Educação

CLASSIFICAÇÃO PERIÓDICA: UM EXEMPLO DIDÁTICO PARA ENSINAR ANÁLISE DE COMPONENTES PRINCIPAIS

Wellington da Silva Lyra, Edvan Cirino da Silva, Mario Cesar Ugulino de Araújo e Wallace Duarte Fragoso*
Departamento de Química, Universidade Federal da Paraíba, 58051-970 João Pessoa – PB, Brasil
Germano Veras
Departamento de Química, Universidade Estadual da Paraíba, 58109-753 Campina Grande – PB, Brasil

Recebido em 16/9/09; aceito em 3/4/10; publicado na web em 20/7/10

PERIODIC CLASSIFICATION: A DIDACTIC EXAMPLE TO TEACH PRINCIPAL COMPONENT ANALYSIS. A dataset of chemical properties of the elements is used herein to introduce principal components analysis (PCA). The focus in this article is to verify the classification of the elements within the periodic table. The reclassification of the semimetals as metals or nonmetals emerges naturally from PCA and agrees with the current SBQ/IUPAC periodic table. Dataset construction, basic preprocessing, loading and score plots, and interpretation have been emphasized. This activity can be carried out even when students with distinct levels of formation are together in the same learning environment.

Keywords: principal components analysis; semimetals; chemical properties of the elements.

INTRODUÇÃO

A análise de componentes principais (PCA) encontra-se certamente entre as mais importantes ferramentas da análise multivariada, inclusive por constituir a base onde se fundamentam a maioria dos outros métodos multivariados de análise de dados. Como uma ferramenta de análise exploratória a PCA permite revelar a existência ou não de amostras anômalas, de relações entre as variáveis medidas e de relações ou agrupamentos entre amostras. Além disto, métodos eficientes de classificação, como a modelagem independente para analogia de classes (SIMCA) e de calibração, como a regressão em componentes principais (PCR) ou a regressão por mínimos quadrados parciais (PLS), são derivados da PCA.

No contexto da aplicação em problemas químicos estes métodos estatísticos são denominados métodos quimiométricos. A importância da quimiometria para os laboratórios modernos de química cresceu com a capacidade dos instrumentos analíticos de produzirem conjuntos de dados cada vez maiores e mais complexos e com a evolução dos computadores que permitem tratá-los agilmente. Assim, aplicações de PCA em problemas de química são cada vez mais comuns e podem ser encontradas facilmente na literatura.[1-10]

Para o aprendizado desta ferramenta estão disponíveis vários livros[11,12] e artigos, [13-15] tanto para quem pretende apenas usar a PCA como para quem pretende se aprofundar na álgebra e nos algoritmos empregados. Um problema recorrente que não pode ser negligenciado é o risco que os usuários correm de perder o sentido químico de seus estudos, muitas vezes preocupados apenas com as tabelas e os gráficos de excelente qualidade produzidos pelos programas. Este risco é minimizado quando o sistema estudado é bem conhecido e as interpretações dos resultados da PCA são feitas fundamentadas neste conhecimento prévio.

Nos cursos de quimiometria comumente enfrentam-se dificuldades para encontrar uma aplicação para a análise de componentes principais que seja suficientemente didática para um primeiro contato. Aplicações envolvendo os mais diversos sistemas[1-10] podem ser empregadas, mas estes sempre serão dominados por alguns alunos e não por outros. Assim, um banco de dados oriundo de espectroscopia é um bom exemplo para um estudante já adiantado em química analítica, mas torna-se incompreensível para um estudante que ainda não conhece a técnica. Cursos de quimiometria costumam ter um público com interesses bastante diversificados e em diferentes níveis de formação, o que dificulta ainda mais encontrar um exemplo ideal.

Uma proposta que surgiu em um de nossos cursos é a análise da classificação dos elementos tradicionalmente conhecidos como semimetais na Tabela Periódica, como metais ou ametais, de acordo com a recomendação da IUPAC e adotada na Tabela Periódica da SBQ desde 2001. Nessa nova classificação, germânio, antimônio e polônio são metais, enquanto boro, silício, arsênio e telúrio são ametais. Por se tratar de uma recomendação recente, muitos estudantes ainda trazem do ensino médio a noção do grupo dos semimetais, agora extinto. Assim pode-se simultaneamente introduzir esta atualização da Tabela Periódica e aplicar a análise de componentes principais a uma série de propriedades dos elementos e verificar a consistência das atribuições dos elementos aos grupos como proposto. No presente trabalho, este exemplo de aplicação da PCA é apresentado e proposto para a utilização em cursos de quimiometria, uma vez que tanto as propriedades empregadas quanto a caracterização do problema em si são facilmente reconhecidas por quaisquer estudantes de química, mesmo que iniciantes, de modo que as discussões e a conclusão da análise sejam facilmente compreendidas por todos.

O BANCO DE DADOS

Os elementos que constituem o banco de dados são os naturais, ou seja, aqueles que vão do H até o U (com exceção de Tc, Fr, Pm e At), um total de 88 elementos. Os elementos artificiais não são incluídos por não possuírem valores disponíveis para todas as propriedades consideradas.

Para realização da PCA foram utilizados valores tabelados[16] das seguintes propriedades: primeira energia de ionização (EI), raio atômico (RA), afinidade eletrônica (AE), eletronegatividade de Pauling (EN), densidade (D), calor específico (CE), entropia padrão (Sº) e condutividade térmica (CT).

Os alunos são orientados a construir uma tabela de dados (matriz) pesquisando em um *Handbook*[16] os valores das propriedades para cada elemento. Os elementos (objetos) são dispostos nas linhas da matriz

*e-mail: wallace@quimica.ufpb.br

e as propriedades (variáveis) nas colunas. É importante que os alunos construam o banco de dados para se familiarizar com a estrutura da matriz e com o significado das linhas e colunas.

Pré-processamento

A matriz de dados construída usando os valores das propriedades dos elementos apresenta variáveis com significados físicos, magnitudes e unidades distintas. Fica fácil mostrar para os estudantes que estes valores como estão não podem ser combinados.

Os dados precisam ser pré-processados e o pré-processamento adequado é o autoescalonamento, ou seja, a matriz é centrada na média dos valores subtraindo-se o valor de cada elemento da matriz da média de cada variável (coluna) e, em seguida, é normalizada pelo desvio padrão dividindo-se o valor de cada elemento centrado na média pelo desvio padrão da variável. O propósito dessa transformação é permitir que todas as variáveis possam exercer influências equitativas nos resultados além de torná-las adimensionais.

Análise de componentes principais – PCA

A realização da PCA consiste em fatorar a matriz de dados **X**, de modo que $\mathbf{X}=\mathbf{TL}^T+\mathbf{E}$, onde **L** é a matriz dos pesos, **T** a matriz dos escores e **E** a matriz dos resíduos. O símbolo ^{T} (T sobrescrito) é o operador de transposição de matriz. A primeira componente principal é $\mathbf{PC1}=\mathbf{t}_1\mathbf{l}_1^T$, que é a melhor aproximação de posto 1 para **X** e corresponde à direção de maior variância no espaço multivariado. $\mathbf{E}_1=\mathbf{X}-\mathbf{t}_1\mathbf{l}_1^T$ é o resíduo de **X**, descontado **PC1**. A segunda componente principal é $\mathbf{PC2}=\mathbf{t}_2\mathbf{l}_2^T$, que é a melhor aproximação de posto 1 para $\mathbf{E}_1$ e corresponde à direção de maior variância no espaço multivariado não modelada por **PC1**, ou seja, ortogonal a ela. $\mathbf{E}_2=\mathbf{E}_1-\mathbf{t}_2\mathbf{l}_2^T$ é o resíduo deixado por **PC1** e **PC2**. As componentes subsequentes modelam sempre a direção de maior variância no espaço multidimensional não modelado pelas PCs anteriores e são sempre ortogonais a todas elas. É possível realizar uma truncagem na sequência das componentes principais mantendo apenas um número pequeno de PCs que já respondem por uma parcela significativa da informação total contida na estrutura de dados.

Na prática, para fazer uma análise de componentes principais, calculamos inicialmente a matriz de covariância, **C**, para dados centrados na média, ou a matriz de correlação, **R**, para dados autoescalonados.

$$\mathbf{C}\ (\text{ou}\ \mathbf{R}) = \mathbf{E}_0^T\mathbf{E}_0/(n-1)$$

onde $\mathbf{E}_0$ é a matriz pré-processada e n é o número de linhas da matriz. Em seguida calculamos os autovalores e autovetores normalizados de **C** ou **R**.

$$\mathbf{CL}=\Lambda\mathbf{L}$$

onde Λ é a matriz diagonal dos autovalores. Cada autovetor **l** é um vetor de pesos de uma componente principal. Cada autovalor λ fornece a quantidade de variância explicada pela respectiva componente, de modo que PC1 tem o maior autovalor, PC2 o segundo maior, e assim por diante. Outra operação algébrica, a decomposição em valores singulares, e um algoritmo numérico, o NIPALS, também podem ser usados para executar uma análise de componentes principais. Para uma discussão mais detalhada destes métodos sugerimos o artigo de Ferreira e colaboradores.[15]

Programas computacionais

Hoje em dia é grande a variedade de programas voltados para a análise multivariada e a análise de componentes principais está presente em todos eles. Dentre os programas comerciais podemos citar o *Unscrambler®*,[17] *o Statistica®*,[18] *o Minitab®*,[19] *e o Pirouette®*.[20] Também não é difícil encontrar, ou mesmo escrever, rotinas para PCA para o *Matlab®*[21] e para o programa livre GNU *Octave®*.[22] O professor poderá utilizar qualquer um destes ou outro que tenha disponível para ministrar esta aula. Em qualquer um deles os gráficos que serão apresentados aqui e, certamente, muitos outros poderão ser construídos.

RESULTADOS E DISCUSSÃO

Sob o ponto de vista formal, fazer uma análise de componentes principais é realizar uma mudança da base do espaço vetorial do conjunto de dados. Cada objeto (no nosso caso cada elemento) que era então representado num espaço N-dimensional definido pelas N variáveis (no nosso caso o espaço das 8 propriedades), passa a ser representado por N componentes principais. A Tabela 1 mostra as variâncias explicadas e cumulativas para cada uma das 8 componentes principais. Como podemos observar, as primeiras componentes respondem pela maior parte da variância. Então podemos simplificar a análise truncando a base em um número de variáveis menor que N, sem perda significativa de informação. Algumas vezes não precisamos nos preocupar em recuperar uma quantidade elevada da informação, mas sim recuperar a parte da informação relevante ao problema que estamos estudando.

Tabela 1. Variância explicada e cumulativa ao longo das componentes principais

PC	Variância explicada (%)	Variância cumulativa (%)
1	33,897	33,897
2	24,237	58,134
3	13,321	71,455
4	9,545	81,000
5	8,542	89,542
6	5,667	95,209
7	3,141	98,350
8	1,650	100,000

A Figura 1 mostra o gráfico de pesos para as duas primeiras componentes principais. Geometricamente, os pesos correspondem aos cossenos dos ângulos que as componentes principais fazem com as variáveis originais. São os pesos das variáveis originais na combinação linear que definem cada Componente Principal.

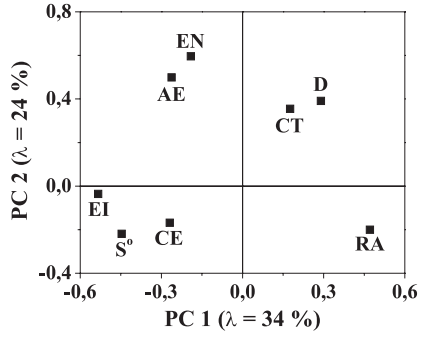

Figura 1. *Gráfico dos pesos de PC1 e PC2 para oito propriedades: primeira energia de ionização (EI), raio atômico (RA), afinidade eletrônica (AE), eletronegatividade de Pauling (EN), densidade (D), calor específico (CE), entropia (Sº) e condutividade térmica (CT)*

No gráfico dos pesos observamos a relação entre as variáveis. Com base nestas relações podemos inicialmente tentar inferir algu-

ma interpretação física para as componentes principais. Na Figura 1, é interessante notar a disposição das variáveis ao longo de PC1, que modela 34% da variância da matriz de dados. O raio atômico (RA) tem sinal contrário ao da eletronegatividade (EN), da afinidade eletrônica (AE) e da primeira energia de ionização (EI). Isto está condizente com a variação destas propriedades periódicas, já que o RA varia exatamente no sentido inverso das outras três propriedades, tanto nos períodos quanto nas famílias. A EI apresenta o maior peso, contribuindo mais para a PC1, refletindo a maior variabilidade desta propriedade ao longo dos elementos, quando comparada à EN e AE. Assim, temos razões para acreditar que a primeira componente principal (**PC1**) modela o comportamento periódico dos elementos. Adicionalmente S°, que não é uma propriedade periódica, mas uma propriedade termodinâmica, tem um peso (negativo) bastante elevado. Como os metais têm entropia em geral mais baixa que os não metais, o valor de S° também ajudará a discriminá-los.

Neste ponto o professor pode chamar atenção para o sinal dos pesos das PCs. É importante notar que a Figura 1 obtida em aula pode ser um pouco diferente desta porque os pesos podem ter todos os sinais trocados para cada componente principal. Isto ocorre porque os pesos são autovetores normalizados da matriz de correlação (ou da matriz de covariância se os dados são apenas centrados na média) e se todos os sinais do vetor de peso forem trocados, ainda teremos um autovetor normalizado da matriz de correlação. A interpretação do gráfico de pesos não muda, uma vez que é a posição relativa dos pesos no gráfico que tem significado físico.

A Figura 2 mostra o gráfico dos escores para as duas primeiras componentes principais e a ampliação da região onde se vê os tradicionais semimetais. Os escores são as projeções dos objetos originais no espaço das componentes principais, ou seja, são as novas coordenadas dos objetos nas novas variáveis que são as PCs. É no gráfico dos escores que procuramos alguma relação entre os objetos, no caso do nosso conjunto de dados as relações entre os elementos. Uma tarefa importante quando da construção de um gráfico de escores é propor um sistema de rotulagem dos objetos, pois muitas vezes apenas com uma rotulagem adequada certos padrões poderão aparecer. No caso da Figura 2, para o problema em questão, a rotulagem escolhida foi destacar os elementos em função das suas classes: metais, semimetais, ametais, gases nobres e hidrogênio.

Conforme pode ser observado na Figura 2a, os elementos de uma mesma classe tendem a agrupar-se, ocupando a mesma região no gráfico dos escores. Enquanto os ametais estão mais à esquerda os metais se encontram à direita. Os outrora ditos semimetais encontram-se entre essas duas classes. Agora podemos atribuí-los a uma ou outra classe. Se a primeira componente principal de fato modela o comportamento periódico como acreditamos, então os elementos mais à direita devem ser reclassificados como metais e os elementos mais à esquerda como ametais (isto para os pesos em **PC1** exatamente com os mesmos sinais mostrados na Figura 1. Se os seus pesos têm os sinais trocados o gráfico de escores também os terá).

A análise de componentes principais é uma técnica de reconhecimento de padrões e não uma técnica de classificação. Ela apenas ilustra a relação entre os elementos no gráfico de escores, mas não dirá em absoluto como classificá-los. O telúrio é o elemento mais à esquerda, como pode ser visto na Figura 2b, ou seja, é o mais próximo dos ametais. Se algum elemento deve ser classificado como ametal, então seguramente este é o telúrio. Da mesma forma, o polônio é o elemento mais à direita (Figura 2b) e pode ser classificado como um metal, pois está mais próximo destes. Por outro lado, a PCA não irá nos dizer até onde vai a classe dos ametais e a partir de onde se inicia a dos metais. Esta decisão precisará de critérios químicos para ser tomada. No entanto, o resultado da PCA mostra-se consistente com a decisão da IUPAC, uma vez que

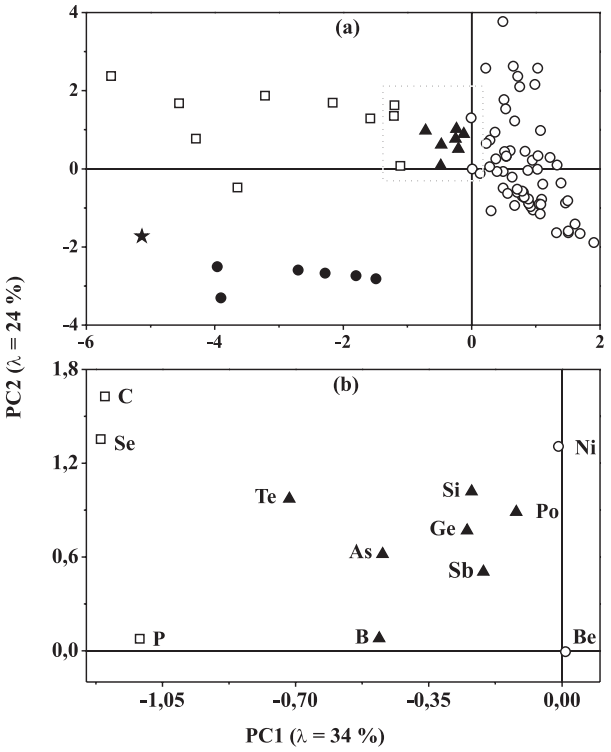

Figura 2. *(a) Gráfico dos escores de PC1 e PC2 para 88 elementos químicos: ○ metais, □ ametais, ▲ semimetais, ● gases nobres e ★hidrogênio. (b) Ampliação da região onde se vê os semimetais*

arsênio e boro, ambos classificados como ametais, são os próximos elementos mais à esquerda, e que o antimônio, classificado como metal, é o segundo elemento mais à direita entre os até então semimetais. Silício (ametal) e germânio (metal) são praticamente coincidentes e com esta análise de componentes principais não temos resolução para identificar o silício como ametal. De fato, dadas as suas propriedades e semelhanças com os metais e apesar da IUPAC ter preferido classificá-lo como ametal, não é incomum o silício ser tratado como metal, sobretudo na indústria, na física e eletrônica. Cabe ao professor chamar a atenção dos alunos que este padrão observado é o máximo que podemos conseguir para estas variáveis escolhidas. Se as variáveis fossem outras, teríamos um padrão diferente, que talvez concordasse melhor com as atribuições da IUPAC. De qualquer forma, o resultado obtido é bastante satisfatório e fácil de ser trabalhado e interpretado em aula. Além disso, é importante expor não só as potencialidades da técnica, mas também seus limites.

Note que apenas a **PC1** foi necessária para que se pudesse agrupar os elementos que seriam reclassificados como metais ou ametais. Esta componente, no entanto, modela apenas 34% da informação associada às variáveis originais. A segunda componente principal modela mais 24% da informação e também pode guardar algum padrão interessante sobre a classificação dos elementos.

Podemos perceber na Figura 2a que em valores mais negativos de PC2 aparecem os gases nobres e o hidrogênio, ficando os demais elementos deslocados para valores mais positivos. Observando os pesos na Figura 1 vemos que a densidade, a afinidade eletrônica e a eletronegatividade têm pesos elevados em **PC2**, puxando os escores para valores mais positivos. Para os gases nobres a densidade é baixa e a afinidade eletrônica e eletronegatividade são nulas. Por outro lado, a entropia, que é alta para gases tem peso negativo. Todos estes fatores contribuem para deslocar os gases nobres para valores bastante negativos de **PC2**.

CONCLUSÃO

Um banco de dados de propriedades periódicas e aperiódicas de elementos químicos foi sugerido como um exemplo simples para um primeiro contato com PCA, mostrando-se acessível e de fácil compreensão para estudantes de química em vários níveis de aprendizado. Isto se deve à simplicidade das questões levantadas e à necessidade apenas de conhecimento básico de química. Os gráficos de pesos mostraram relações consistentes entre as propriedades químicas e o gráfico de escores mostrou a separação das classes dos elementos. Uma reclassificação dos semimetais como metais ou ametais foi sugerida e mostrou-se de acordo com a reclassificação observada desde 2001 na Tabela Periódica da SBQ. Assim, o problema proposto atualiza o conhecimento dos estudantes com relação à classificação dos elementos químicos e fornece uma oportunidade para se revisar as relações entre diversas propriedades dos elementos.

MATERIAL SUPLEMENTAR

Está disponível em http://quimicanova.sbq.org.br, na forma de arquivo .PDF e com acesso livre. Apresenta a tabela com as propriedades: primeira energia de ionização (EI), raio atômico (RA), afinidade eletrônica (AE), eletronegatividade de Pauling (EN), densidade (D), calor específico (CE), entropia padrão (Sº) e condutividade térmica (CT) para os 88 elementos que constituem a base de dados desta atividade.

REFERÊNCIAS

1. Bona, I. A. T.; Sarkis, J. E. S.; Salvador, V. L. R.; Soares, A. L. R.; Klamt, S. C.; *Quim. Nova* **2007**, *30*, 785.
2. 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.
3. de Moura, M. C. S.; Lopes, A. N. C.; Moita, G. C.; Moita Neto, J. M.; *Quim. Nova* **2006**, *29*, 429.
4. Godinho, M. da S.; Pereira, R. O.; Ribeiro, K. de O.; Schimidt, F.; de Oliveira, A. E.; de Oliveira, S. B.; *Quim. Nova*. **2008**, *31*, 1485.
5. Magalhães, D.; Bruns, R. E.; Vasconcelos, P. C.; *Quim. Nova*. **2007**, *30*, 577.
6. Moreira, R. F. A.; Trugo, L. C.; de Maria, C. A. B.; *Quim. Nova* **1997**, *20*, 5.
7. de Sena, M. M.; Poppi, R. J.; Frighetto, R. T. S.; Valarini, P. J.; *Quim. Nova* **2000**, *23*, 547.
8. Silva, F. L. do N.; dos Santos Jr., J. R.; Moita Neto, J. M.; da Silva, R. L. G. do N. P.; Flumignan, D. L.; de Oliveira, J. E.; *Quim. Nova* **2009**, *32*, 56.
9. de Sousa, R. A.; Borges Neto, W.; Poppi, R. J.; Baccan, N.; Cadore, S.; *Quim. Nova* **2006**, *29*, 654.
10. Zimmermann, C. M.; Guimarães, O. M.; Peralta-Zamora, P. G.; *Quim. Nova* **2008**, *31*, 1727.
11. Gemberline, P. J.; *Practical Guide to Chemometrics*, 2nd ed., CRC Press: USA, 2006.
12. Brereton, R. G.; *Chemometrics: data analysis for laboratory and chemical plant*, John Wiley & Sons: England, 2003.
13. Brereton, R. G.; *Analyst* **2000**, *125*, 2125.
14. Correia, P. R. M.; Ferreira, M. M. C.; *Quim. Nova* **2007**, *30*, 481.
15. Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O.; *Quim. Nova* **1999**, *22*, 724.
16. Lide, D. R.; *Handbook of Chemistry and Physics*, 89th ed., CRC Press: USA, 2008.
17. http://www.camo.com/rt/Products/Unscrambler/unscrambler.html, acessada em Agosto 2009 e Julho 2010.
18. http://www.statsoft.com, acessada em Agosto 2009 e Julho 2010.
19. http://www.minitab.com/en-BR/default.aspx, acessada em Agosto 2009 e Julho 2010.
20. http://www.infometrix.com/software/pirouette.html, acessada em Agosto 2009 e Julho 2010.
21. http://www.mathworks.com/products/matlab/, acessada em Agosto 2009 e Julho 2010.
22. http://www.gnu.org/software/octave/, acessada em Agosto 2009 e Julho 2010.