Aplicação da Análise por Componentes Principais (PCA) na identificação de marcas de canetas esferográficas – uma introdução à quimiometria.

Rosylane Elaine Costa Lopes[1] (IC)*, Iraci Pereira dos Santos[1] (IC), Frederico Luis Felipe Soares[1] (IC), Eduardo Ferreira Pereira[1] (PG), Jez Willian Batista Braga[1] (PQ).

***rosylane.lopes@yahoo.com.br**

[1]Laboratório de Química Analítica e Ambiental – Instituto de Química, UnB, Campus Universitário Darcy Ribeiro, Brasília-DF, CEP: 70910-900.

Palavras-Chave: Quimiometria, PCA, caneta.

RESUMO: Na Quimiometria a Análise dos Componentes Principais (PCA) é uma das ferramentas mais utilizadas, que visa principalmente à redução do número de variáveis, eliminação de dados redundantes e facilitar a interpretação dos dados. Atualmente o desenvolvimento de experimentos que introduzam as principais ferramentas dessa área em nível de graduação é de grande importância. O trabalho presente tem como objetivo apresentar um experimento para ilustrar a aplicação do PCA na identificação de padrões existentes na composição de 5 modelos de canetas azuis por espectroscopia na região do infravermelho. Os resultados mostram que depois da derivação dos espectros foi possível reduzir de 3528 variáveis para 2 componentes principais e identificar os grupos e padrões de cada caneta. O experimento ainda ilustra a importância de pré-processamentos dos dados, foi realizado com programas desenvolvidos no laboratório ou gratuitos e pode ser executado em aproximadamente 3 horas, restando tempo suficiente para discussão com os alunos.

INTRODUÇÃO

A Quimiometria é uma das áreas mais recentes da Química, que tem se mostrado de grande importância na interpretação e análise de dados obtidos pelos diversos métodos instrumentais disponíveis hoje em laboratório. Pode ser definida como sendo o desenvolvimento e a aplicação de métodos estatísticos e matemáticos no planejamento, otimização de procedimentos ou na obtenção de informações químicas através da análise de dados[1]. Essa área, iniciou-se na primeira metade da década de 70, porém só com o decorrer do desenvolvimento dos recursos computacionais e da utilização de métodos instrumentais nos laboratórios químicos a sua importância e expansão foi comprovada. Também se acredita que ela começou muito antes, junto com os trabalhos do químico Student, conhecido pelo teste analítico *t.* Com a chegada dos computadores aos laboratórios químicos, a combinação da química com a estatística começou a ganhar mais notoriedade, onde o grupo do Prof. Dr. Bruce Kowalski foi o primeiro a produzir um programa quimiométrico, chamado ARTHUR[2].

A extração de informações dos dados de um experimento normalmente envolve a análise de um considerável número de variáveis. Sendo que frequentemente apenas um pequeno o número destas variáveis apresentam maior importância, resultando em um grande conjunto de dados que podem ser redundantes ou que não apresentem relevância para o objetivo do experimento. A análise de componentes principais (PCA, do inglês *Principal Component Analysis*) é um dos principais métodos utilizados em quimiometria, onde seu objetivo é reduzir o número de dimensões do

conjunto de dados sem a perda das informações relevantes, de modo a se obter um número menor de novas variáveis (componentes principais) que facilite a interpretação dos dados. O PCA pode propiciar, através de gráficos, a identificação da existência de padrões de similaridade existentes em dados de um conjunto das amostras analisadas.

Mesmo com sua importância e aplicabilidade ainda são poucas as universidades que possuem disciplinas de Quimiometria nos currículos de graduação, além de ser relativamente pequeno o número de experimentos que ilustram métodos quimiométricos para os alunos de graduação. Tendo em vista essa deficiência, o presente trabalho tem como objetivo apresentar um experimento para ilustrar e introduzir a alunos de graduação o método de análise de componentes principais (PCA) para a análise multivariada de dados. Para esse fim, é apresentada uma aplicação voltada para a identificação de tintas de canetas esferográficas de cor azul utilizando espectroscopia na região do infravermelho médio e medidas por reflectância total atenuada (ATR, do inglês Attenuated Total Reflectance).

ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

O PCA é uma ferramenta quimiométrica que, a partir de um determinado conjunto de dados, é capaz de condensar as informações mais relevantes em um número reduzido de novas variáveis[3]. Esse método está baseado transformação das variáveis originais de uma matriz dados, onde as linhas representam as amostras e as colunas as variáveis, em novas variáveis não correlacionadas, chamadas componentes principais (PC, do inglês *Principal Components*), que são combinações lineares das variáveis originais. Quando o número de PC é significativamente menor que o número de variáveis inicial obtém-se uma redução substancial de informação, proporcionando uma melhor visualização do conjunto de dados através das PC. Sendo assim este método pode ser utilizado na redução de informações, para reconhecimento de padrões, na seleção de amostras, na construção de modelos para calibração multivariada, entre diversas outras aplicações.

O cálculo utilizado na PCA baseia-se na decomposição de uma matriz qualquer "X" em um produto de duas matrizes menores **T** e **P**, conforme expresso pela equação 1.

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_d\mathbf{p}_d^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad (1)$$

Sendo **X** a matriz original com *n* linhas e *m* colunas; **T** a matriz escores com *n* linhas e *d* colunas (número de PC escolhido), $\mathbf{P}^T$ a transposta da matriz de pesos com *m* colunas e *d* linhas e **E** a matriz de resíduos que contém a fração da informação não é modelado/explicado pelas PC. A matriz de pesos é onde se encontra a relevância das variáveis originais em cada PC, onde cada elemento de **P** é matematicamente igual ao cosseno do ângulo entre o eixo da cada variável original e a PC. A matriz de escores representa a disposição das amostras no espaço das PC, isto é, a projeção dos pontos experimentais nos eixos definidos pelas PC.

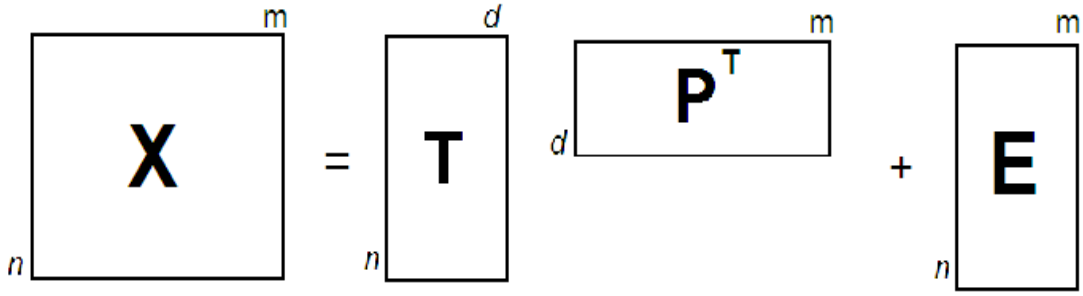

Figura 1: Representação esquemática da decomposição de uma matriz X em PCA.

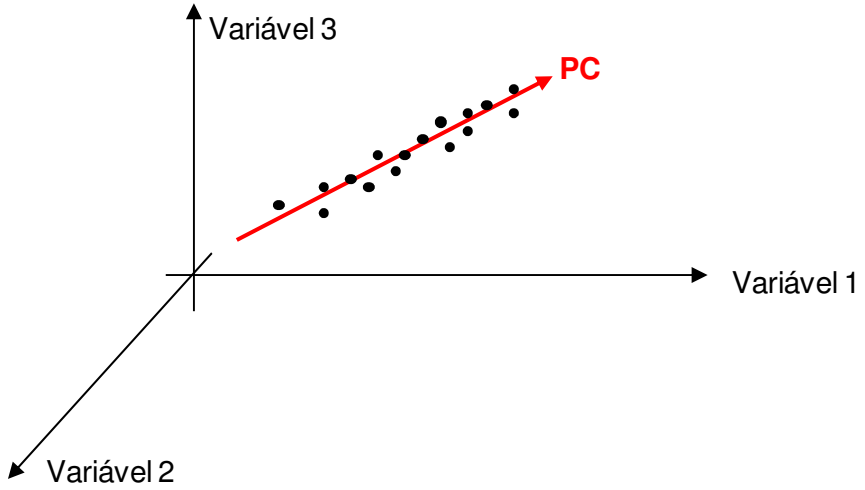

Figura 2: Exemplificação da redução de um espaço de três dimensões para um de uma dimensão através da utilização de PCA.

A princípio, espera-se que grande parte da variância (informação) dos dados seja explicada por um número pequeno de PC[4]. A percentagem de variância explicada pela PC *d* (%VE) pode ser calculada a partir da equação 2:

$$\%VE = \frac{1 - \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i,j} - t_{i,d}p_{i,d}^{T})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i,j})^2} x100 \qquad (2)$$

Onde o numerador representa a variância explicada pela PC *d* e o denominador a variância total dos dados.

A maneira mais comum de se representar graficamente o resultado da decomposição em PCA é plotar, entre si em um gráfico, os escores e pesos das componentes principais escolhidas, na maioria das vezes criando um gráfico bi ou tridimensional que permite observar com uma maior clareza a disposição das amostras e a participação das variáveis naquele padrão observado e identificação dos possíveis agrupamentos presentes nos dados[5,6].

As distâncias entre os escores podem ser calculadas baseadas na posição espacial destes, possibilitando um estudo quanto à existência similaridade ou não entre as amostras. Assim ocorrerá a formação de um agrupamento de dados pelas amostras que apresentam maior similaridade (pontos que se encontram mais próximos). Caso exista uma amostra com características distintas do grupo ocorrerá distanciamento do ponto referente a essa amostra em relação ao grupo, o que faz a PCA também uma importante ferramenta para identificar amostras anômalas.

ESPECTROSCOPIA NA REGIÃO DO INFRAVERMELHO

A região do espectro infravermelho se estende desde a extremidade final da região do visível (cerca de 13333 cm^{-1}) até o início da região de microondas (cerca de 33 cm^{-1}). Porém, uma das regiões mais úteis de trabalho se encontra na região de 4000 a 400 cm^{-1}, uma vez que esta apresenta bandas características de grande parte dos grupos funcionais orgânicos.

A espectroscopia na região do infravermelho envolve as transições de ordem vibracional e rotacional em moléculas que apresentam variação de momento dipolar em consequência dos movimentos vibracional e rotacional. A radiação eletromagnética é formada por duas componentes, um elétrico e outro magnético, que se propagam perpendicularmente. Quando ocorre ama variação na distância entre dois núcleos atômicos, como ocorre durante uma vibração em uma ligação química por causa da movimentação das cargas, fica estabelecido um campo elétrico. Este campo elétrico formado tem capacidade de interagir com o campo elétrico da radiação eletromagnética. Quando a frequência natural de vibração da molécula se iguala à frequência da radiação incidida, ocorre uma transferência de energia da radiação para a molécula causando uma mudança amplitude da vibração molecular.

Os espectros de infravermelho gerados são bastante complexos com inúmeros picos e vales e são muito utilizados para fins de comparação. Por isso as principais aplicações da espectroscopia na região do infravermelho envolvem a identificação e a elucidação de estruturas de moléculas desconhecidas, além de encontrar atualmente uma forte aplicação na analise quantitativa com a utilização de métodos quimiométricos[7].

PARTE EXPERIMENTAL

Descrição das amostras:

Para a realização do experimento foram utilizados cinco modelos de canetas esferográficas de quatro marcas distintas com a tinta de cor azul, sendo elas: BIC, Faber-Castell, CIS Pro, CIS Glycer e Compactor.

Equipamentos e programas utilizados:

Para aquisição dos espectros na região do Infravermelho foi utilizado um espectrômetro de infravermelho com transformada de Fourier da marca Jasco modelo 4100 com acessório para medidas por Reflectância Total Atenuada (ATR).

Para os cálculos foi utilizado o programa Octave versão 3.0.2 instalado em um computador pessoal com processador core2 duo 2.0 Ghz, 2 Gb de memória RAM e com sistema operacional Windows Vista. O Octave foi empregado para o desenvolvimento de três programas utilizados para o cálculo da PCA, derivada dos espectros e cálculo da % variância explicada pelo modelo. O Octave foi escolhido por ser um software livre de linguagem de programação de alto nível, utilizado principalmente para cálculos numéricos. Os cálculos são realizados através do prompt comando de Octave, janela semelhante ao MSDOS, a qual é conveniente para resolver tanto problemas lineares quanto não-lineares e executar experimentos numéricos em um ambiente relativamente fácil[8].

Procedimento experimental:

Com cada uma das canetas foi desenhado um retângulo totalmente preenchido em uma tira papel, recortada de folha branca modelo A4. Após feito o desenho aguardou-se por cerca de 10 minutos para que a tinta secasse e não houvesse a contaminação do cristal do acessório de ATR. As tiras de papel foram posicionadas com o retângulo sobre o cristal de seleneto de zinco (ZnSe) e os espectros de infravermelho registrados no modo de absorbância. Como referencia do equipamento (Background) foi utilizada a leitura do espectro do ar. Os espectros foram registrados como uma média de 32 varreduras e resolução de 4 cm^{-1}.

Os dados obtidos de cada espectro foram exportados e, em seguida, organizados em uma planilha onde constavam os números de onda, a identificação das amostras e os valores de absorbância em cada número de onda para cada amostra. Essa planilha foi importada para o programa Octave e realizados os cálculos.

RESULTADO E DISCUSSÃO

Partindo dos espectros foi gerada uma matriz de 50 colunas por 3528 linhas perfazendo um total de 176400 valores. Na figura 3 podem-se observar os espectros das 50 amostras em um único gráfico, onde se observa que apenas a marca Bic pode ser diferenciada visualmente por apresentar bandas bem distintas das demais marcas e modelos. Logo, é necessária a aplicação de PCA para verificar a possibilidade de uma visualização mais fácil dos dados e tentar discriminar as marcas e modelos de caneta que possuem espectros muito semelhantes.

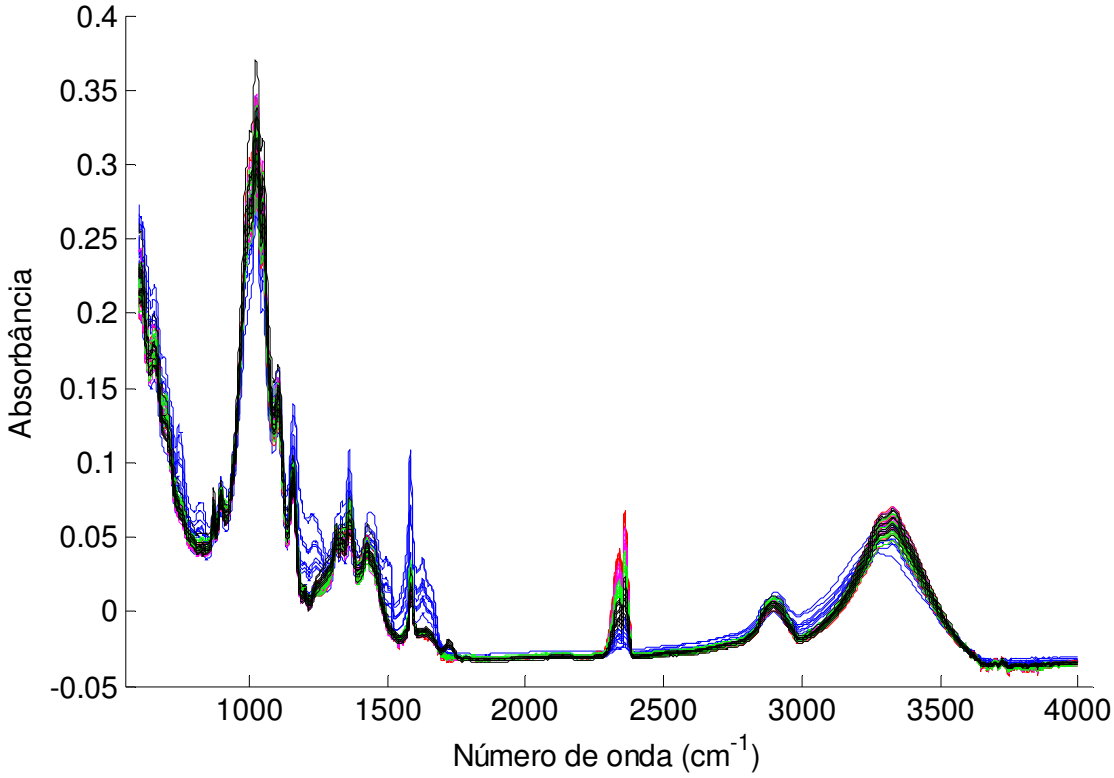

Figura 3: Espectros de FT-IR/ATR das 50 amostras dos 5 modelos de canetas. (azul) Bic, (vermelho) Cys-Glycer, (rosa) Cys-Pro, (verde) Compactor e (preto) Faber Castel.

Na tabela 1 são apresentados os resultados da decomposição em PCA dos dados da figura 3. A primeira coluna indica o número da PC, a segunda indica a porcentagem da variância explicada por aquela PC e na terceira tem-se a porcentagem de variância acumulada (a soma da porcentagem de variância explicada daquela PC e das anteriores). Observa-se que a matriz de dados pode ser reduzida a uma 50x3, mantendo-se cerca de 95% das informações, ou seja, com o auxilio do PCA os dados que estavam num espaço multidimensional (3528 dimensões) foram convertidos para um espaço de três dimensões, facilitando a interpretação dos dados, uma vez que agora se apresentam na forma de um gráfico tridimensional. Considerando que existe uma variação de linha de base nos espectros e o ruído instrumental inerente do equipamento, pode-se esperar que 95 % da informação representa praticamente toda a informação relevante dos dados, sendo os 5 % das outras componentes possivelmente variação aleatória sem importância.

Tabela 1: Variância explicada pelas PCs obtidas pela decomposição dos dados originais.

Número de PC	Variância de cada PC (%)	Variância acumulada (%)
1	54,97	54,97
2	32,62	87,59
3	7,58	95,17
4	1,49	96,66

5	1,03	97,69
6	0,68	98,37
7	0,49	98,86
8	0,36	99,21
9	0,19	99,40
10	0,16	99,57

Com o gráfico de escores apresentado na figura 4 é possível se observar que apenas a marca Bic se mostra separada das demais e que as outras marcas ainda se apresentam de forma agrupada. Esta diferenciação pode ser explicada pelas bandas distintas que essa marca possui em seu espectro de infravermelho, como ressaltado anteriormente.

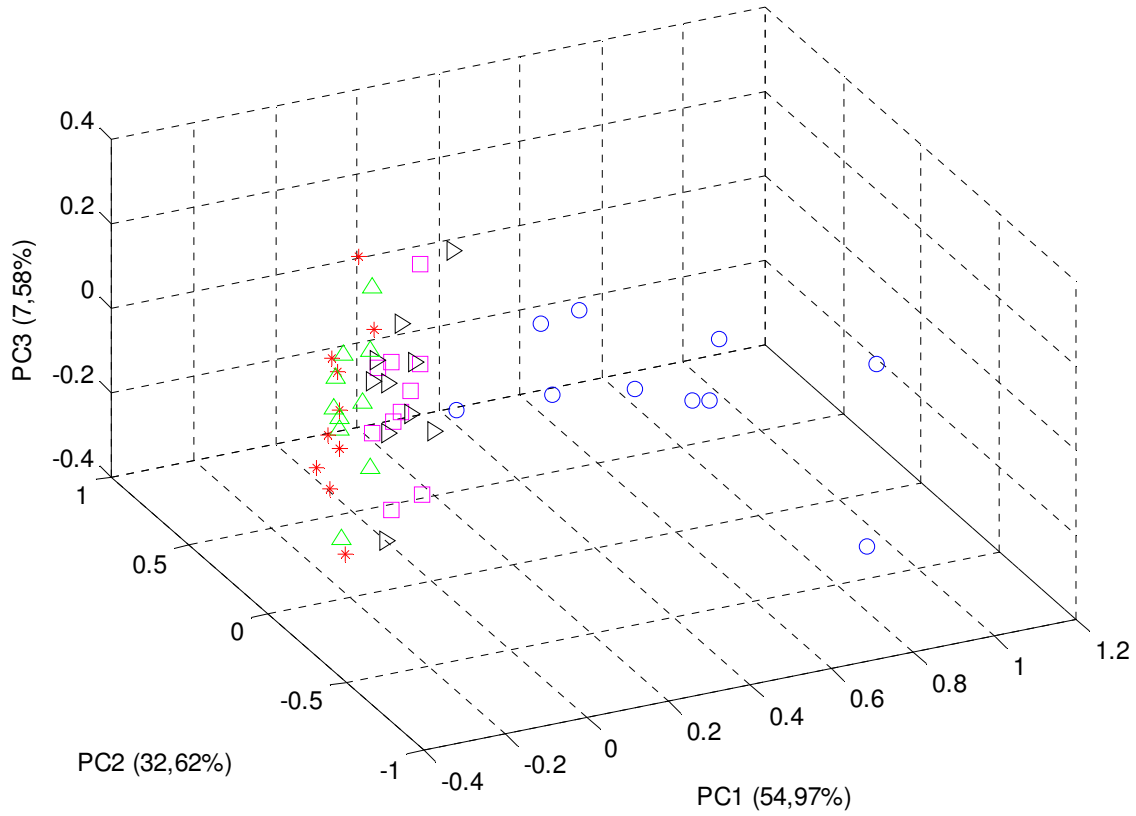

Figura 4: Plotagem dos escores PC1xPC2xPC3 - Bic (o), Faber-Castell (Δ), Compactor (□), Cis Pro (Δ) Cis glycer (*).

Com a finalidade de se corrigir o efeito da linha de base dos espectros e verificar o efeito de um pré-processamento na análise dos dados, foi realizad0 cálculo da primeira derivação dos espectros (figura 5). Observa-se na figura 5 que a derivada ajustou a linha de base e, apesar de pouco visível na figura 5, também tornou mais nítido os sinais característicos de cada marca, mas também causa um aumento na intensidade de ruídos. A partir destes novos dados fez-se um novo cálculo de PCA, cujos resultados são apresentados na tabela 2.

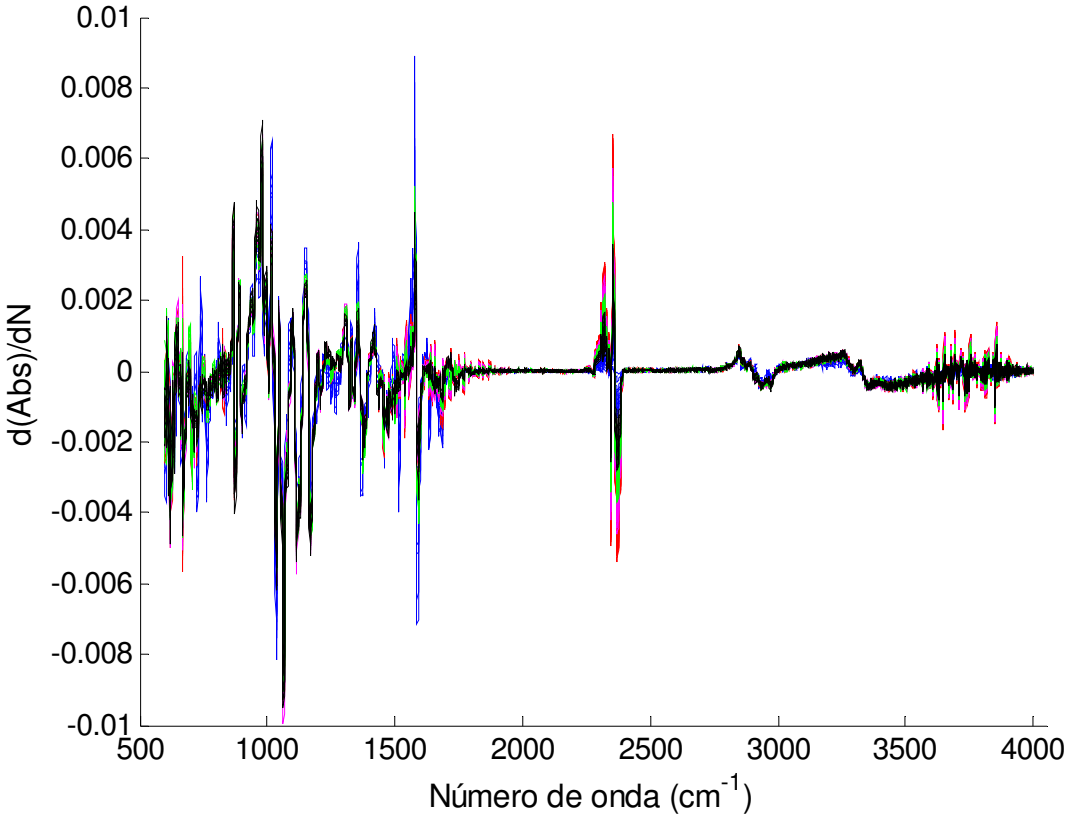

Figura 5: Primeira derivada dos espectros de infravermelho das 50 amostras de tinta de caneta.

Tabela 2: Variâncias explicada pelas PCs obtidas da decomposição dos dados derivados.

Número de PC	Variância de cada PC (%)	Variância acumulada (%)
1	55,77	55,77
2	16,34	72,11
3	6,98	79,09
4	3,99	83,08
5	1,880	84,96
6	1,70	86,73
7	1,49	88,22
8	1,29	89,51
9	0,99	90,49
10	0,89	91,39

Percebe-se que há uma redução da variância explicada pelas primeiras componentes, a partir dos dados derivados observa-se que as duas primeiras PC agora explicam apenas 72,11% da variância. Entretanto, na figura 6 que apresenta o gráfico de escores de PC1 versus PC2 observa-se que um melhor agrupamento de cada modelo de tinta de caneta e distinção de cada modelo entre si. Logo, observa-se

que mesmo com apenas 72 % da informação já temos o suficiente para observar um agrupamento de amostras, o que é explicado pela intensificação de alguns sinais característicos de cada modelo de tinta de caneta no espectro derivado. A figura 6 mostra que a PC1 é responsável pela separação da tinta da caneta Bic das demais, uma vez que todas as amostras que apresentam escores com valores abaixo de 0,01 nessa componente são dessa marca. Observa-se ainda que foi obtida uma boa separação de todas as quatro marcas de caneta, as quais são identificadas por elipses. Pode-se ainda observar uma subdivisão que separa as canetas da marca Cis (Pró e Glycer) mostrando que os espectros de IR e a análise PCA torna possível identificar e discriminar as diferentes modelos de uma mesma marca de caneta.

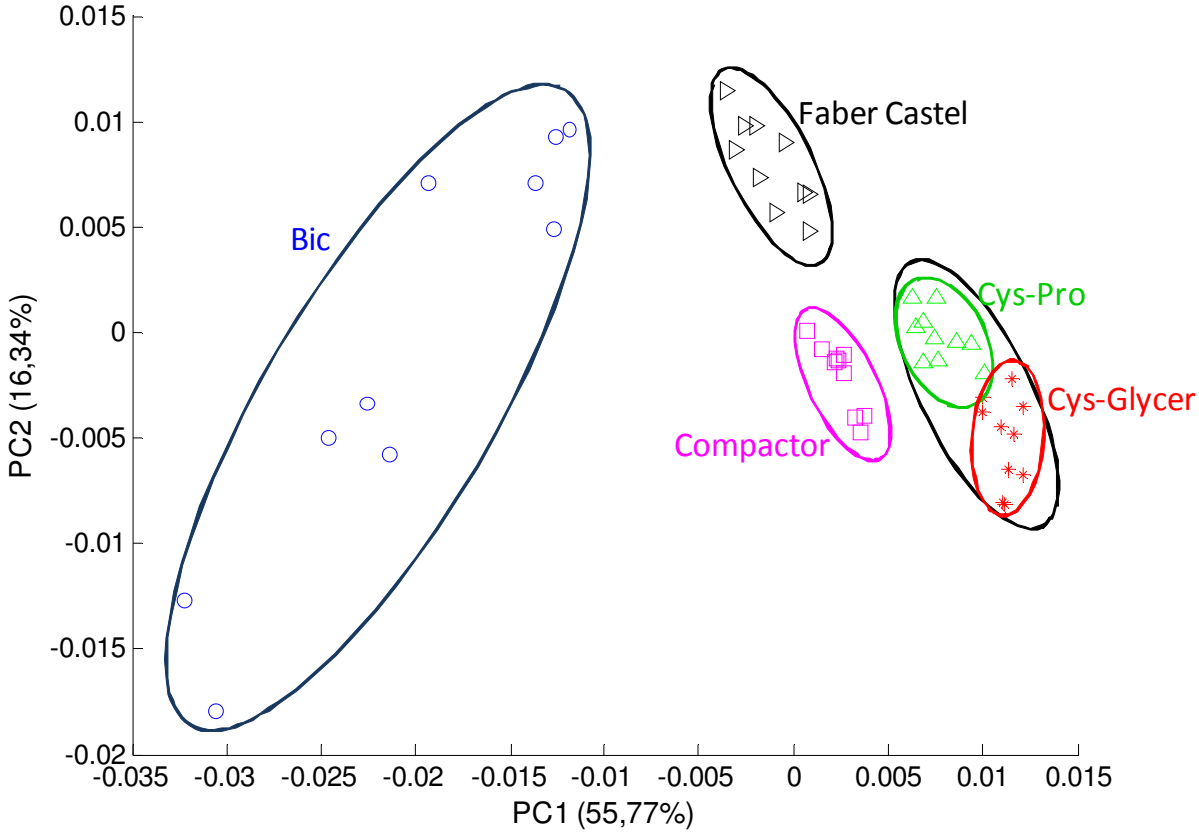

Figura 6: Gráfico PC1xPC2 para os dados obtidos dos espectros derivados. Bic (o), Faber-Castell (Δ), Compactor (□), Cis Pro (Δ) Cis Glycer (*).

O quadro 1 mostra a distribuição do tempo estimado para cada atividade proposta para a realização do experimento, onde se observa que o experimento pode ser executado em uma aula experimental de quatro horas. É importante observar que entre as atividades propostas estão reservados 30 minutos no início do experimento para introdução à PCA e à técnica de espectroscopia no infravermelho e 40 minutos no final para a discussão dos resultados com os alunos.

Quadro 1: Distribuição das atividades propostas para o experimento e o tempo estimado para cada atividade.

Atividades propostas	Tempo estimado
Explicação inicial do experimento	30 minutos
Preparo das amostras	30 minutos
Configuração do equipamento	15 minutos
Aquisição dos espectros	60 minutos
Exportação e organização dos espectros	20 minutos
Realização dos cálculos e figuras no Octave	45 minutos
Interpretação e discussão dos dados	40 minutos
Total	4 horas

Conclusões

O experimento proposto contribui para a inserção da quimiometria nos cursos de graduação. Além da introdução da PCA aos alunos é realizado um experimento que apresenta simplicidade no preparo das amostras, análise rápida, ressalta a importância do pré-processamento dos dados, utiliza um software livre e subprogramas desenvolvidos no laboratório e não produzir resíduos para serem descartados. Esse, ainda, ilustra a aplicação do PCA em dados espectroscópicos obtidos a partir da espectroscopia de absorção na região do infravermelho enfocando um estudo de caso com relevância em química forense, na qual a identificação da marca de tintas utilizada em uma assinatura ou outras finalidades muitas vezes é de grande importância.

O experimento proposto apresenta resultados que possibilita sua aplicação em cursos de graduação, esse pode ser realizado em uma aula experimental de aproximadamente quatro horas, permitindo uma ampla discussão do professor com os alunos a cerca dos conteúdos abordados. Comprova que a PCA é uma ferramenta importante na análise multivariada dos dados obtidos. Os resultados demonstraram que a PCA foi capaz possibilitar a separação e identificação de tintas azuis de quatro marcas diferentes de caneta e de dois modelos dentro de uma mesma marca.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Vogel, A.I., et al. **Análise Química Quantitativa**. 6. ed. Rio de Janeiro: LTC Editora, 2002. p. 77.

2. Neto, B.B; Scarminio, I.S.; Bruns, R.E. **25 Anos de Quimiometria no Brasil.** Química Nova, v. 29, n. 6, p. 1401-1406, 2006.

3. Matos, G.D., et al. **Análise Exploratória em Química Analítica com Emprego de Quimiometria: PCA E PCA de Imagens**. Revista Analytica, n. 6, p. 38-50, Ago./Set. 2003.

4. Otto, M., **Chemometrics Statistics and Computer Application in Analytical Chemistry**. 2. ed. Darmstadt: WILEY-VCH, 2007. p. 127.

5. Breton, R. G., **Chemometrics: Data Analysis for the Laboratory and Chemical Plant.** 2. ed., John Wiley & Sons, Ltd., 2003. p. 187.

6. SOUSA, R. A., et. al. **Classificação de água-de-coco processada e natural por meio de HCA, PCA e teores de íons metálicos determinados por ICP OES.** Química Nova, v. 29, n. 4, p. 654-656, 2006.

7. Ohlweiler, O.A., **Fundamentos de Análise Instrumental**, ed. Livros Técnicos e Científicos ed. S.A., 1981, p. 111.

8. Wikipédia. **GNU Octave**. Disponível em: <http://pt.wikipedia.org/wiki/GNU_Octave>. Acesso em: 23 abril 2010.