[5] M. Gell-Mann, Phys. Lett. 8, 214 (1964); G. Zweig, CERN preprint TH 412 (1964).
[6] Veja, por exemplo, M.J. Longo, "Fundamentals of Elementary Particle Physics", MacGraw-Hill, Tokyo, 1973; J.S. Trefil, "From Atoms to Quarks", Charles Scribiner's Sons – New York, 1980.
[7] Veja artigo de G.B. Levi, Phys. Today 36, 17 (April 1983) sobre a descoberta do boson W que comprova a teoria de Weinberg – Salam.
[8] J.M.F. Bassalo, Revista de Ensino de Física 3, 13 (1981); F.E. Close, "An Introduction to Quarks and Partons", Academic Press – 1979.
[9] N. Isgur e G. Karl, Phys. Today 36, 36 (november 1983).
[10] H. Harari, Scientific American, April 1983, p. 48.
[11] D.B. Lichtemberg, Contem. Phys. 22, 311 (1981); W. Marciano e H. Pagels, Nature 279, 479 (1979); S. Gasiorowics e J.L. Rosner, Am. J. Phys. 49, 954 (1981).
[12] A. De Rújula, R. C. Giles e R.L. Jaffe, Phys. Rev. D. 17, 285 (1978); J.F. Donoghue, Comm. Nucl. Part. Phys. 10, 277 (1982).
[13] H. Enge, "Introduction to Nuclear Physics", Addison – Wesley Publishing Company, 1966.
[14] R.A. Millikan, Phil. Mag. 19, 209 (1919).
[15] G. LaRue, J.D. Phillips e W.M. Fairbank, Phys. Rev. Lett. 46, 967 (1981).
[16] M. Marinelli e G. Morpurgo, Phys. Rep. 85, 162 (1982).
[17] M.A. Lindgren e colaboradores, Phys. Rev. Lett. 51, 1621 (1983).
[18] Veja, por exemplo, S. Weinberg, "Os Três Primeiros Minutos", Guanabara Dois, 1980; D.N. Schramm, Phys. Today 36, 27 (April 1983).
[19] K.S. Lackner e G. Zweig, Phys. Rev. D28, 1671 (1983).
[20] C.K. Jørgensen, Structure and Bonding 43, 1 (1981).
[21] W.M. Fairbank Jr., "Measurements of Very Low Sodium Vapor Densities By Laser Fluorescence Spectroscopy", PhD Thesis, Stanford University, 1974.
[22] A.C. Pavão – "Quark – Antiquark pair stabilization and electronic structures of fractionally Charged atoms", submetido para publicação no Phys. Rev. D.
[23] L.J. Schaad e Colaboradores, Phys. Rev. A23, 1600 (1981).

ARTIGO

QUIMIOMETRIA

Roy E. Bruns e J.F.G. Faigle

Instituto de Química
Universidade Estadual de Campinas
C.P. 6154 – 13.100 Campinas, - S.P.

Recebido em 02/04/84

INTRODUÇÃO

A quimiometria é a parte da química que utiliza métodos matemáticos e estatísticos para:

a) definir ou selecionar as condições ótimas de medidas e experiências e

b) permitir a obtenção do máximo de informações a partir da análise de dados químicos.

Com a sofisticação sempre crescente das técnicas instrumentais, impulsionada pela invasão de microprocessadores e microcomputadores no laboratório químico, tornam-se necessários tratamentos de dados mais complexos do ponto de vista matemático e estatístico, a fim de relacionar os sinais obtidos (intensidades, por exemplo) com os resultados desejados (concentrações). Muita ênfase tem sido dada aos sistemas multivariados, nos quais se pode medir muitas variáveis simultaneamente (ou de forma seqüencial, com grande eficiência) ao se estudar uma amostra qualquer. Nesses sistemas, a conversão da resposta instrumental no dado químico de interesse, requer a utilização de técnicas de estatística multivariadas, álgebra matricial e análise numérica. Essas técnicas se constituem no momento na melhor alternativa para a interpretação de dados e para a aquisição do máximo de informações sobre o sistema.

De todos os ramos da química clássica, talvez a química analítica tenha sido a mais afetada pelo desenvolvimento recente da instrumentação química, associada a computadores. De fato, a "Chemometrics Society", organização internacional dedicada ao uso e desenvolvimento de métodos em quimiometria, é composta principalmente por químicos interessados em problemas analíticos. Atualmente é muito raro se encontrar qualquer periódico respeitável sobre pesquisa em química analítica, que não traga artigos reportando dados obtidos com o auxílio de microprocessadores, ou tratados por matemática multivariada ou métodos estatísticos, sempre com o objetivo de melhorar a qualidade dos resultados ou facilitar a sua interpretação.

Outras áreas em química vêm sendo também influenciadas pela quimiometria. Os métodos espectrométricos comumente empregados em química orgânica e inorgânica, tornam-se muito mais poderosos quando acoplados a técnicas

de matemática multivariada, o que permite a análise quantitativa de grande quantidade de dados, os quais até agora, só podiam ser analisados de forma qualitativa. Sistemas de reações que são por natureza multivariados, dependentes de temperatura, pH, solvente, etc., podem ser otimizados de forma sistemática utilizando métodos como o SIMPLEX[2]. Muitos campos da físico-química são afetados pela quimiometria. Vários modelos universais rígidos, como a química quântica, podem ser aplicados apenas a sistemas simples. As técnicas de reconhecimento de padrões podem se constituir em métodos complementares ou alternativos, na solução desses problemas e de outros na área de físico-química.

O emprego de computadores na aquisição e interpretação de dados, deverá sofrer no futuro um aumento muito significativo. O aprimoramento técnico de microprocessadores e microcomputadores, vem ocorrendo mais rapidamente que o desenvolvimento da instrumentação química. Ademais, o equipamento utilizado em computação, apesar de se tornar a cada dia mais eficiente e flexível, tem seus custos reduzidos, contrastando com a escalada de preços da instrumentação química convencional.

A quimiometria é um ramo recente da química. Excetuando-se alguns esforços isolados, a década de 70 marca o início da atividade de pesquisa concentrada nesse campo. É ainda difícil prever que tipo de pesquisa em quimiometria obterá maior êxito no futuro.

Os artigos de revisão listados na bibliografia[3], cobrem a maioria dos aspectos ligados às atividades da quimiometria. Nesse trabalho, pretendemos apresentar breves descrições didáticas dos diferentes tópicos da quimiometria, que são objeto de investigação ativa em todo o mundo. Serão enfatizados dois tópicos, em estudo no Brasil:

1) Calibração multivariada por meio do método geral de adição de padrões[4].

2) Classificação de espécies químicas e interpretação de dados multivariados por reconhecimento de padrões.

Espera-se com a discussão desses tópicos fornecer ao leitor uma idéia sobre cada uma das atividades gerais em quimiometria, conforme a definição apresentada no início desse artigo. Os leitores com maior interesse no assunto devem se reportar aos livros e revisões mais exaustivas[2,5], escritos por alguns dos pioneiros nesse campo.

Ao encerrar esta introdução, afirmamos entender que a quimiometria só poderá dar seu total impacto no pensamento químico, após a publicação de textos sobre o assunto e sua inclusão nos currículos de graduação e pós-graduação dos cursos de química. Os alunos poderão então ser educados para pensar em termos de sistemas multivariados em espaços multidimensionais, ao invés de pensar em termos de espaços a três dimensões ou menos. Isto deverá ampliar sua habilidade em resolver os problemas mais complexos, cuja natureza é em si multivariada.

ESTATÍSTICA

Todas as medidas feitas em química analítica apresentam uma incerteza. A despeito disso, uma grande quantidade de análise é feita sem grande preocupação em expressar a incerteza dos resultados, através das ferramentas da análise estatística. Isso é particularmente verdadeiro quando os resultados analíticos são utilizados na solução de problemas em outras áreas da química, como na química ambiental, química forense, etc. Recentemente, Currie[6] observou a relação entre as incertezas de alguns dados químicos e as decisões de impacto social tomadas com base nesses dados. Os estudos sobre precisão e exatidão, propagação de erros, caracterização da relação sinal/ruído, teoria de amostragem, otimização de experiências utilizando esquemas fatoriais e análises de séries temporais, estão todos fundamentados na estatística. Mais informações podem ser encontradas na referência 3b.

RESOLUÇÃO

Parte substancial das atividades em quimiometria está voltada à resolução de picos superpostos, tanto na análise de substâncias puras como na de misturas. No caso de substâncias puras, são utilizados diferentes tipos de funções matemáticas para representar os picos principais, cuja resultante fornece um conjunto de picos superpostos, do tipo que costuma preocupar os químicos analíticos. O método multivariado de análise de componentes principais[7] (veja a secção de reconhecimento de padrões) e o de análise fatorial[8], muito semelhante ao primeiro e largamente empregado em ciências sociais, vêm se tornando método padrões na separação de picos superpostos em análises de misturas de muitos componentes. A grande abrangência do método levou à publicação de um texto introdutório de fácil leitura[5d], que pacientemente expõem seu funcionamento e ilustram aplicações em espectroscopia de absorção e de emissão, cromatografia e espectrometria de massa.

Aqui apresentaremos os conceitos básicos envolvidos na análise de componentes principais, não apenas para ilustrar sua aplicação na solução de problemas numéricos, mas como uma introdução ao seu emprego nas técnicas de classificação por reconhecimento de padrões.

Suponha a análise de uma mistura por uma técnica espectrométrica, por exemplo, espectroscopia no visível e ultravioleta. Os espectros são obtidos a vários comprimentos de onda selecionados, numa série de misturas onde as concentrações relativas dos componentes variam. A matriz de dados corresponde a um agrupamento sistemático das intensidades registradas a diferentes comprimentos de onda, para todas as misturas em estudo. Sua forma geral pode ser representada pela matriz A:

$$\underset{\sim}{A} = \begin{bmatrix} A_{11} & A_{12} \ldots . A_{1p} \\ A_{21} & A_{22} \ldots . A_{2p} \\ . & . \quad . \\ . & . \quad . \\ . & . \quad . \\ . & . \quad . \\ A_{n1} & A_{n2} \ldots . A_{np} \end{bmatrix} \qquad (1)$$

onde os elementos A_{ki} representam as intensidades obtidas no i-ésimo comprimento de onda para a mistura k ($i = 1,2, \ldots p$; $k = 1,2, \ldots n$). A matriz de covariança é

obtida multiplicando-se a matriz de dados pela sua transposta:

$$\underset{\sim}{COV} = \underset{\sim}{A}^t \underset{\sim}{A} \quad (2)$$

Essa matriz é posteriormente diagonalizada, fornecendo como solução:

$$(\underset{\sim}{COV})\ \underset{\sim}{\beta}^t = \underset{\sim}{\beta}^t\ \underset{\sim}{\Lambda} \quad (3)$$

ou

$$(\underset{\sim}{COV})\ \vec{\beta}_j^t = \lambda_j\ \vec{\beta}_j^t \quad (4)$$

onde $\vec{\beta}_j^t$ é o autovetor correspondente ao j-ésimo autovalor, compondo a j-ésima coluna na matriz $\underset{\sim}{\beta}^t$. Ademais, o j-ésimo autovalor traduz a fração da variança original expressa pelo j-ésimo autovetor. Assim, a diagonalização da matriz de covariança permite a projeção de um espaço de p variáveis num outro de ordem inferior, em geral de duas ou três dimensões, a fim de facilitar a visualização, o que é de grande importância em problemas de classificação de dados. Os "A" primeiros autovetores, quando ordenados de forma decrescente segundo seus autovalores, expressam uma percentagem da variança dada por: $\sum_{i=1}^{A} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ (100).

O valor de A costuma ser pequeno nos problemas classificatórios, ficando em geral em torno de dois ou três, dependendo da complexidade do conjunto de dados. Em problemas numéricos, o valor de A é determinado pelo número de componentes principais necessários para descrever o conjunto de dados. Os outros (A-p) autovetores não são significativos, uma vez que expressam uma percentagem de variança comparável aos valores dos erros experimentais. A matriz $\underset{\sim}{A}$ é então descrita por:

$$\underset{\sim}{A} = \underset{\sim}{\Theta}\ \underset{\sim}{\beta} \quad (5),$$

onde os autovetores contidos em $\underset{\sim}{\beta}$ são chamados "loadings" e os valores de Θ são chamados "scores", ou seja, valores das variáveis das *n* amostras, após a transformação para o espaço descrito pelos autovetores. Os valores de $\underset{\sim}{\Theta}$ podem ser calculados pela equação:

$$\underset{\sim}{A}\ \underset{\sim}{\beta}^t = \underset{\sim}{\Theta} \quad (6),$$

como a matriz $\underset{\sim}{\beta}$ é ortogonal, $\underset{\sim}{\beta}^t = \underset{\sim}{\beta}^{-1}$.

As matrizes $\underset{\sim}{\Theta}$ e $\underset{\sim}{\beta}$ apresentam alguns fatores abstratos, cuja importância está em predizer o número de componentes principais que determina a estrutura dos dados da matriz $\underset{\sim}{A}$. Normalmente, procura-se obter uma matriz de transformação apropriada, mediante a rotação dos fatores abstratos no espaço *p* ou através do "target testing", o que converte tais fatores em fatores reais.

$$\underset{\sim}{A} = \underset{\sim}{C}\ \underset{\sim}{\epsilon} \quad (7)$$

ou

$$A_{ki} = \sum_{j=1}^{m} C_{kj}\, \epsilon_{ji} \quad (8),$$

onde ϵ_{ji} é a absortividade molar por unidade de comprimento do caminho ótico do componente j no comprimento de onda i, C_{kj} é a concentração molar do componente j na mistura k e m é o número de componentes na mistura.

Os fatores reais e abstratos se relacionam entre si através de uma matriz linear de transformação. Cada linha da matriz $\underset{\sim}{\epsilon}$ corresponde às absorvâncias de um dos componentes, nos p comprimentos de onda estudados. Cada coluna da matriz $\underset{\sim}{C}$ corresponde às concentrações de um dos m componentes de cada uma das n misturas.

Essa transformação para componentes principais reais, obviamente depende da validade da lei de Beer para o sistema estudado.

Em resumo, a análise de componentes principais determina:

1) m, o número de componentes principais do sistema, que é o número de constituintes ativos da absorção.
2) C_{kj}, as concentrações de cada componente em cada mistura e
3) ϵ_{ji}, relacionado com o espectro de cada componente.

O preço a ser pago pela variedade de informações obtida por análise fatorial, está no grande número de medidas que se deve fazer, num número significativo de misturas. Podemos ter uma idéia desse problema, analisando os pontos principais de um dos trabalhos recentes no campo. Kankare[5d], estudando complexos, mediu espectros de absorção de soluções formadas por 8.10^{-5} M em íons de bismuto, 1 M em ácido perclórico e quantidades variáveis de cloreto de sódio e perclorato de potássio, a fim de manter constante a força iônica do meio. Os espectros foram obtidos contra soluções de referência, que continham a mesma composição, mas sem íons de bismuto. Foram medidas as absorvâncias de 17 soluções, entre 230 e 360 nm, a intervalos de 5 nm, o que resultou em 27 comprimentos de onda nas 17 soluções, gerando portanto uma matriz $\underset{\sim}{A}$ 17 x 27.

Para determinar os valores de *m* nessa matriz de 459 elementos, Kankare comparou os desvios padrão residuais dos dados experimentais, calculados a partir das diferenças entre as absorvâncias obtidas da equação 8 e as absorvâncias experimentais, com o desvio padrão estimado de seus valores experimentais.

Como o desvio padrão residual calculado por análise fatorial deve ser menor que o desvio estimado (o número mínimo de fatores que satisfaz esse critério é m), foi possível concluir que haviam 7 espécies ativas na absorção das soluções. Essas espécies foram propostas como sendo: Bi^{3+}, $BiCl^{2+}$, $BiCl_2^+$, $BiCl_3$, $BiCl_4^-$, $BiCl_5^{2-}$ e $BiCl_6^{3-}$.

O programa para computador denominado FACTANAL, que faz análise fatorial e inclui o "target testing", pode ser obtido pelo "Quantum Chemistry Program Exchange[9]".

CALIBRAÇÃO

Um método ideal de calibração é dado por R = KC, onde R representa a medida da resposta de um sensor, C é uma concentração e K uma constante de resposta (a inclinação de curva de calibração). Para que esse modelo seja verdadeiro e portanto útil, o sensor não deve responder a espécies interferentes na amostra e os efeitos matriciais[4] devem estar ausentes. A primeira restrição leva a uma mudança no modelo, que deve ser generalizado na forma:

$$R = K_A C_A + \sum_{I=1}^{W} K_I C_I \qquad (9)$$

onde o índice A refere-se à espécie de interesse e I aos interferentes. Se os termos K_I e C_I forem conhecidos, o somatório do seu produto pode ser subtraído de R, eliminando o efeito interferente. Essa operação é idêntica a uma correção de contagem de fundo.

Em análises de misturas, as interações entre as espécies e seus interferentes são menos claras, já que cada espécie pode interferir na determinação de uma outra. O GSAM* considera os sinais interferentes como informações deslocadas, que não devem ser eliminadas por subtração, mas sim utilizadas na determinação das concentrações das espécies. Como esse método é uma generalização do método simples de adição de padrões, os efeitos matriciais estão incluídos nas curvas de calibração utilizadas para determinar as concentrações das espécies.

Consideremos o caso de se analisar *r* espécies com *p* sensores analíticos, sendo r < p. São feitas m adições de padrão de cada componente a ser analisado, sendo tomadas as leituras de cada um dos p sensores após cada adição. Sendo n o número total de adições de padrão (n = m.r), a equação matricial básica será:

$$\underset{\sim}{R} = \underset{\sim}{C}\,\underset{\sim}{K} \qquad (10),$$

onde $\underset{\sim}{R}$ é uma matriz (n.p) contendo os valores das leituras de cada sensor, depois de cada adição de padrão e $\underset{\sim}{C}$ é uma matriz (n.r), contendo os valores das concentrações de cada espécie após cada adição de padrão. A matriz $\underset{\sim}{K}$ (r.p) contém as constantes de resposta para todas as concentrações das espécies, em cada sensor analítico. Se r = p, $\underset{\sim}{K}$ será uma matriz quadrada que admite uma inversa, através da qual se pode calcular as concentrações originais, usando:

$$\underset{\sim}{R}\,\underset{\sim}{K}^{-1} - \underset{\sim}{\Delta C} = {}_{o}\underset{\sim}{C} \qquad (11)$$

Aqui, $\underset{\sim}{\Delta C}$ é uma matriz (n.r) cujos elementos são as variações de concentração de cada espécie, a cada adição de padrão. ${}_{o}\underset{\sim}{C}$ é uma matriz (n.r) que apresenta o mesmo número de linhas que $\underset{\sim}{\Delta C}$, cada linha contendo a concentração original das r espécies presentes na amostra em estudo.

Pode-se melhorar a qualidade dos resultados utilizando-se um número de sensores maior que o de espécies a serem analisados (p > r), o que torna K uma matriz retangular. Nesse caso, a equação (11) se torna:

$${}_{o}\underset{\sim}{C} = [\underset{\sim}{R}\,(\underset{\sim}{\Delta C}^{t}\,\underset{\sim}{\Delta C})^{-1}\,\underset{\sim}{\Delta C}^{t} - \underset{\sim}{I}]\,\underset{\sim}{\Delta C} \qquad (12),$$

onde $\underset{\sim}{I}$ representa a matriz identidade.

Na elaboração de um método analítico para vários componentes, é útil se conhecer os elementos da matriz $\underset{\sim}{K}$, definida por:

$$\underset{\sim}{K} = (\underset{\sim}{\Delta C}^{t}\,\underset{\sim}{\Delta C})^{-1}\,\underset{\sim}{\Delta C}\,\underset{\sim}{\Delta R} \qquad (13)$$

Assim, pode-se por exemplo utilizar um detector mais sensível com melhor limite inferior de detecção, a despeito de uma eventual perda de seletividade, quando se utiliza o GSAM.

O GSAM tem sido aplicado à espectroscopia no visível e ultravioleta[10], espectroscopia de plasma de argônio acoplado indutivamente[11] e na voltametria extrativa (stripping voltammetry)[12].

No Brasil, um esforço considerável tem sido feito no sentido de minimizar o tempo de análise para utilização do GSAM. O grande número de adições de padrões necessários para a análise de muitos componentes, tem sido um sério impedimento para o uso do método em análises de rotina. Para evitar o grande consumo de tempo que representam as várias adições manuais de padrões, tem sido sugerido o uso de injeção em fluxo (FIA-sigla inglesa para "Flow Injection Analyses"), em sistemas acoplados diretamente ao detetor (nesse caso, em espectroscopia de emissão de plasma[13]), o que permite a coleta de dados com alta freqüência, sem perda da precisão[14]. Alguns aperfeiçoamentos dessa técnica, além da aplicação do FIA, têm ainda levado a um aumento da exatidão e da precisão dos resultados[15-18].

Pode-se obter a baixo custo o programa GSAM, que faz todas as operações matriciais necessárias nos cálculos de adições de padrões.[19]

PROBLEMAS DE CLASSIFICAÇÃO E PRÉ-PROCESSAMENTO

Os problemas classificatórios são muito comuns em ciência e engenharia. Problemas dessa natureza, há muito tempo estão sendo resolvidos em biologia, psicologia, economia e engenharia, utilizando reconhecimento de padrões (RP).

Na última década, vários grupos de pesquisa começaram a aplicar RP aos problemas de natureza química. Muitas drogas já foram classificadas como biologicamente ativas ou não ativas, com base em seus parâmetros físico-químicos[20]. Amostras de vinho foram classificadas de acordo com o ano e a região onde as uvas foram colhidas[21]. Petróleo[22], água mineral[23], artefatos arqueológicos[24] e amostras minerais[25], já foram classificados com respeito a suas origens. Com base em intensidades de picos cromatográficos, foi possível classificar em categorias várias marcas de whiskey[26]. Doenças de fígado foram classificadas em dois

GSAM (sigla inglesa para "General Standard Addition Method").

tipos, com base nas concentrações de sete enzimas analisadas no sangue de pacientes[27]. Produtos manufaturados de natureza química foram classificados como dentro ou fora das especificações, utilizando as concentrações dos componentes em maior proporção nesses produtos[28].

O reconhecimento de padrões e sua área complementar de conhecimento de padrões*são ramos da inteligência artificial, como se vê na figura 1. Se a distribuição de probabilidade estatística dos dados referentes a uma amostra é conhecida, pode-se trabalhar no modo parametrizado. O mais comum em problemas químicos é que essa distribuição não seja conhecida, o que exige a aplicação do reconhecimento e conhecimento de padrões não parametrizados. Havendo a disponibilidade de um conjunto de treinamento, ou seja um conjunto para o qual se conhece a categoria a que pertence cada amostra, é possível derivar regras de classificação, com base em medidas das variáveis relativas a cada espécie. Esse processo é conhecido como aprendizagem supervisionada. A seguir, a partir de medidas relativas às amostras desconhecidas, pode-se classificá-las pelas regras estabelecidas a partir do conjunto de treinamento.

Em alguns casos não se possui um conjunto de treinamento, ou ainda dispomos de informação suficiente sobre o sistema, para se prever o número de categorias esperado para um grupo de amostras. Nesse caso utiliza-se a conhecimento de padrões, que é uma forma de aprendizagem não supervisionada. O método de conhecimento de padrões mais empregado é o de análise de agrupamentos de dados, que procura classificar num mesmo subgrupo as amostras mais semelhantes entre si.

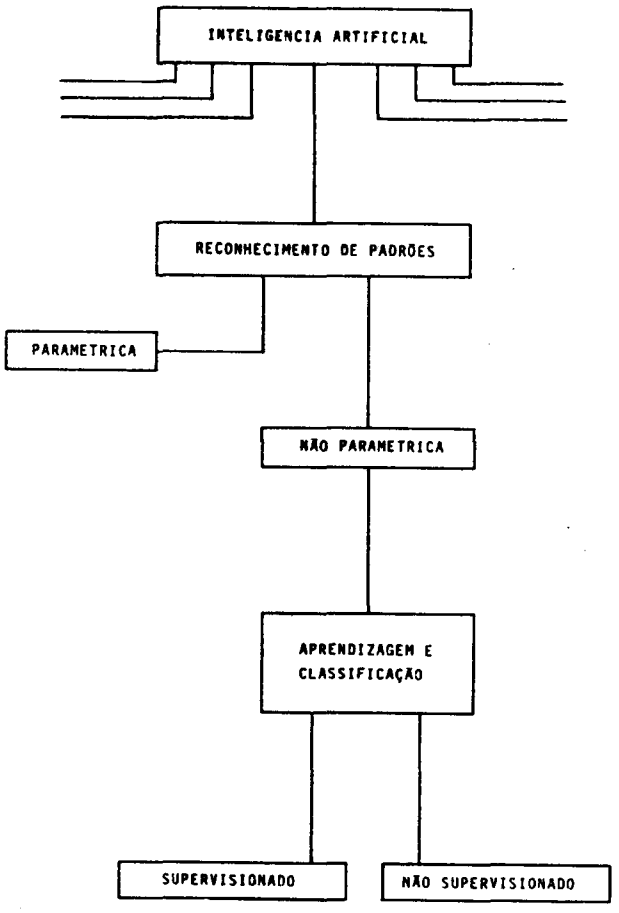

Fig. 1 – A posição do reconhecimento de padrões no campo da inteligência artificial.

A classificação não é o único objetivo do reconhecimento de padrões. Em geral, deseja-se obter uma compreensão mais profunda da natureza do problema estudado. Por exemplo, através do reconhecimento de padrões, pode-se facilmente descobrir quais parâmetros físico-químicos (caso existam) são importantes para discriminar entre drogas ativas e não ativas biologicamente. Entretanto, essa informação não satisfaz plenamente. O que se deseja, em geral, é saber por que esses parâmetros são importantes e como eles variam quantitativamente com a atividade biológica da droga. Para tratar esses problemas mais complexos, envolvendo sistemas biológicos, misturas de produtos naturais, problemas ambientais e outros, existem poucas teorias bem estruturadas e que sejam suficientemente exatas para contribuir para a solução dos problemas. Nesses casos, as teorias mais flexíveis mas apenas localmente válidas, como as fornecidas pelo reconhecimento de padrões, podem resultar em maiores contribuições.

A idéia básica de reconhecimento de padrões é de fácil compreensão. Cada objeto ou amostra, k, de um conjunto de dados, pode ser descrito por um conjunto de propriedades ou características, representadas pelas variáveis, i = 1, 2. . . . . p. O objeto k corresponde ao ponto (X_{k1}, X_{k2}, ... X_{ki} . . . . X_{kp}), num espaço p-dimensional, conforme ilustra a figura 2. Quanto mais próximos estejam dois pontos num espaço p-dimensional, mais semelhantes serão as amostras relativas a esses dois pontos. Matematicamente, isso é expresso pela equação:

$$S_{k1} = 1{,}000 - d_{k1}/(d_{k1})_{max} \qquad (14),$$

onde S_{k1} representa a similaridade entre os pontos *k* e *1* e d_{k1} a distância (Euclidiana ou não) entre esse ponto[30]. O termo $(d_{k1})_{max}$ representa a maior distância entre qualquer par de pontos no conjunto em estudo; a escala de similaridade vai então de zero (nenhuma similaridade) até um (identidade).

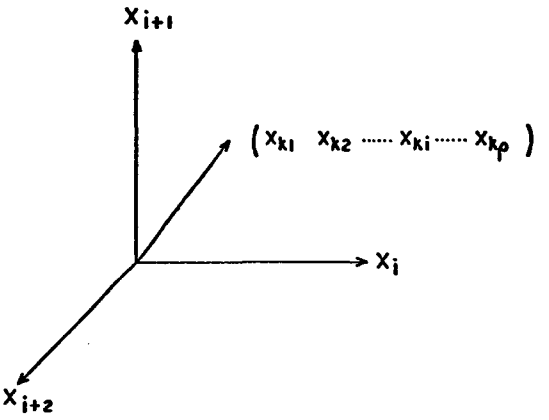

Fig. 2 – Representação gráfica do ponto relativo à k-ésima amostra, com variáveis (x_{k1}, x_{k2}, . . . x_{ki} . . .x_{kp}).

* do inglês "pattern recognition and pattern cognition".

O químico, talvez sem se dar conta, faz rotineiramente medidas multivariadas com grandes valores de p. Por exemplo, os espectros de massa, de RNM, Raman, infravermelho e outros, que são manifestações em infinitas dimensões das variáveis relacionadas à natureza da amostra, podem ser facilmente expressos, de forma aproximada, por 100 ou mais valores discretos de variáveis. Existem instrumentos analíticos, como o espectrômetro de emissão de plasma já mencionado, que fornecem rotineiramente as concentrações de 20 elementos diferentes, simultaneamente. Inúmeras técnicas utilizadas em química fornecem dados multivariados e pode-se esperar que o número delas cresça no futuro.

Com valores crescentes de p e com um número cada vez maior de amostras a serem analisadas diariamente, é necessário utilizar um computador digital para armazenar, manipular e ordenar a informação quantitativa recebida. O homem, embora sua memória deixe muito a desejar, excede em muito o computador na capacidade de identificar tendências ou padrões. Por essa razão, um bom método de reconhecimento de padrões combina a infalível memória do computador (para cálculos no espaço p), com a inata habilidade humana em reconhecer padrões (por meio de projeções em espaço bi ou tri-dimensional), a fim de resolver problemas multivariados. Os programas elaborados para grandes computadores ou para microcomputadores, não fazem outra coisa senão otimizar a capacidade do sistema homem-máquina para resolver problemas de reconhecimento de padrões.

A última versão do programa ARTHUR[31] para computadores de grande porte e o SIMCA/3B[32] para micros, são indicados para problemas de reconhecimento de padrões em química. Esses programas podem ser adquiridos a preços relativamente baixos. O ARTHUR/75, o primeiro programa de reconhecimento de padrões utilizado em nosso laboratório, não é protegido por direitos autorais; por essa razão, enviaremos com prazer cópias do ARTHUR/75 a laboratórios interessados na América Latina.

Para a manipulação apropriada dos dados multivariados, deve-se armazená-los numa matriz de dados do tipo indicado na figura 3. Cada linha contém todos os valores das variáveis para uma dada amostra, enquanto cada coluna fornece o valor de uma variável para todas as amostras. Em outras palavras, o índice das linhas indica a amostra e o das colunas, a variável.

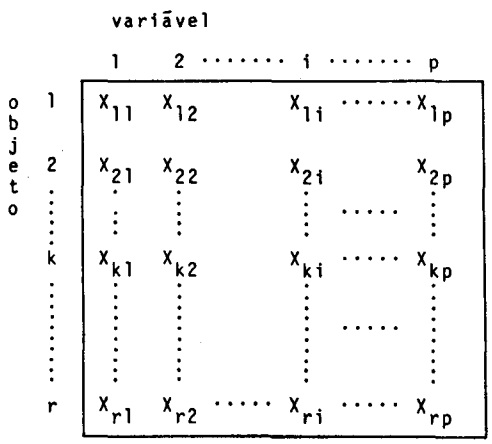

Fig. 3 – A matriz de dados, $\underline{X}$. O elemento X_{ki} corresponde ao valor da variável i para a amostra k.

Antes de aplicar os métodos de reconhecimento de padrões aos elementos da matriz de dados, esses dados são em geral pré-processados. Esse estágio do reconhecimento de padrões costuma ser complicado, podendo incluir redução de variáveis e detecção de pontos deslocados (outliers). Normalmente, o primeiro passo do pré-processamento envolve um escalamento, a fim de que cada variável receba o mesmo peso inicial nos cálculos. Isto se consegue por meio de uma transformação das variáveis, de forma que o valor médio e a variança (o quadrado do desvio padrão) sejam 0 e 1, respectivamente. Esse tipo de escalamento assegura que as influências relativas das diferentes variáveis sobre os cálculos, sejam independentes das unidades dessas variáveis.

A redução de variáveis permite eliminar os valores que não sejam relevantes para a classificação desejada. Uma maneira de se fazer essa redução, consiste em decidir se um valor é mantido ou eliminado da matriz de dados. Pode-se também aplicar um método mais suave, calculando o peso de variança[30] ou o peso de Fisher[33] para cada variável, num par de categorias.

O peso de Fisher para a variável i e para amostras das categorias p e q, pode ser calculado da seguinte maneira: divide-se o quadrado da diferença dos valores médios da variável i para as classes p e q, pela soma das variâncias dessa variável nas duas categorias, ou seja:

$$W_{pq}(i) = \frac{(\bar{X}_i(p) - \bar{X}_i(q))^2}{S_i^2(p) + S_i^2(q)} \qquad (15)$$

O conceito do peso de Fisher está ilustrado na figura 4. Seu valor aumenta com o aumento do numerador da expressão 15. Em outras palavras, quanto mais separados os valores da variável para as duas classes, mais importantes será essa variável para discriminar os objetos das classes p e q. O valor do peso de Fisher aumenta também quando diminui a soma das variâncias dentro de cada classe. Variâncias pequenas correspondem a picos mais estreitos na figura 4.

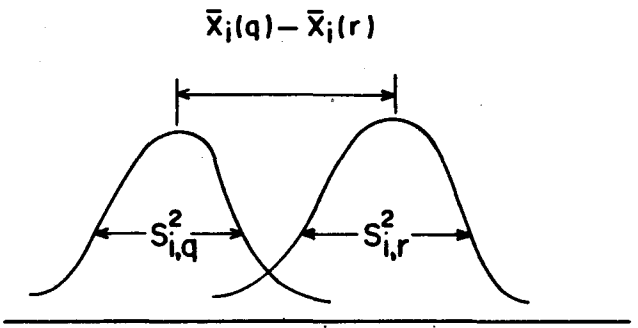

Fig. 4 – O peso de Fisher para a variável i e para as classes q e r, é dado pela equação 15.

As variáveis que exibem picos mais estreitos apresentam maior resolução na separação das categorias ou classes. Assim, quanto maior o peso de Fisher, maior o potencial discriminatório da variável i para as classes p e q. Pode-se calcular o peso de Fisher para cada variável, simplesmente tomando a média dos pesos obtidos na equação (15), para

TABELA I

Valores de peso de variança para a discriminação de três categorias de água mineral.

SN x L [a]		SN x V [a]		L x V [a]	
Elemento	Peso	Elemento	Peso	Elemento	Peso
Na	2,6	K	7,0	K	3,9
Si	2,4	Na	2,3	Si	1,5
K	2,1	Ca	1,5	Mg	1,1
Ca	1,4	Si	1,2	P	1,1
Mg	1,2	Mg	1,1	Ca	1,0
P	1,0	P	1,0	Na	1,0

a) Categorias: SN = Serra Negra; L = Lindóia e V = Valinhos
Para cada par de categorias os pesos de variança dos elementos são dados em ordem de valor decrescente.

todos os pares de categorias. Esses pesos médios podem então ser multiplicados pelos valores das variáveis auto-escaladas:

$$X'_{ki} = W_i X_{ki} \qquad (16),$$

para se obter valores ponderados das variáveis, X'_{ki}, o que pode facilitar a separação das amostras em suas categorias apropriadas.

O peso de variança tem a mesma função que o de Fisher. Para as classes p e q, o peso de variança é calculado dividindo a variança interclasse pela soma das variânças intraclasse.

Os pesos de variança calculados para concentrações de seis elementos encontrados em águas minerais correspondentes a três categorias, Serra Negra, Lindóia e Valinhos, estão na Tabela 1. Note que os pesos para as concentrações de sódio e potássio são relativamente elevados para quase todos os pares de categorias, indicando o elevado poder de discriminação dessas variáveis. Por outro lado, as concentrações de magnésio e fósforo são de pouca relevância para a separação de amostras das três categorias.

Em geral, consideramos recomendável o escalamento em quase todas as aplicações de reconhecimento de padrões. Calculamos sempre os pesos de variança e de Fisher, o que nos permite julgar a importância relativa de cada variável na separação das categorias. Entretanto, temos evitado o emprego das variáveis ponderadas nos cálculos, o que levaria a uma maior percentagem de classificações corretas, mas poderia forçar um tipo de resultado.

MÉTODOS DE RECONHECIMENTO DE PADRÕES

Análise de Discriminante Linear – LDA (Linear Discriminant Analysis). A análise de discriminante linear (LDA), consiste em procurar uma função linear que permite distinguir os pontos relativos a dados de duas categorias diferentes[2]. Se existe tal função, dizemos que os pontos pertencentes às duas categorias são linearmente separáveis. Não é difícil estender essa análise a n categorias, utilizando (n-1) funções lineares capazes de separar cada uma das n categorias das demais.

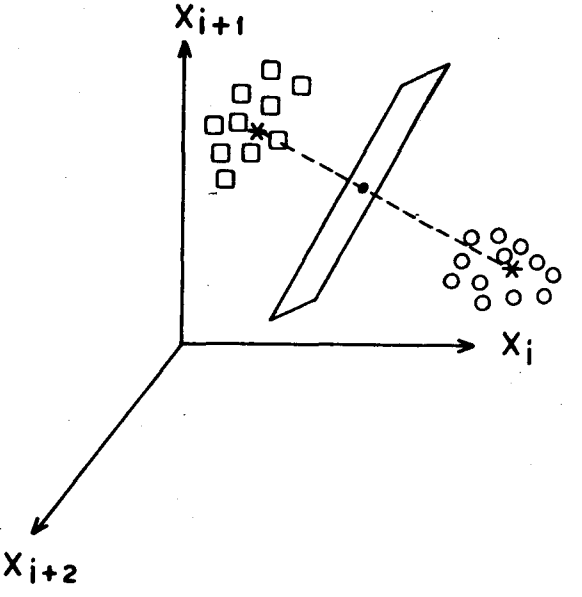

Fig. 5 – A técnica de análise de discriminante linear. Determina-se os centros de gravidade para as duas categorias, definidas por círculos e quadrados. O plano de decisão é traçado perpendicularmente à linha que une os centros de gravidade, numa posição eqüidistante a esses centros.

Uma das maneiras mais simples para se obter uma função linear discriminatória está ilustrada na figura 5. Determina-se os centros de gravidade para as categorias, unindo-os posteriormente por uma linha reta. O plano de decisão ou função de discriminante é o perpendicular a essa reta, eqüidistante dos centros de gravidade. O espaço fica então dividido em duas regiões, R_1 e R_2. Se um ponto referente a uma categoria ignorada aparece na região R_1, é classificado como pertencente à categoria 1; se aparece na região R_2, à categoria 2. Num espaço tridimensional, temos um plano de discriminação. Num espaço de quatro ou mais dimensões, a superfície de decisão será um hiperplano m-1, onde m é o número de dimensões (ou variáveis). Note-se que as amostras são sempre classificadas como pertencentes a uma ou outra categoria. Esse tipo de análise discri-

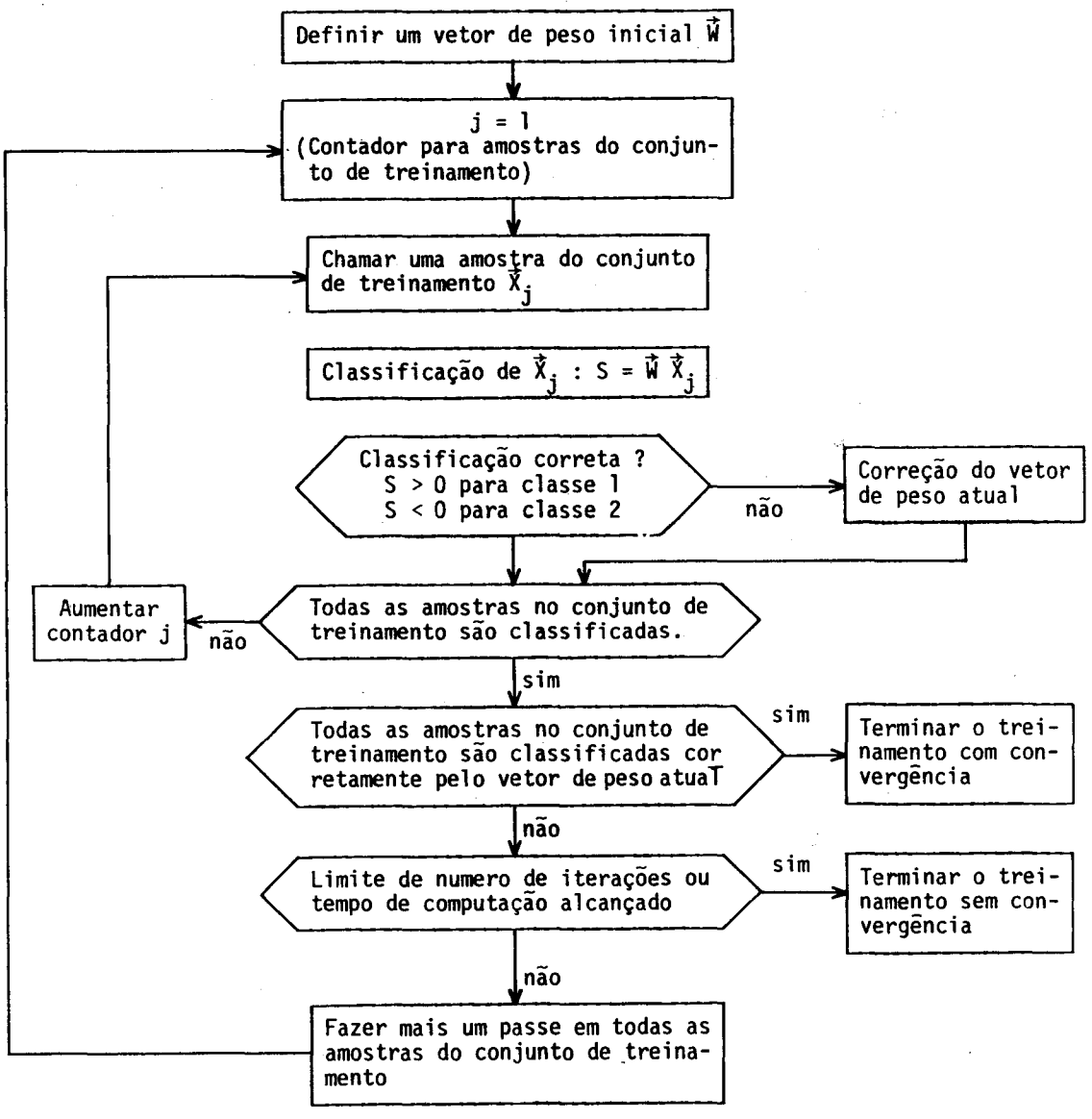

Fig 6 – Por meio de métodos iterativos pode-se traçar um plano ou hiperplano, que separa as duas categorias. O esquema deste método iterativo para um classificador linear e binário é dado nesta figura.

minatória não contempla a possibilidade da amostra em estudo pertencer a uma categoria ainda não definida. Muitas variantes foram propostas para o cálculo de funções de discriminantes lineares, descritas em vários textos.

Máquina de Aprendizagem Linear – LLM[34] (Linear Learning Machine). Conforme se observa na figura 5, os pontos relativos a duas categorias podem ser separados linearmente por um hiperplano, cuja posição é determinada por um processo iterativo, cujo esquema computacional é apresentado na figura 6.

O termo "máquina de aprendizagem" tem origem nos esforços iniciais, na década de 60, para se obter circuitos eletrônicos capazes de discriminar categorias. Presentemente, todos os classificadores em uso no reconhecimento de padrões, constituem-se em algorítimos computacionais. O LLM apresenta as mesmas desvantagens já apontadas no LDA. Ademais, tanto um como outro método requerem que a relação entre o número de amostras e o de variáveis seja de pelo menos 5, para uma separação entre duas categorias. Apesar disso, dados que apresentam variações aleatórias (ruído) podem ser separados em duas categorias, com essa relaçao mantida em 2^{35}.

Regra do Vizinho Mais Próximo – KNN[36]-(K-Nearest Neighbor).

A idéia básica do KNN está ilustrada na figura 7. Utiliza-se um conjunto de treinamento para distribuir os pontos entre as classes. Entretanto, nesse estágio de treinamento, não se determina nenhuma função matemática, como se fez no LDA e no LLM. Aqui, o que se faz é classificar um ponto na mesma categoria que seu vizinho mais próximo. É interessante utilizar mais que um vizinho mais próximo, pelo que as rotinas de classificação dos programas de computador determinam normalmente K vizinhos mais próximos, onde K pode variar de 1 a 10.

Existem muitas vantagens no KNN, com relação a outros métodos de reconhecimento de padrões. Por exemplo, distribuições bimodais que não são linearmente separáveis, podem ser classificadas com sucesso pelo KNN. Além disso, não é necessário aqui manter a relação entre o número de amostras e o de variáveis em 5, como nos dois casos anterio-

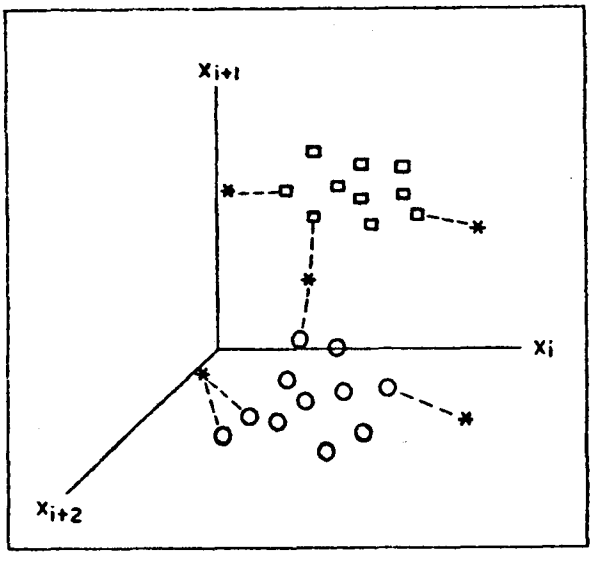

Fig. 7 – A classificação de uma amostra com categoria desconhecida é feita com base nas categorias dos seus vizinhos mais próximos.

res. Entretanto, o KNN não é capaz de sugerir que uma amostra possa pertencer a uma categoria ainda não definida.

A filosofia do KNN também é diferente do LDA e do LLM. Nesses dois métodos, o que se busca é uma separação entre categorias. O KNN, por outro lado, é um exemplo de modelo de similaridade. As amostras do conjunto de treinamento mais similares à amostra em estudo, servem como referência para sua classificação.

O problema de identificação de fontes de água mineral do estado de São Paulo, pode servir como exemplo da capacidade de classificação da regra do vizinho mais próximo. Durante o período de um ano, foram colhidas 116 amostras de água mineral das regiões de Serra Negra, Lindóia e Valinhos[37]. Em cada uma dessas regiões existem várias fontes. Entretanto, a distância entre as três regiões ($\geqslant$ 20 km), é bem superior à que existe entre as fontes de cada região ($\approx$ 2 km).

Se os ambientes geológicos das fontes dessas três regiões forem diferentes, deve ser possível identificar as fontes de água mineral, com base na concentração de metais presen-

TABELA II

Resultados da regra do vizinho mais próximo para as amostras do conjunto de treinamento e conjunto de teste.

CONJUNTO DE TREINAMENTO

Categoria[a]	Nº de amostras[b]	Nº de amostras classificadas incorretamente				
		1NN[c]	3NN	5NN	7NN	9NN
SN	46	4	4	3	3	3
L	31	2	3	4	3	3
V	39	0	5	4	6	6
Total	116	6	12	11	12	12
% de classificação correta		95	90	91	90	90

CONJUNTO DE TESTE [d]

Fonte	Embalagem	Categoria Correta	Classificação obtida				
			1NN	3NN	5NN	7NN	9NN
N. Sra. Aparecida	garrafa plástica	SN	SN	SN	SN	SN	SN
N. Sra. Aparecida	garrafa plástica	SN	SN	SN	SN	SN	SN
N. Sra. Aparecida	garrafa plástica	SN	SN	SN	SN	SN	SN
São José	garrafa plástica	L	L	L	L	L	L
São José	garrafa plástica	L	L	L	L	L	L
São José	copo plástico	L	L	L	L	L	L
São José	copo plástico	L	L	L	L	L	L
São Sebastião	garrafa de vidro	L	L	L	L	L	L

a) As categorias SN, L e V referem-se a águas minerais de Serra Negra, Lindóia e Valinhos, respectivamente.
b) Número de amostras obtido diretamente das fontes correspondentes às três categorias, durante um período de um ano.
c) KNN representa a regra de vizinho mais próximo utilizando K = 1, 3, 5, 7 e 9 vizinhos mais próximos.
d) As amostras foram obtidas em supermercados, bares e distribuidoras de água mineral. As fontes e as categorias supostamente corretas são aquelas indicadas nos rótulos das garrafas e copos.

tes nas amostras de água. Utilizando espectroscopia de emissão de plasma de argônio, foram analisados 18 elementos nas amostras de água, e suas concentrações determinadas. Apenas seis elementos se apresentaram em concentrações acima do limite de detecção em todas as amostras, e foram então utilizados para o estabelecimento de um conjunto de treinamento. Foi possível obter 90% de classificação correta das amostras das três categorias, para até K = 10 vizinhos mais próximos (Tabela II).

Como conjunto de teste foram utilizadas amostras de água mineral comercializada em garrafas de vidro e de plástico e em copos plásticos. Os resultados da classificação, mostrados na Tabela II, claramente atestam a autenticidade das amostras estudadas.

Regra de Alocação – ALLOC (Allocation Rule)[38]

Esse método é semelhante ao KNN, mas ao invés de calcular distâncias entre os pontos referentes a amostras desconhecidas e os referentes ao conjunto de treinamento, o método cria um potencial em cada ponto do espaço definido pelas variáveis, resultante da contribuição de cada ponto do conjunto de treinamento[38]. Cada amostra do conjunto de teste ocupará um ponto nesse espaço e será classificada como pertencente à categoria que apresentar maior potencial nesse ponto.

SIMCA

Entre todos os métodos de reconhecimento de padrões discutidos no Seminário de Estudos Avançados em Quimiometria, patrocinado pela OTAN em Setembro último, o SIMCA foi o que despertou maior interesse[39-42].

SIMCA é a sigla para "Soft Independent Modelling by Class Analogy". Entretanto, um título mais descritivo seria "Modelos Independentes de Similaridade Utilizando Componentes Principais". Esse método apresenta notáveis vantagens para a classificação por categorias, quando comparado aos outros métodos descritos. Baseia-se na análise de componentes principais, já discutida na seção RESOLUÇÃO.

Considere o arranjo de pontos ilustrado na figura 8. No sistema de coordenadas original, tanto a dimensão X_1 como a X_2 são importantes, uma vez que uma parte significativa da variança das amostras está representada em cada eixo. Por outro lado, os pontos se dispõem segundo uma relação linear, o que permite o uso de um modelo de componente principal, utilizando apenas um eixo capaz de descrever adequadamente a estrutura dos dados. Isso é uma forma de pré-processamento dos dados, podendo também ser entendida como um método de redução de variáveis. O novo sistema de coordenadas obtido pode ser relacionado com o original, por uma simples rotação.

A posição dos eixos do novo sistema de coordenadas é determinada através da diagonalização da matriz de covariança, como descrito anteriormente; os valores dos dados no novo sistema de coordenadas, θ, são fornecidos pelo autovetor $\vec{\beta}_j$, obtido de uma das colunas da matriz $\underset{\sim}{\beta}^t$, em:

$$\underset{\sim}{\theta} = \underset{\sim}{X}\,\underset{\sim}{\beta}^t \qquad (17)$$

Na seção RESOLUÇÃO, a matriz de dados representada por $\underset{\sim}{A}$ dizia respeito ao caso específico do uso de absorvâncias. Aqui, a matriz $\underset{\sim}{X}$ diz respeito a uma matriz de dados genérica. Para o caso bidimensional ilustrado na figura 8

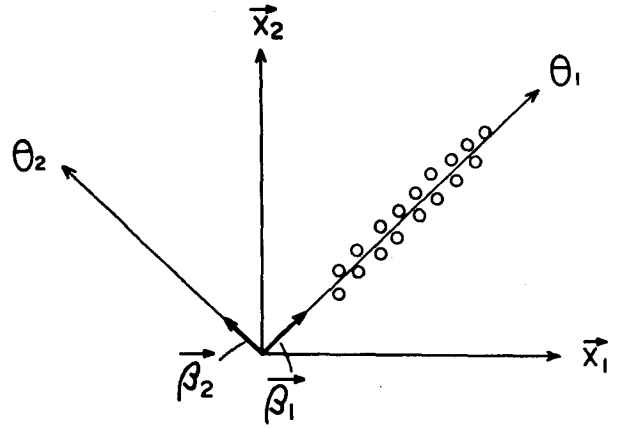

Fig. 8 – A diagonalização da matriz de covariança resulta em um novo sistema de coordenadas, especificadas pelos vetores unitários $\vec{\beta}_1$ e $\vec{\beta}_2$. O primeiro autovalor reflete quase toda a variança dos dados na direção especificada por $\vec{\beta}_1$. O valor transformado do ponto (X_1, X_2) é dado por (θ_1, θ_2).

o autovalor λ_1 é cerca de 10 vezes maior que λ_2, ou seja, cerca de 90% da variança total é expressa pelo eixo do componente principal, $\vec{\beta}_1$, sendo os restantes 10% atribuídos a $\vec{\beta}_2$. Como $\vec{\beta}_1$ e $\vec{\beta}_2$ são autovetores, devem ser mutuamente ortogonais e os eixos devem guardar entre si um ângulo de 90°.

O método SIMCA é um esquema de classificação por modelo de similaridade. Se repetirmos medidas multivariadas de um mesmo objeto, os valores obtidos para uma variável devem diferir entre si apenas pelo erro experimental, e podem ser expressos por:

$$x_{ki} = \alpha_i + e_{ki} \qquad (18),$$
$$k = 1,2,\ldots n$$
$$i = 1,2,\ldots p$$

onde α_i representa o valor médio para a variável i e e_{ki} é o erro experimental da variável i para o objeto k. Esse modelo, de pouca aplicação prática, é um modelo de zero componentes e pode ser representado no espaço p por uma hiperesfera de raio S_0, conforme se vê na figura 9. S_0 é o desvio padrão dos resíduos,

$$S_0^2 = \sum_{i=1}^{p} \sum_{k=1}^{n} e_{ki}^2 / [(p\text{-}A)(n\text{-}A\text{-}1)] \qquad (19),$$

onde o denominador da equação representa o número de graus de liberdade. Se este modelo for aplicado para objetos similares mas não idênticos e_{ki} e S_0 aumentarão porque

os resíduos conterão erros de modelagem além de erros de medida. Nesse caso é necessário um modelo mais sofisticado de um componente, a fim de descrever a estrutura mais complexa dos dados:

$$x_{ki} = \alpha_i + \theta_k \beta_i + e_{ki} \quad (20)$$

Os valores de θ e β nessa equação, compõem inicialmente as matrizes (n.l) θ e β^t. Geometricamente, o modelo pode ser representado por um hipercilindro no espaço p (figura 9). O eixo do cilindro corresponde nesse modelo ao componente principal, que é o eixo de maior variança, $\vec{\beta}_j$; o raio do cilindro é dado por S_0. A posição relativa de cada ponto ao longo do eixo do componente principal é dada pelo valor de θ.

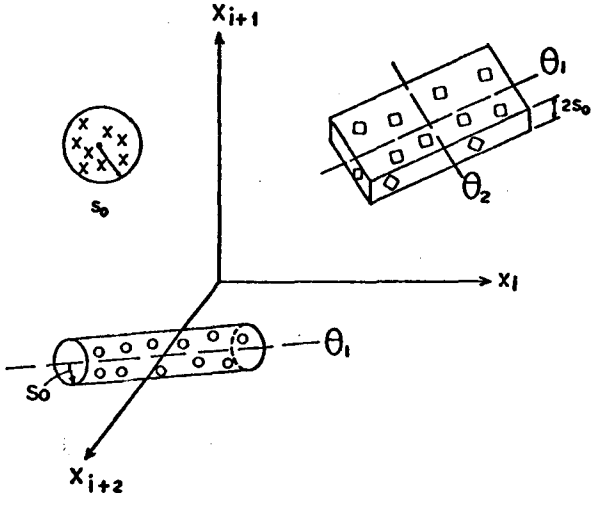

Fig. 9 – Representação geométrica dos modelos de zero (hiperesfera), um (hipercilindro) e dois (hipercaixa) componentes, utilizados na classificação SIMCA.

Para dados cuja estrutura é ainda mais complexa, é necessário um modelo de dois componentes.

$$x_{ki} = \alpha_i + \theta_{k1} \beta_{1i} + \theta_{k2} \beta_{2i} + e_{ki} \quad (21)$$
$$i = 1,2 \ldots p$$
$$k = 1,2 \ldots n$$

Essa equação é representada geometricamente na figura 9 por uma caixa no espaço p cujo comprimento e largura são determinados pelos desvios padrões dos pontos ao longo dos dois componentes principais, e a profundidade é $2\,S_0$.

Em geral, para estruturas de dados muito complexas, utiliza-se um método com A componentes.

$$x_{ki} = \alpha_i + \sum_{a=1}^{A} \theta_{ka} \beta_{ai} + e_{ki} \quad (22)$$

Essa equação representa uma hipercaixa no espaço p. O valor do objeto k projetado no eixo do componente principal a é dado por θ_{ka}, enquanto β_{ai} representa a velocidade de variação da variável i devida a variações unitárias no a-ésimo componente principal.

Os dados do conjunto de treinamento são utilizados para determinar o modelo de cada categoria. Para a determinação do número apropriado de componentes principais, pode-se utilizar técnicas de "Cross-Validation". O modelo aplicado a cada categoria é totalmente independente do das outras. Também o número de componentes necessários para descrever os dados pode variar de uma categoria para outra, dependendo do grau de complexidade da estrutura dos dados em cada caso, como se vê na figura 9.

Terminada a modelagem, resta classificar os pontos correspondentes a amostras desconhecidas. Esse processo é exemplificado na figura 10, por projeções bidimensionais de dois modelos de um componente. Os valores $Sp^{(1)}$ e $dp^{(2)}$ mostrados na figura, representam a menor distância entre um ponto de teste e o eixo do componente principal de cada categoria. Esses valores, por conveniência computacional, são determinados por técnicas de regressão linear.

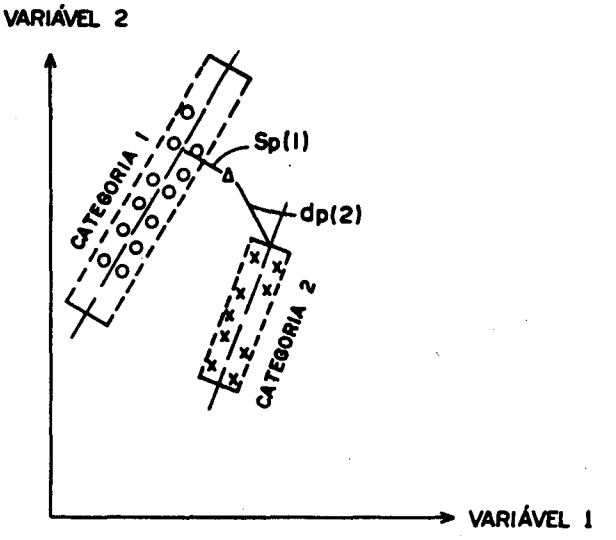

Fig. 10 – Uma projeção de dois modelos de um componente (A = 1), num plano de duas variáveis. O ponto de teste, representado pelo triângulo, pertence a uma categoria desconhecida. O modelo irá classificá-lo em uma das categorias, por comparação entre as distâncias do ponto aos componentes principais, $Sp^{(1)}$ e $dp^{(2)}$, com os desvios padrões, S_0.

Podemos agora, uma vez descrito o método SIMCA, ilustrar suas vantagens com relação aos outros métodos. O SIMCA é capaz, por exemplo, de indicar quando um ponto referente a uma amostra não pertence a nenhuma das categorias classificadas, indicando-o como um ponto deslocado (outlier) ou membro potencial de uma categoria ainda não definida. Esse é o caso mostrado na figura 11. Todos os pontos situados dentro das hipercaixas, são classificados como pertencentes à categoria definida por seu hipervolume; o ponto indicado pelo triângulo, é acusado como um ponto deslocado e não pertencente a nenhuma categoria definida. Todos os outros métodos de reconhecimento de padrões descritos, classificariam incorretamente esse ponto, como pertencente a uma das categorias definidas no conjunto de treinamento.

Outra vantagem do SIMCA é a sua aplicação a problemas do tipo ilustrado na figura 12. Os pontos (indicado por

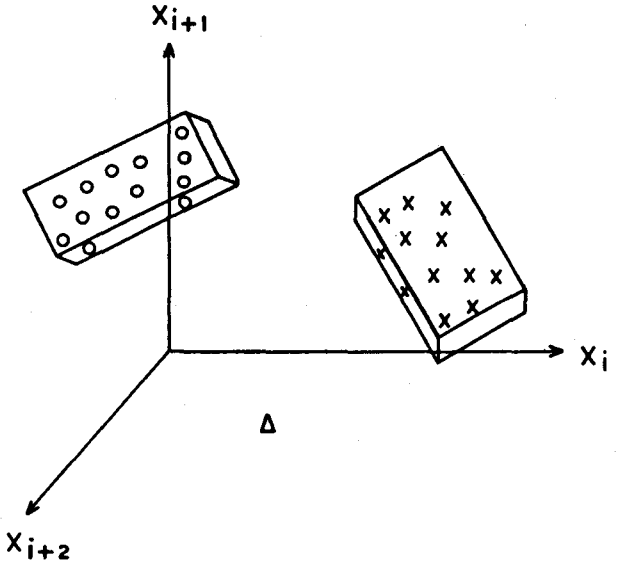

Fig. 11 – A idéia de um ponto deslocado. As amostras assinaladas com círculos e cruzes nas hipercaixas, pertencem a categoria definidas. A amostra assinalada pelo triângulo situa-se fora das hipercaixas, e é considerada como pertencente a uma terceira categoria ainda não definida. Todos os outros métodos de reconhecimento de padrões considerariam o triângulo como pertencente a uma das categorias definidas a priori.

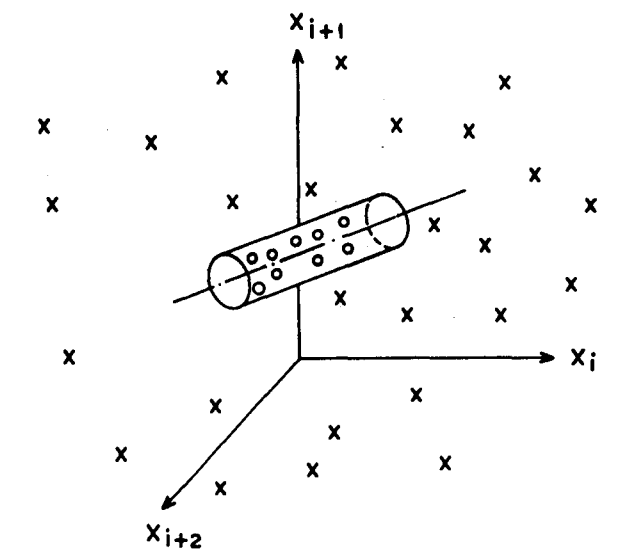

Fig. 12 – Uma categoria bem definida (representando, por exemplo, um produto industrial de alta qualidade) circundada por uma outra, espalhada no espaço p (como seria de se esperar para produtos de qualidade não controlada). O método SIMCA pode ser aplicado a esse tipo de problema assimétrico, enquanto outros métodos, exceto o KNN e o ALLOC, não podem ser aplicados.

"o") aqui representados podem ser relativos a um produto manufaturado de especificações bem definidas, ou a um produto que apresente atividade biológica específica. É razoável esperar que os produtos de boa qualidade sejam resultantes de um controle de qualidade severo, e seus pontos se disponham no espaço definido por um hipercilindro ou uma hipercaixa. Por outro lado, os produtos de má qualidade ou de baixa atividade biológica devem estar dispersos numa faixa ampla, e seus pontos (indicados por "x") muito provavelmente deverão se situar fora dos limites especificados pelo modelo do SIMCA. Nos métodos de reconhecimento de padrões descritos anteriormente, só é possível definir uma classificação quando o conjunto de treinamento apresenta duas ou mais categorias bem definidas.

Finalmente, o SIMCA pode ser aplicado a casos onde o número de amostras, n, é bem menor que o de variáveis, p, uma vez que o número de graus de liberdade, dado por (p-A) (n-A-1), aumenta tanto com o número de amostras como com o de variáveis. Conforme já vimos, o LDA e o LLM são especificamente limitados nesse sentido.

Temos encontrado grande eficiência no método SIMCA, na classificação de álcoois primários, secundários e terciários, com base nas intensidades e freqüências de seus espectros, no infravermelho. Há um considerável interesse em saber se as freqüências ou as intensidades contêm maior informação relevante na classificação desses álcoois. As tabelas de freqüências características no infravermelho têm sido muito utilizadas em análise orgânica qualitativa; entretanto, pouco tem sido feito no sentido de relacionar o comportamento de intensidades com o tipo de molécula estudado.

Dados de espectro infravermelho para 33 álcoois alifáticos (10 primários, 11 secundários e 12 terciários) foram obtidos da biblioteca Sadtler e digitalizados manualmente. Foram utilizadas como variáveis, oito freqüências e quatro intensidades relativas. A matriz de dados obtidos foi portanto de dimensão 33 x 12. O método SIMCA aplicado a essa matriz de dados, forneceu 97% de classificação correta. Os pesos de Fisher indicam que os valores de freqüência são extremamente importantes na separação dos álcoois terciários das outras duas categorias. Em particular, as freqüências que mais pesam nessa separação estão na região de deformação -OH e do estiramento -OH. Os dados relativos a intensidade, a despeito de sua importância secundária nesse exemplo, contribuem na separação dos álcoois primários e secundários.

Relação Entre Dois Blocos. Considere o caso de termos duas matrizes de dados, $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ como se vê na figura 13. É interessante pesquisar alguns fatores interblocos, ou tendências principais de variações que sejam comum aos dois blocos de variáveis, desprezando-se os fatores não comuns, específicos do bloco $\underset{\sim}{X}$ ou $\underset{\sim}{Y}$, onde se inclui os erros experimentais aleatórios característicos de cada variável de um bloco. Sob condições ideais, e considerando as variáveis do bloco Y como dependente, é possível prever o valor dessas variáveis, em função das variáveis independentes do bloco $\underset{\sim}{X}$. A tabela II traz vários exemplos de blocos $\underset{\sim}{X}$ e $\underset{\sim}{Y}$.

Se a matriz de dados $\underset{\sim}{Y}$ contém apenas uma variável (por exemplo, a atividade biológica de uma droga) e o bloco $\underset{\sim}{X}$ apresenta um conjunto multivariado (parâmetros físico-químicos da droga, por exemplo), o primeiro impulso que se tem é o de tentar uma regressão linear múltipla para se obter Y a partir dos valores de X, como:

$$y = b_1x_1 + b_2x_2 + \ldots\ldots b_px_p \qquad (23)$$

TABELA III

Alguns exemplos de relacionamento entre variáveis, nos Blocos $\underset{\sim}{X}$ e $\underset{\sim}{Y}$

Bloco $\underset{\sim}{X}$	Bloco $\underset{\sim}{Y}$
Concentrações totais de metais em amostras de solos.	Concentrações de metais solúveis nas amostras de solos.
Intensidades espectrométricas relativas às concentrações.	Concentrações químicas para multicomponentes determinados usando análises clássicas.
Concentrações de agronutrientes em solos.	Concentrações desses nutrientes em plantas cultivadas nesses solos.
Valores Teóricos calculados para energia, momento de dipolo, parâmetros espectroscópicos, etc.	Parâmetros moleculares experimentais (energia, momento de dipolo, parâmetros espectroscópicos, etc.).
Concentrações de espécies que mostrem ser importantes na determinação da qualidade dos alimentos.	Indicadores multivariados de qualidade de alimentos (sabor, textura, doçura, aspecto, etc.).

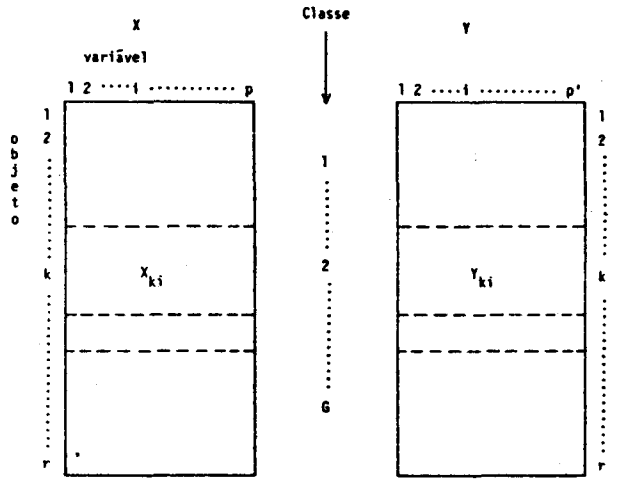

Fig. 13 – A matriz $\underset{\sim}{Y}$ é montada de tal forma que os objetos apresentem a mesma ordem que na matriz $\underset{\sim}{X}$. Assim, todos os objetos da categoria 1 são seguidos de todos os da categoria 2, e assim por diante. As variáveis dos blocos $\underset{\sim}{Y}$ e $\underset{\sim}{X}$ são normalmente consideradas dependentes e independentes, respectivamente.

Entretanto, a regressão múltipla só pode ser rigorosamente aplicada quando as variáveis X forem ortogonais entre si. Do ponto de vista teórico, a melhor solução é uma análise de componentes principais com os dados do bloco $\underset{\sim}{X}$ e uma regressão para Y pelos valores do componente principal

$$y = b'_1\theta_1 + b'_2\theta_2 + \ldots\ldots b'_a\theta_a \qquad (24)$$

Os valores de θ não são apenas mutuamente ortogonais, como também apresentam variança maximizada, permitindo a utilização de um número menor de variáveis ($a < p$), sem perda de informação estatística significativa, o que parcialmente resolve a difícil questão de decidir quais variáveis independentes devem ser incluídas na regressão. Se o bloco $\underset{\sim}{Y}$ contiver mais que uma variável, a análise de componentes principais deverá ser efetuada para os dados da matriz $\underset{\sim}{X}$ e da matriz $\underset{\sim}{Y}$. A relação entre os componentes principais de cada bloco será ilustrada na figura 14.

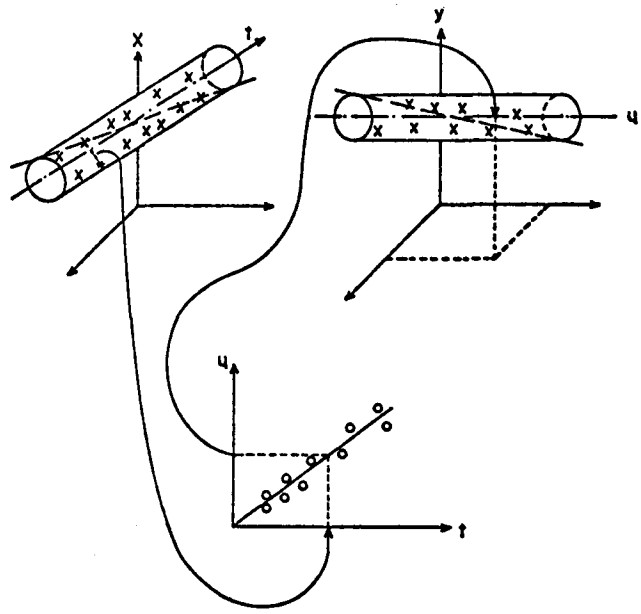

Fig. 14 – Na fase de treinamento, são montados modelos semelhantes ao de componentes principais, para descrever as classes $\underset{\sim}{X}$ e Y. Seus eixos são levemente deslocados, para melhorar a correlação entre os valores de t e u.
Na fase de teste ou de previsão, o novo objeto (asterisco) é classificado ou não numa categoria, de acordo com sua posição dentro ou fora do intervalo de tolerância da categoria, no espaço $\underset{\sim}{X}$. A posição do objeto na parte X do modelo, ou seja, ao longo do eixo t, é utilizada na curva u x t, fornecendo uma previsão do valor de u para o objeto. Esse valor é utilizado no espaço Y para fornecer previsões dos valores das variáveis y.

Suponha a análise SIMCA realizada separadamente nas matrizes $\underset{\sim}{X}$ e $\underset{\sim}{Y}$. Cada categoria de dados é agora tratada independentemente. Por razões didáticas, vamos ainda supor que os dados de $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ sejam adequadamente descritos por modelos de um componente. Dessa forma, os conjuntos de dados de $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ podem ser representados por hipercilindros no espaço p, como se vê na figura 14. Os eixos dos componentes principais estão indicados pela linha pontilhada, e não coincidem com o eixo do cilindro, por razões que serão apontadas mais adiante. As equações do método SIMCA para os dois blocos são:

$$x_{ki} = \alpha_i + \sum_{a=1}^{A} \theta_{ka}\, \beta_{ai} + e_{ki} \qquad (25)$$

e

$$y_{kj} = \alpha_j + \sum_{a=1}^{A} \eta_{ka}\, \beta_{aj} + e_{kj} \qquad (26)$$

No nosso exemplo simplificado, A = 1 e podemos então fazer um gráfico dos valores de θ_{ka} contra η_{ka}. Em casos favoráveis podemos esperar uma relação linear entre os valores de η_{ka} e θ_{ka}, o que nos permite relacionar os dados dos blocos $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ ou seja, podemos obter valores de η por regressão dos dados de θ.

O método dos mínimos quadrados parciais (PLS-partial least squares) proposto por H. Wold[43], permite a relação interblocos. Os dados do bloco $\underset{\sim}{X}$ são ponderados a partir das informações obtidas do bloco $\underset{\sim}{Y}$ e vice-versa, pelo que os eixos originais dos componentes principais são ligeiramente inclinados, maximizando a correlação entre os componentes dos dois blocos. Por essa razão o eixo do cilindro da figura 14 não coincide com o eixo original do componente principal. Os valores das projeções dos dados de $\underset{\sim}{X}$ e $\underset{\sim}{Y}$ sobre os eixos dos cilindros, são designados por t e u, respectivamente.

Considere agora um objeto para o qual se conhece as variáveis X, mas não Y, (como por exemplo o indicado por*, na figura 14). As variáveis Y podem ser estimadas da maneira que segue. Um ponto relativo a um dado de X é projetado sobre o eixo do cilindro em seu próprio espaço, fornecendo um valor de t. Esse valor é transposto para o gráfico de u contra t (η contra θ, antes da rotação dos eixos), de onde se obtém o valor de u correspondente ($u_{ka} = d_a\, t_{ka} + e_k$). Uma vez localizado o valor de u no eixo do componente principal do bloco $\underset{\sim}{Y}$, pode-se prever os valores das variáveis y correspondentes, por projeção sobre os eixos originais.

CONHECIMENTO DE PADRÕES

Nem sempre é possível utilizar um conjunto de treinamento como referência para resolver um problema de classificação. Pode ocorrer, por exemplo, que não se disponha de informações necessárias à classificação, tais como a origem da amostra e outras. Mesmo assim, pode ser importante determinar quais amostras são mais semelhantes entre si. Existem métodos, baseados no conceito de similaridade já mencionado, capazes de realizar a classificação sem utilizar um conjunto de treinamento; são os métodos de aprendizagem não supervisionada.

Métodos do tipo do SIMCA podem ser utilizados nesses casos. Pode-se, por exemplo, utilizar uma projeção da análise de componentes principais no espaço p, num espaço de duas ou três dimensões. Por simples inspeção visual, podemos identificar os arranjos dos dados em grupos distintos. Fazemos então uma classificação tentativa dos dados, que pode a seguir ser verificada pelo método SIMCA.

Há no entanto métodos aplicáveis apenas a problemas de aprendizagem não supervisionada. Um exemplo são as técnicas de agrupamento hierárquico[30].

É muito comum a utilização da equação 14 como uma medida de similaridade. Isto é feito por inspeção dos dados da matriz de similaridade*, até se encontrar dois pontos de similaridade máxima (menor distância), no espaço p. Esses dois pontos são eliminados e substituídos pelo seu centro de gravidade. A matriz de similaridade é reduzida em uma dimensão, e então se procura pelo próximo par de pontos com a mais alta similaridade. Há várias maneiras de tratar o centro de gravidade já obtido, mas matematicamente ele é simbolizado como um outro dado. O processo continua até que se tenha obtido um grande agrupamento de dados.

Talvez a idéia fique mais clara observando-se a figura 15, que se refere a 30 amostras de água mineral obtidas de seis fontes distintas na cidade de Lindóia[23,37]. Cada linha vertical à base refere-se a uma amostra; a ordenada apresenta valores de similaridade calculados pela equação 14.

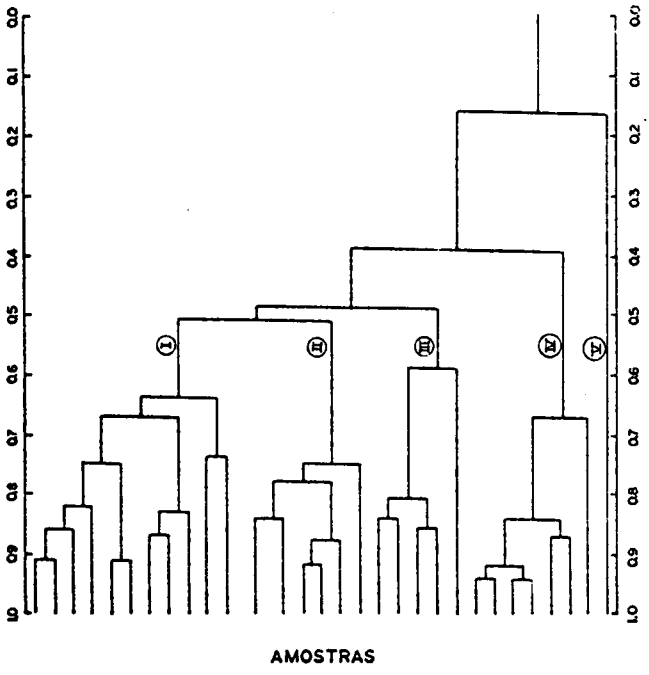

Fig. 15 – Diagrama correspondente às amostras de água mineral da cidade de Lindóia.

* (Veja a equação 14).

Deve-se notar que a amostra V é muito diferente das outras, no diagrama. Seu valor de similaridade é de ~ 0,15 com relação ao centro de gravidade dos outros pontos**. O grupo de pontos IV tem um valor de similaridade de ~0,4, relativamente a outros grupos. É interessante que todas as amostras obtidas de uma fonte, a fonte São Jorge, estão no grupo IV. Do mesmo modo todas as amostras no grupo III são da fonte BC 16A, e nenhuma amostra dessa fonte aparece em outro grupo. O valor de similaridade desse grupo é de ~0,5, com relação aos grupos I e II.

A medida que aumentam os valores de similaridade, torna-se mais difícil relacionar os grupos matemáticos com fatores reais (fontes de água mineral, nesse caso).

Entretanto, estudos mais detalhados sobre as águas minerais da região de Lindóia, utilizando espectroscopia de emissão de plasma, permitiram compreender um pouco melhor a hidrologia das reservas de águas minerais.

Outro método de estudo de agrupamentos é o da árvore de varredura mínima. Enquanto o agrupamento hierárquico funciona bem para conglomerados de pontos, o método de varredura é excelente para agrupar conjuntos de dados de estrutura laminar. O método simula a construção de uma árvore, ligando cada ponto ao seu vizinho mais próximo de forma que os ramos da árvore sejam mínimos. O valor médio das distâncias entre os pontos é utilizado como um valor limite para a separação dos grupos. A árvore é partida em grupos nas posições onde a distância entre dois pontos seja de duas ou mais vezes o valor do limite.

Finalmente, vale mencionar o método CLUPOT (*Clustering with Pot*entials)[45], que faz uso da superfície de potencial gerada pelo método ALLOC. O agrupamento dos dados é estudado a partir de atenuações da função do potencial.

CONSIDERAÇÕES FINAIS

A química não é imune a influências exteriores, tais como a revolução de minicomputadores que se verifica na maioria dos países do primeiro mundo. O uso de computadores pessoais para uso doméstico, como o controle de contas bancárias e outras atividades, tem sido considerado um luxo no conceito do terceiro mundo, sem que isso traga grandes prejuízos nesse nível. Entretanto, os laboratórios químicos, que exigem maior sofisticação computacional, não se podem permitir negligenciar as inovações tecnológicas nesse campo, que permitem não só a resolução de problemas práticos como uma compreensão mais profunda dos conceitos da Química.

AGRADECIMENTOS

Um de nós (REB) agradece a FAPESP pelo apoio financeiro para participar no "NATO Advanced Study Institute on Chemometrics" em Cosenza, Itália.

** Trata-se provavelmente de um ponto deslocado, que poderia ser eliminado por teste estatístico.

Referências:

[1] The Chemometrics Society, Prof. D.L. Massart, Urije Universeteit Brussel, Farmaceutisch Institut, Taarbeeklaan 103, Brussel, Bélgica.

[2] Varmuza K., "Lecture Notes in Chemistry No. 21 Pattern Recognition in Chemistry", G. Berthier e outros, eds Springer Verlag, Berlim, 1980.

[3] a) Kowalski, B.R., Anal. Chem. 52 112R (1980);
b) Frank I.E., e Kowalski, B.R. Anal. Chem. 54 232R (1982).

[4] B.E.H. Saxberg e B.R. Kowalski, Anal. Chem., 51, 1031 (1979).

[5] a) Kateman, G., Pipers, F.W. "Quality Control in Analytical Chemistry, Wiley-Interscience, New York, 1981
b) Taylor, J.K. Anal., Chem. 53 1655A 1981
c) Eckschlager, K, Stepanek, V. "Information Theory as Applied to Chemical Analysis", Wiley-Interscience, New York, 1979.
d) Malinowski, E.R., Howery, D.G., "Factor Analysis in Chemistry", Wiley-Interscience, New York, 1980.

[6] Currie, L.A., Anal. Lett. 13 1 (1980).

[7] K. Fukunaga e W.L.G. Koontz, IEEE Transactions on Computers C 19 311 (1970).

[8] T.H. Wonnacott e R.J. Wonnacott, Introductory Statistics, Wiley, New York (1969).

[9] E.R. Malinowski, D.G. Howery, P.H. Weiner, J.H. Soroka, P.T. Funke, R.S. Selzer e A. Livingstone, Quantum Chemistry Program Exchange, Program 320 (1976).

[10] Jochum, C., Jochum, P., Kowalski, B.R., Anal. Chem. 53 85 (1981).

[11] Kalivas, J.H. e Kowalski, B.R., Anal. Chem., 53 2207 (1981).

[12] Gerlach, R.W. e Kowalski, B.R., Anal. Chim. Acta, 133 119 (1981).

[13] A.O. Jacintho, E.A.G. Zagatto, H. Bergamin Fº, F.J. Krug, B.F. Reis, R.E. Bruns e B.R. Kowalski, Anal. Chim. Acta, 130, 243 (1981).

[14] E.A.G. Zagatto, A.O. Jacintho, F.J. Krug, B.F. Reis R.E. Bruns e M.C.U. Araujo, Anal. Chim. Acta 145 169 (1983).

[15] J.H. Kalivas, Anal. Chem. 55 567 (1983).

[16] J.H. Kalivas e B.R. Kowalski, Anal. Chem. 55 532 (1983).

[17] J.H. Kalivas e B.R. Kowalski, Anal. Chem. 54 560 (1982).

[18] I.E. Frank, J.H. Kalivas e B.R. Kowalski, Anal. Chem. 55 1800 (1983).

[19] O programa computacional pode ser obtido de Infometrix, Inc., P.O. Box 25888, Seattle, Washington 98125, USA.

[20] Wold, S. e Dunn III, W.J., J. Chem. Info. and Comp. Sci. 23 6 (1983).

[21] W.V. Kawn e B.R. Kowalski, J. Food Science, 43 1320 (1978).

[22] D.L. Duewer e B.R. Kowalski, Anal. Chem. 47 1573 (1975).

[23] I.S. Scarminio, R.E. Bruns e E.A.G. Zagatto, Energia Nuclear e Agricultura 4 99 (1982).

[24] B.R. Kowalski, T.F. Schatzki e F.H. Stross, Anal. Chem. **44** 2176 (1972).
[25] M.L. Bisani, D. Faraone, S. Clementi, K.E. Esbenson e S. Wold, Anal. Chim. Acta **150** 129 (1983).
[26] B.E.H. Saxburg, D.L. Duewer, J.L. Booker e B.R. Kowalski, Anal. Chim. Acta **103** 201 (1978).
[27] B.R. Kowalski, Anal. Chem. **47**, 1152A (1975).
[28] Kowalski, B.R., Chemtech **1974** 300.
[29] Duewer, D.L. e Kowalski, B.R., Anal. Chem. **47** 526 (1975).
[30] Kowalski, B.R. e Bender, C.F., J. Am. Chem. Soc. **94** 5632 (1972).
[31] O programa computacional ARTHUR pode ser obtido de Infometrix, Inc. P.O. Box 25888 Seattle, WA. 98125.
[32] O programa SIMCA-3B pode ser obtido de Prof. S. Wold, Institute of Chemistry, Umeå University S-90187, Umeå, Suécia ou Dr. David Stalling, Columbia National Fisheries Research Laboratory, U.S. Fish and Wildlife Service, Route 1, Columbia, MO 65201, EUA.
[33] Fisher, R.A., Ann. Eugen. 7, 179 (1936).
[34] Nilsson N.J., "Learning Machines, Mc Graw-Hill, New York 1965.
[35] Meisel, W.S. "Computer Oriented Approach to Pattern Recognition", Academic Press, New York (1972).
[36] Cover, T.M. e Hart, P.E., IEEE Trans. Info. Theory, **IT-13** 21 (1967).
[37] Ieda S. Scarmino, Tese de Mestrado, Universidade Estadual de Campinas, (1981).
[38] D. Coomans, D.L. Massart, I. Broeckaert, e A. Tassin, Anal. Chim. Acta., **133**, 216 (1981).
[39] S. Wold, Pattern Recognition 8 127 (1976).
[40] S. Wold e M. Sjölström, ACS Symposium Series, **52**, 245 (1977).
[41] Albano, C., Blomqvist, G., Coomans, D., Dunn III, W.J., Edlund, U., Eliasson, B., Hellberg, S., Johansson, E., Nordén, Bo, Johnels, D., Sjöström, M., Söderström, B., Wold, H. e Wold, S., Proceedings of the Symposium on Applied Statistics (A. Hoskuldsen et al. Eds.) **1981**, 183.
[42] Wold, S., Albano, C. Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E. e Sjöström, em "Data Analysis in Food Research, H. Martens e Russwurm, eds., Applied Science Publishers Ltd., London, (1983).
[43] Wold, H. em "Systems Under Indirect Observation", Jöreskog, K.G. e Wold, H., eds., North-Holland, Amsterdam, 1982.
[44] Andrews, H.C., "Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, 1972.
[45] Coomans, D. e Massart, D.L. Anal. Chim. Acta, 133 226 (1981).

ARTIGO

QUÍMICA DE ZEÓLITAS E CATÁLISE

Shantappa Sidramappa Jewur

Departamento de Química
Universidade Federal do Rio Grande do Norte
59000 – Natal – RN – Brasil

Recebido em: 26/07/84

RESUMO

As zeólitas são aluminossilicatos hidratados que contêm metais alcalinos e alcalinos terrosos. Existem vários tipos de zeólitas naturais, com composições químicas e estruturas cristalográficas variáveis. Desde 1955 vêm sendo feitas tentativas no sentido de preparar zeólitas sintéticas. Cerca de 15 zeólitas já foram sintetizadas, caracterizadas e testadas para atividades catalíticas. As zeólitas sintéticas apresentam uma série de vantagens sobre as naturais em diversas aplicações industriais. Elas estão sendo bastante usadas como catalisadores, dessecantes e purificadores de gases e líquidos. As sínteses das zeólitas A, X, Y, ZSM-5 e mordenita e suas respectivas formas modificadas abriram novos caminhos para a tecnologia catalítica. Os métodos de síntese e as naturezas dos reagentes usados na obtenção de zeólitas não foram padronizados. O presente artigo descreve a química das zeólitas naturais e das sintéticas incluindo suas aplicações em catálise heterogênea industrial.

1. INTRODUÇÃO

Os aluminossilicatos que ocorrem na natureza são classificados como feldspatos, feldspatóides, escapolitas e zeólitas. O grupo de zeólitas pode ser distinguido dos demais grupos de aluminossilicatos pela presença das moléculas de água e também por causa das suas estruturas abertas características. A existência das zeólitas naturais passou a ser conhecida após a descoberta da zeólita denominada estilbita, em 1756, por Cronsted[1].