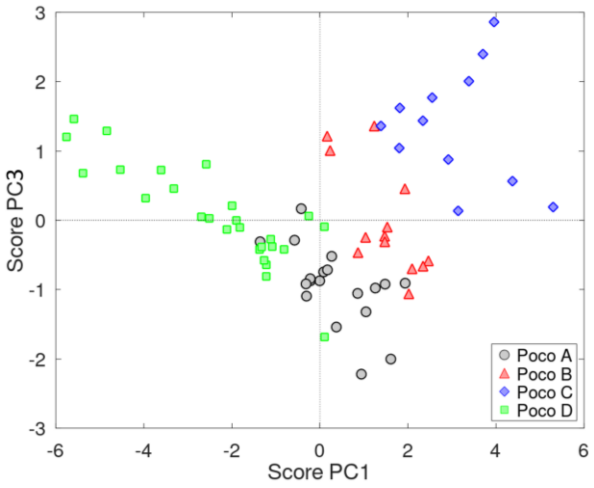
```
>> set(gca,'FontSize',16);
```

Os eixos também podem ser editados, inserindo o nome de interesse. Neste trabalho, pode ser interessante adicionar, “Score PC1” e “Score PC3”, como nome dos eixos. O tamanho da fonte também pode ser alterado, a escolhida foi de 24. A mudança foi realizada com os comandos:

```
>>ylabel('ScorePC2','FontSize',18);
>>xlabel('ScorePC1','FontSize',18);
```

Como resultado, é obtida a **Figura 9**.

Figura 9. Gráfico de *scores* da PC1 *versus* PC2 para PCA criada a partir da matriz de propriedades físico-químicas.



Fonte: Autor.

5 CONCLUSÕES

O tutorial didático para utilização da PCA em software gratuito GNU Octave é uma alternativa didática e de baixo custo para possibilitar a maior disseminação da quimiometria entre todos os níveis acadêmicos. Todas as rotinas e funções criadas dentro do GNU Octave se assemelham ao software Matlab (programa de maior frequência de uso em trabalhos utilizando quimiometria). Ou seja, possibilita maior interação e aprendizado entre os estudantes e professores de química. Além disso, os dados experimentais também estão disponíveis virtualmente de forma gratuita para integralizar melhor todo o conteúdo abordado nesse trabalho.

AGRADECIMENTOS

Os autores agradecem às empresas de fomento Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES) [88887.487966/2020-00], Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) [356/18;83552723, 442/2021, 3530.503.20537.12092017 e 76459934/16], e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [422515/2016-7, 445987/2014-6, 310349/2021-4 e 465450/2014-8] por todo suporte financeiro, ao LABPETRO-UFES e à Petróleo Brasileiro S.A. (PETROBRAS) pelo fornecimento da amostra de petróleo.

REFERÊNCIAS

ADAMS, M. J. **Chemometrics in analytical spectroscopy**. Cambrigde: The Royal Society of Chemistry, 1995. ISBN: 978-0-85404-555-6.

BRERETON, R. G. **Chemometrics: Data Analysis for the Laboratory and Chemical Plant.** Ltd, John Wiley & amp; Sons; 2003. Doi: 10.1002/0470863242

CENTNER, V.; DE NOORD, O. E.; MASSAR, D. L. **Detection of nonlinearity in multivariate calibration**. Anal Chim Acta. 1998;376(2):153-168. doi:10.1016/S0003-2670(98)00543-1. Disponível em: https://www.sciencedirect.com/science/article/pii/S0003267098005431. Acesso em: 29 de Jun. de 2022.

DHANOA, M.S.; LISTER, R. S.; BARNES, R.J. **The Link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) Transformations of NIR spectra**. Journal of Near Infrared Spectroscopy 1994, 2(43). ISSN 0967-0335.

GITHUB. **Tutorial-PLS-DA-Quimiometria**. Disponível em: https://github.com/felipebachion/Tutorial-PLS-DA-Quimiometria. Acesso em: 21 de Jul. de 2022.

JOLLIFFE, J. T.; CADIMA, J. **Principal component analysis: a review and recent developments**. Phil. Trans. R. Soc. 2016, 374, 1-16. Doi: 10.1098/rsta.2015.0202. Disponível em https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202. Acesso em: 23 de Jun. de 2022.

KU, W.; STORER, R. H.; GEORGAKIS, C. **Disturbance detection**

and isolation by dynamic principal component analysis. Chemometrics and Intelligent Laboratory Systems. 1995, 30(1), 179-196. Doi: 10.1016/0169-7439(95)00076-3. Disponível em https://www.sciencedirect.com/science/article/pii/0169743995000763. Acesso em: 13 de Jul. de 2022.

FERREIRA, M. M. C. **Quimiometria – Conceitos, Métodos e Aplicações**. Ed. Unicamp, 1, 2015.

NETO, B. B.; SCARMINIO, I. S.; BRUNS, R. E.; J. **25 anos de quimiometria no Brasil**. Química Nova. 2006, 29(6), 1401-1406. Doi: 10.1590/S0100-40422006000600042. Disponível em . Acesso em: https://www.scielo.br/j/qn/a/mQNsqf68QY9TmMw3KytvdvN/?lang=pt. 15 de Jul. de 2022.

PEREIRA, P. C. S; FREITAS, C. F.; CHAVES, C. S.; ESTEVAO, B. M; PELLOSI, D. S.; TESSARO, A. L.; BATISTELA, V. R.; SCARMINIO, I. S.; CAETANO, W.; HIOKA, N. **A quimiometria nos cursos de graduação em química: proposta do uso da análise multivariada na determinação de pKa**. Quim. Nova. 2014, 37(8), 1417-1425, 2014. Doi: 10.5935/0100-4042.20140216. Disponível em https://www.scielo.br/j/qn/a/ZG9sYng77z5sywb5FMvx6Vx/abstract/?lang=en. Acesso em: 23 de Jul. de 2022.

WOLD, S.; ESBENSEN, K.; GELADI, P. **Principal Component Analysis**. Chemometrics and Intelligent Laboratory Systems, 2 (1987), pp. 37-52. Doi: 10.1016/0169-7439(87)80084-9. Disponível em https://www.sciencedirect.com/science/article/pii/0169743987800849. Acesso em: 03 de Ago. de 2022.