UMA INTRODUÇÃO À ANÁLISE EXPLORATÓRIA DE DADOS MULTIVARIADOS

José Machado Moita Neto e Graziella Ciaramella Moita
Departamento de Química - Universidade Federal do Piauí - 64.049-550 - Teresina - PI

Recebido em 19/3/97; aceito em 10/10/97

AN INTRODUCTION ANALYSIS EXPLORATORY MULTIVARIATE DATE. The modern technological ability to handle large amounts of information confronts the chemist with the necessity to re-evaluate the statistical tools he routinely uses. Multivariate statistics furnishes theoretical bases for analyzing systems involving large numbers of variables. The mathematical calculations required for these systems are no longer an obstacle due to the existence of statistical packages that furnish multivariate analysis options. Here basic concepts of two multivariate statistical techniques, principal component and hierarchical cluster analysis that have received broad acceptance for treating chemical data are discussed.

Keywords: cluster analysis; principal component analysis; dendrogram.

INTRODUÇÃO

A extração de informações dos resultados de um experimento químico envolve a análise de grande número de variáveis. Muitas vezes, um pequeno número destas variáveis contém as informações químicas mais relevantes, enquanto que a maioria das variáveis adiciona pouco ou nada à interpretação dos resultados em termos químicos. A decisão sobre quais variáveis são importantes é feita, geralmente, com base na intuição química ou na experiência, ou seja, baseado em critérios que são mais subjetivos que objetivos.

A redução de variáveis através de critérios objetivos, permitindo a construção de gráficos bidimensionais contendo maior informação estatística, pode ser conseguida através da análise de componentes principais. Também é possível construir agrupamentos entre as amostras de acordo com suas similaridades, utilizando todas as variáveis disponíveis, e representá-los de maneira bidimensional através de um dendrograma. A análise de componentes principais e de agrupamento hierárquico são técnicas de estatística multivariada complementares que têm grande aceitação na análise de dados químicos.

Antes de apresentar as duas técnicas é necessário discutir alguns termos e conceitos básicos:

A MATRIZ DE DADOS

Os dados consistem em n medidas de diferentes propriedades (variáveis) executadas sobre m amostras (objetos), de modo que a matriz de dados D é formada por mxn elementos (m linhas correspondentes as amostras e n colunas correspondentes as variáveis).

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1j} & \cdots & d_{1n} \\ d_{21} & & & & \vdots & \vdots \\ \vdots & & & & \vdots & \\ d_{i1} & \cdots & \cdots & d_{1j} & & \\ & & & \vdots & \vdots & \\ d_{m1} & \cdots & \cdots & \cdots & \cdots & d_{mn} \end{pmatrix}$$

A j-ésima variável é representada por um vetor coluna. O i-ésimo objeto, ou seja, uma amostra qualquer, é representado por um vetor linha chamado vetor resposta e pode ser descrito como um ponto no espaço n-dimensional.

PADRONIZAÇÃO E ESCALONAMENTO

A finalidade da padronização e escalonamento dos dados originais é expressar cada observação em termos de variações inerentes ao sistema (autoescalonamento).

Para exemplificar a importância deste pré-tratamento da matriz de dados, vejamos o comportamento de algumas variáveis que podem ser medidas para o óleo de soja refinado[1]:

propriedade	intervalo
densidade relativa	0,919 - 0,925
índice de refração	1,466 - 1,470
índice de saponificação	189 - 195
índice de iodo	120 - 143

A amplitude da densidade é 0,006 enquanto que a do o índice de iodo é de 23. Uma diferença de densidade 0,003 entre duas amostras de óleo de soja corresponde a uma variação de 50% em relação a amplitude. Uma variação do índice de iodo desta mesma ordem de grandeza é desprezível (~0,01%). Além disso, o valor numérico entre as variáveis diferem acentuadamente de modo que a comparação direta entre variáveis levaria a uma ponderação maior das variáveis com maior valor numérico (p. ex.: índices de iodo e saponificação).

Uma maneira de resolver estes problemas, mantendo a informação estatística dos dados, é realizar uma transformação sobre o conjunto original dos dados de modo que cada variável apresente média zero e variância igual a um (autoescalonamento). Esta transformação (*z transformation*) expressa cada observação como o número de desvios padrões da média:

$$z_{ij} = \frac{d_{ij} - \bar{d}_j}{s_j}, \text{ onde } \bar{d}_j = \frac{1}{m}\sum_{i=1}^{m} d_{ij} \text{ e } s_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}(d_{ij} - \bar{d}_j)^2$$

O exemplo mostrado acima (autoescalonamento) é apenas uma das várias opções de transformações sobre o conjunto de dados que podem ser feitas.

MEDIDAS DE SIMILARIDADE

Cada objeto é representado por um ponto no espaço n-dimensional e, portanto, pode ser agrupado com outros que estejam

próximos e mais se assemelham a ele. Dois critérios de melhor associação podem ser utilizados[2]:

Covariância e Correlação

Partindo da matriz de dados D (***mxn***), obtém-se a matriz de covariância C, onde seus elementos são dados por:

$$c_{kl} = \frac{1}{m-1}\sum_{i=1}^{m}(d_{ik} - \bar{d}_k)(d_{il} - \bar{d}_l) \quad \text{onde} \quad \bar{d}_k = \frac{1}{m}\sum_{i=1}^{m} d_{ik}$$

c_{kl} é grande e positivo quando, para a maior parte das amostras, os valores das variáveis k e l desviam da média na mesma direção. Portanto, a covariância de duas variáveis é uma medida de sua associação. Para cada elemento da matriz de covariância pode ser calculado o coeficiente de correlação, consequentemente a matriz de covariância pode ser transformada numa matriz de correlação R, onde seus elementos são dados por:

$$r_{kl} = \frac{c_{kl}}{s_k \cdot s_l}$$ onde s_k e s_l são os desvios padrões das variáveis K e l

Os valores de r_{kl} são uma covariância padronizada entre -1 e +1.

Medidas de distâncias

Na análise de agrupamentos (*cluster analysis*) a similaridade entre duas amostras pode ser expressa como uma função da distância entre os dois pontos representativos destas amostras no espaço n-dimensional. A maneira mais usual de calcular a distância entre dois pontos **a** e **b** no espaço n-dimensional é conhecida por distância euclidiana (x_{ab}) e é dada por:

$$x_{ab}^2 = \sum_{j=1}^{n}\left(d_{aj} - d_{bj}\right)^2$$

Existem outras maneiras de calcular distâncias, como a distância de Mahalanobis, que não discutiremos aqui.

ANÁLISE DE AGRUPAMENTO HIERÁRQUICO

A técnica de agrupamento hierárquico interliga as amostras por suas associações, produzindo um dendrograma onde as amostras semelhantes, *segundo as variáveis escolhidas*, são agrupadas entre si. A suposição básica de sua interpretação é esta: quanto menor a distância entre os pontos, maior a semelhança entre as amostras. Os dendrogramas são especialmente úteis na visualização de semelhanças entre amostras ou objetos representados por pontos em espaço com dimensão maior do que três, onde a representação de gráficos convencionais não é possível.

Existem muitas maneiras de procurar agrupamentos no espaço n-dimensional. A maneira matematicamente mais simples consiste em agrupar os pares de pontos que estão mais próximos, usando a distância euclidiana, e substituí-los por um novo ponto localizado na metade da distância entre eles. Este procedimento, quando repetido até que todos os pontos sejam agrupado em um só ponto, leva a construção do dendrograma, onde, no eixo horizontal são colocadas as amostras e, no eixo vertical, o índice de similaridade, s_{ij}, entre os pontos i e j, calculado segundo a seguinte expressão:

$$s_{ij} = 1 - \frac{d_{ij}}{d_{máx}}$$

onde d_{ij} é a distância entre os pontos i e j e $d_{máx}$ é a distância máxima entre qualquer par de pontos. Os dendrogramas, portanto, consistem em diagramas que representam a similaridade entre pares de amostras (ou grupos de amostras) numa escala que vai de um (identidade) a zero (nenhuma similaridade).

Os dendrogramas são construídos diretamente por todos os programas estatísticos que fazem classificação dos dados através de agrupamento hierárquico *(Hierarchical Analysis ou Cluster Analysis).*

ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais consiste essencialmente em reescrever as coordenadas das amostras em outro sistema de eixo mais conveniente para a análise dos dados. Em outras palavras, as n-variáveis originais geram, através de suas combinações lineares, n-componentes principais, cuja principal característica, além da ortogonalidade, é que são obtidos em ordem decrescente de máxima variância, ou seja, a componente principal 1 detém mais informação estatística que a componente principal 2, que por sua vez tem mais informação estatística que a componente principal 3 e assim por diante.

Este método permite a redução da dimensionalidade dos pontos representativos das amostras pois, embora a informação estatística presente nas n-variáveis originais seja a mesma dos n-componentes principais, é comum obter em apenas 2 ou 3 das primeiras componentes principais mais que 90% desta informação. O gráfico da componente principal 1 *versus* a componente principal 2 fornece uma janela privilegiada (estatisticamente) para observação dos pontos no espaço n-dimensional.

A análise de componentes principais também pode ser usada para julgar a importância das próprias variáveis originais escolhidas, ou seja, as variáveis originais com maior peso (*loadings*) na combinação linear dos primeiros componentes principais são as mais importantes do ponto de vista estatístico.

Portanto, a tarefa do químico que trabalha com estatística multivariada, consiste em interpretar a distribuição dos pontos no gráfico de componentes principais e identificar as variáveis originais com maior peso na combinação linear das componentes principais mais importantes.

Existem pacotes computacionais de estatística que fazem todas as operações necessárias à obtenção de componentes principais e agrupamento hierárquico, inclusive o tratamento prévio de padronização e escalonamento dos dados, como é o caso do SPSS, SYSTAT, PIROUETTE, etc. No SPSS (*Statistical Package for the Social Sciences*), a opção de componentes principais aparece no menu através de uma de suas finalidades: a redução de dados. As componentes principais também podem ser obtido como um dos métodos da análise de fatores (*Factor Analysis*). O procedimento matemático para obtenção de componentes principais pode ser facilmente seguido por aqueles que têm conhecimento de álgebra matricial e é encontrado em diversos textos[3,4], inclusive em português e dirigido para químicos[5].

APLICAÇÃO

Para ilustrar a aplicação destas técnicas de estatística multivariada utilizamos uma tabela de composição de alimentos[6] que traz os teores de calorias, glicídios, proteínas, lipídios, cálcio, fósforo e ferro para 20 frutas. Neste caso, portanto, a matriz de dados é representada por 7 variáveis e 20 amostras. O dendrograma e os componentes principais foram obtidas no SPSS.

A figura 1 mostra o dendrograma relativo a similaridade das frutas segundo as variáveis escolhidas. As maiores similaridades são encontradas entre abacaxi e ananás, laranja pêra e tangerina e entre laranja Bahia, limão verde e limão doce. A similaridade entre os abacates, entre as mangas, entre as bananas e entre a ata e a condessa também era esperada devido a proximidade botânica. A maçã vermelha é mais próxima de abacaxi e ananás do que da maçã branca, isto deve ter ocorrido por que a composição das maçãs diferem acentuadamente em fósforo e ferro. O grupo dos abacates diferem dos demais devido ao alto teor de lipídios e calorias. O grupo das bananas se distingue pelo seu alto teor de glicídios.

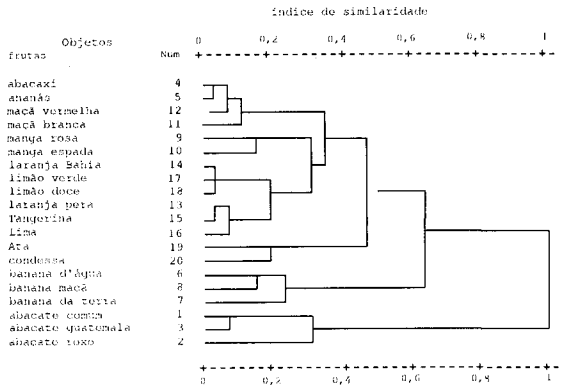

Figura 1. Dendrograma obtido da análise de agrupamento hierárquico utilizando as sete variáveis: calorias, glicídios, proteínas, lipídios, cálcio, fósforo e ferro.

A figura 2 está mostrando o gráfico da componente principal 1 *versus* a componente principal 2. Neste gráfico se distingue facilmente o grupo das bananas e dos abacates. A razão disso pode ser observada nos pesos das componentes principais: na primeira componente os maiores pesos estão em calorias (0,94) e lipídios (0,85), na segunda componente os maiores pesos estão nos glicídios (0,93) e nos lipídios (-0,48).

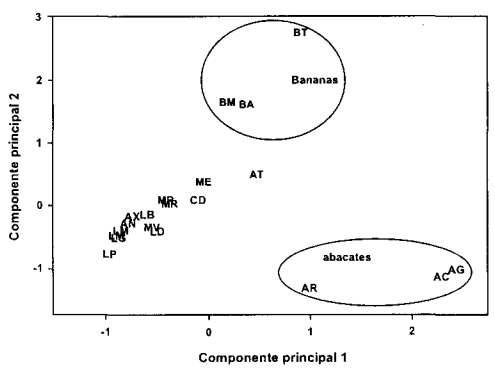

Figura 2. Gráfico da componente principal 1 versus componente principal 2. Abacate comum (AC), abacate roxo (AR), abacate guatemala (AG), abacaxi (AX), ananás (AN), banana d'água (BA), banana da terra (BT), banana maçã (BM), manga rosa (MR), manga espada (ME), maça branca (MB), maçã vermelha (MV), laranja pêra (LP), laranja Bahia (LB), Tangerina (TG), Lima (LM), limão verde (LV), limão doce (LD), Ata (AT), condessa (CD),

CONCLUSÃO

As facilidades computacionais de obtenção de dendrogramas e de gráficos de componentes principais possibilitam uma utilização mais corriqueira destes métodos no ensino e pesquisa em Química, contudo algumas observações finais são pertinentes:

a) a identificação de agrupamento pode ser feita por diversos algoritmos que podem produzir resultados diferentes entre si;

b) as variáveis escolhidas para a identificação dos grupos tem grande importância na interpretação do resultado final;

c) os gráficos da componente principal 1 *versus* componente principal 2 mostra a melhor janela para a observação dos dados, porém a componente principal 3 pode trazer informações estatística relevantes para entendimento do sistema em estudo;

O conhecimento do sistema é importante na análise estatística multivariada, portanto a interpretação destes resultados é uma tarefa dos químicos.

AGRADECIMENTOS

Agradecemos ao professor Benício de Barros Neto da Universidade Federal de Pernambuco e a professora Ieda S. Scarminio da Universidade Estadual de Londrina pelas sugestões apresentadas ao texto de divulgação sobre análise multivariada para os alunos do curso de *estatística aplicada à química* da UFPI, que posteriormente originou este artigo de divulgação.

REFERÊNCIAS

1. Codex Alimentarius Commision. *Codex Standards for Edible Soya Bean Oil.* Roma, FAO/WHO 1992. V. 8, p 9-12 (Codex Stan 20-1981).
2. Auf der Heyder, T. P. E.; *J. Chem. Educ.* **1990**, *67*, 461.
3. Kowalski, B. R.; (Ed.) *Chemometrics: Mathematical and Statistics in Chemistry.* NATO ASI series. Série C; vol. 138. D. Riedel Publishing Company, Dordrecht, 1984.
4. Malinowski, E. R.; e Howery, D. G.; *Factor Analysis in Chemistry.* John Wiley & Sons, Inc. New York, 1980.
5. Bruns, R. E. e Faigle, J. F. G.; *Quím. Nova* **1985**, *8*, 84.
6. Franco, G.; *Tabela de Composição Química de Alimentos.* 9ª edição. Livraria Atheneu Editora, Rio de Janeiro 1992.