QUIMIOMETRIA I: CALIBRAÇÃO MULTIVARIADA, UM TUTORIAL

Márcia M. C. Ferreira, Alexandre M. Antunes, Marisa S. Melgo e Pedro L. O. Volpe
Departamento de Físico-Química - Instituto de Química - Universidade Estadual de Campinas - UNICAMP - 13081-970 - Campinas - SP

Recebido em 13/3/98; aceito em 12/1/99

CHEMOMETRICS I: MULTIVARIATE CALIBRATION, A TUTORIAL. The aim of this work is to present a tutorial on Multivariate Calibration, a tool which is nowadays necessary in basically most laboratories but very often misused. The basic concepts of preprocessing, principal component analysis (PCA), principal component regression (PCR) and partial least squares (PLS) are given. The two basic steps on any calibration procedure: model building and validation are fully discussed. The concepts of cross validation (to determine the number of factors to be used in the model), leverage and studentized residuals (to detect outliers) for the validation step are given. The whole calibration procedure is illustrated using spectra recorded for ternary mixtures of 2,4,6 trinitrophenolate, 2,4 dinitrophenolate and 2,5 dinitrophenolate followed by the concentration prediction of these three chemical species during a diffusion experiment through a hydrophobic liquid membrane. MATLAB software is used for numerical calculations. Most of the commands for the analysis are provided in order to allow a non-specialist to follow step by step the analysis.

Keywords: chemometrics; multivariate calibration; PLS; PCR; liquid membranes.

INTRODUÇÃO

O uso de computadores para analisar dados químicos cresceu drasticamente nos últimos vinte anos, em parte devido aos recentes avanços em "hardware" e "software". Por outro lado, a aquisição de dados principalmente na área de química analítica, atingiu um ponto bastante sofisticado com o interfaceamento de instrumentos aos computadores produzindo uma enorme quantidade de informação, muitas vezes complexa e variada.

Uma das características mais interessantes dos modernos instrumentos é o número das variáveis que podem ser medidas em um única amostra. Um exemplo notável é a intensidade de absorção em mil ou mais comprimentos de onda que é rotineiramente registrada em um único espectro. De posse de tal quantidade de dados, a necessidade de ferramentas novas e mais sofisticadas para tratá-los e extrair informações relevantes cresceu muito rapidamente, dando origem à Quimiometria, que é uma área especificamente destinada à análise de dados químicos de natureza multivariada.

A quimiometria não é uma disciplina da matemática, mas sim da química, isto é, os problemas que ela se propõe a resolver são de interesse e originados na química, ainda que as ferramentas de trabalho provenham principalmente da matemática, estatística e computação. Como citado por Kowalski[1] "as ferramentas quimiométricas são veículos que podem auxiliar os químicos a se moverem mais eficientemente na direção do maior conhecimento". Isto nos leva a uma definição formal de quimiometria: "... uma disciplina química que emprega métodos matemáticos e estatísticos para planejar ou selecionar experimentos de forma otimizada e para fornecer o máximo de informação química com a análise dos dados obtidos"[1].

Experimentos envolvendo a análise espectrofotométrica quantitativa de amostras com muitos componentes cujos espectros sejam superpostos são bastante importantes em disciplinas de química analítica, sejam elas básicas ou mais avançadas. Em geral, as concentrações dos compostos de interesse numa amostra são determinadas através da resolução de um sistema de equações simultâneas obtido pela lei de Beer em tantos comprimentos de onda quantos forem os analitos. Curvas de calibração são construídas em cada comprimento de onda a partir de soluções padrão de cada analito a fim de estabelecer constantes de proporcionalidade individuais entre concentração e intensidade de absorção. No caso de misturas binárias simples, muitas vezes obtemos bons resultados por este método. Entretanto, quando se passa para amostras reais, podem surgir problemas devido a interferências espectrais e desconhecimento da real identidade dos compostos de interesse. Nestas situações, a resolução simultânea das equações já não fornece resultados precisos e por isso foram buscados novos métodos para resolver este tipo de problema.

Atualmente a quimiometria já é suficientemente estabelecida e de uso disseminado para que justifique sua introdução em cursos regulares da área de química. Neste sentido, seria interessante adaptar alguns experimentos para que métodos quimiométricos possam ser introduzidos em cursos de graduação.

Neste tutorial são usados espectros de misturas ternárias registrados na região UV-Vis para as etapas de modelagem e validação do modelo de calibração. Como aplicação, dados experimentais provenientes do transporte das mesmas espécies através de uma membrana líquida hidrofóbica são usados para a etapa de previsão. Tais membranas têm sido utilizadas como modelos de membranas biológicas, para o estudo de mecanismos de transporte e separação de espécies químicas carregadas, sendo de grande utilidade em processos industriais para a remoção de cátions a partir de soluções bastante diluídas de diferentes cátions metálicos[2]. Este é, portanto, um exemplo muito interessante do ponto de vista químico e adequado do ponto de vista pedagógico.

Para a análise de dados os métodos multivariados são os mais adequados, porque permitem um estudo com várias espécies presentes ao mesmo tempo, não importando a existência ou ausência de diferenças espectrais marcantes entre elas nem a existência de alta correlação nos dados. É possível, também, a identificação de problemas eventuais com linha base ou interferentes nas amostras usadas na calibração e nas novas amostras, de previsão[3].

O objetivo deste trabalho é despertar a motivação dos alunos em nível de graduação/pós-graduação para o uso da análise multivariada em química, utilizando para isto um exemplo

marcia@iqm.unicamp.br

simples mas de grande interesse em aplicações. O software MATLAB™ foi escolhido por oferecer um ambiente computacional amigável e de alto nível, especialmente na área de álgebra linear, que tem sido usado amplamente e com muito sucesso em quimiometria e em diversas outras áreas.

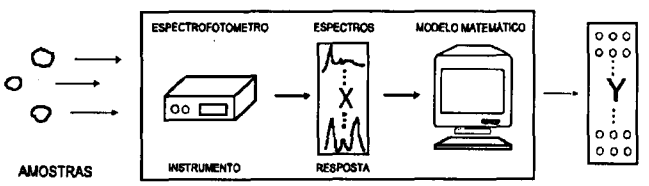
Softwares de domínio público e de "performance" semelhante ao MATLAB tais como SciLab podem ser obtidos gratuitamente através do endereço:

http://www-rocq.inria.fr/scilab/

MÉTODOS

Com grande frequência, o objetivo de uma análise química é determinar a concentração de um determinado analito em amostras, isto é, um composto químico de interesse presente nas mesmas. Os instrumentos de laboratório não produzem diretamente as concentrações como resposta. Por exemplo um espectrofotômetro registra absorbâncias que naturalmente dependem das concentrações dos analitos.

Calibração é o procedimento para encontrar um algoritmo matemático que produza propriedades de interesse a partir dos resultados registrados pelo instrumento. O resultado de uma calibração pode ser representado pelo diagrama abaixo



onde o instrumento associado ao algoritmo funciona como se fosse um novo "instrumento". Absorbância e concentração são apenas exemplos das inúmeras possibilidades que este procedimento oferece para obtenção de informações de interesse a partir de sinais instrumentais. Um exemplo importante em outra área - mas que utiliza as mesmas idéias - é a tomografia, onde sinais de radiação são transformados em imagens.

Uma vez encontrado, este algoritmo poderá ser usado para prever a concentração do componente químico de interesse em amostras de composição desconhecida, usando a resposta instrumental das mesmas. Neste tutorial, os termos espectro e concentração serão genericamente utilizados para se referir às medidas instrumentais e propriedade das amostras respectivamente.

Os espectros, um para cada amostra, são organizados numa matriz, **X** (n x m), de variáveis independentes,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

onde cada linha representa uma amostra e contém as respostas medidas para a mesma.

Por exemplo, para a iésima amostra, x_{i1}, x_{i2}, ..., x_{im} são as absorbâncias registradas, uma para cada comprimento de onda. Cada coluna de **X** corresponde a um comprimento de onda específico. Portanto, x_{1j}, x_{2j}, ..., x_{nj} são as respostas na jésima frequência para as amostras *1, 2, ..., n*. O outro conjunto de dados é constituído das variáveis dependentes e organizado na matriz **Y**, caso haja mais de uma variável (mais de um analito de interesse), ou pelo vetor **y**, no caso de uma única variável. O total de elementos deste vetor é igual a *n*, isto é, o número de amostras, e corresponde às concentrações de um determinado analito de interesse ou alguma outra propriedade que se espera prever no futuro.

O processo geral de calibração[3,4] consiste de duas etapas: MODELAGEM, que estabelece uma relação matemática entre **X** e **Y** no conjunto de calibração e a VALIDAÇÃO, que otimiza a relação no sentido de uma melhor descrição do analito(s) de interesse. Uma vez concluída a calibração, o sistema (instrumento físico + modelo matemático) representado esquematicamente no diagrama acima está apto a ser utilizado para previsão em outras amostras. Neste trabalho usa-se as onze primeiras soluções da tabela 1 como conjunto de calibração. Aplica-se o método de validação cruzada para validação do modelo, bem como as quatro últimas soluções (12 a 15) da Tabela 1 como conjunto-teste. O modelo validado foi aplicado para prever as concentrações dos três analitos em um experimento de transporte através de membranas.

Tabela 1. Misturas usadas na preparação das curvas de calibração multivariada*.

Mistura	[2,4-dnf] x10^5 (mol L^{-1})	[picrato] x10^5 (mol L-1)	[2,5-dnf] x10^5 (mol L^{-1})
01	2,0	1,0	2,0
02	1,5	1,0	1,5
03	1,0	0,8	2,0
04	2,0	0,8	1,0
05	1,0	0,5	0,5
06	0,5	0,5	1,0
07	0,8	0,8	0,8
08	0,4	0,6	0,8
09	0,8	0,6	0,4
10	1,5	1,2	1,0
11	1,4	1,0	1,2
12	1,0	1,0	1,0
13	1,2	0,7	0,8
14	1,5	0,8	1,0
15	1,0	0,7	1,2

*As onze primeiras amostras são do conjunto de calibração e as quatro últimas, do conjunto-teste.

Os dados experimentais originais podem não ter uma distribuição adequada para a análise, dificultando a extração de informações úteis e interpretação dos mesmos. Nestes casos, um preprocessamento[3,5,6] nos dados originais pode ser de grande valia. Medidas em diferentes unidades e variáveis com diferentes variâncias são algumas das razões que levam a estes problemas. Os métodos de preprocessamento mais utilizados consistem basicamente em centrar na média ou autoescalar os dados. No primeiro caso, calcula-se a média das intensidades para cada comprimento de onda e subtrai-se cada intensidade do respectivo valor médio. Autoescalar significa centrar os dados na média e dividí-los pelo respectivo desvio padrão[a], sendo um para cada comprimento de onda. No caso de calibração de dados espectroscópicos, recomenda-se centrar os dados na média.

Os comandos para ambos os preprocessamentos na linguagem do MATLAB (ou SciLab) são muito simples e utilizam funções previamente definidas. Para centrar os dados na média usam-se os comandos

```
[n,m] = size(X);
Xm = mean(X);
Xcm = X-ones(n,1)*Xm;
```

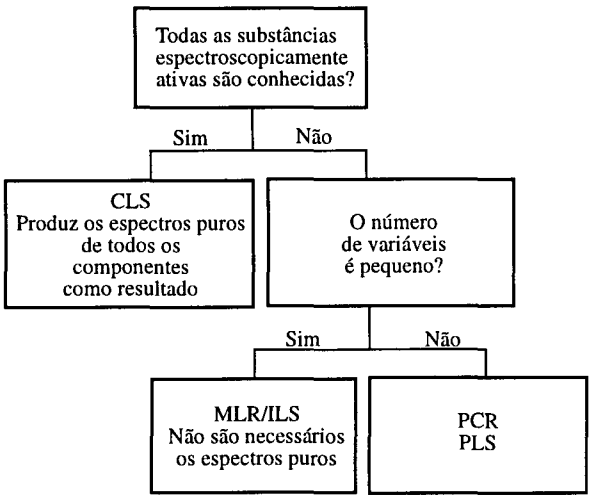
onde **Xm** é o vetor das médias das colunas de **X**, e **Xcm** é a matriz resultante dos dados centrados na média. O sinal ";" nas expressões acima é usado para indicar que as respostas não devem ser mostradas na tela do computador.

Para autoescalar os dados são usados os seguintes comandos

```
Xstd = std(X);
Xa = Xcm./(ones(n,1)*Xstd);
```

onde **Xstd** é o vetor contendo os desvios padrões das colunas de **X**, e **Xa** é a matriz dos dados originais autoescalados. As expressões "mean" e "std" são funções internas do software que calculam as médias e desvios padrões[a] das colunas de **X**. A disponibilidade destas (e inúmeras outras) funções internas otimizadas é que caracterizam um ambiente computacional como sendo de alto nível e torna conveniente o seu uso.

O fluxograma abaixo mostra um esquema geral dos métodos de calibração.



Os métodos tradicionais de calibração CLS (método clássico de mínimos quadrados) e MLR (regressão linear múltipla), têm suas vantagens e desvantagens quando aplicados a problemas químicos. Ambos utilizam toda a informação contida na matriz de dados **X** para modelar a concentração, isto é, toda a informação espectral, incluindo informações irrelevantes (fazem pequena remoção de ruído). O CLS tem como principal problema a necessidade de se conhecer as concentrações de cada espécie espectroscopicamente ativa no conjunto de calibração, o que em geral é impossível nos problemas práticos. Já o método MLR sofre do problema de colinearidade: o número de amostras deve exceder o número de variáveis, que por sua vez devem fornecer predominantemente informação única. Temos neste caso a opção de selecionar um certo número de variáveis que seja menor que o número de amostras e que produzam informação "única", o que pode ser demorado e tedioso. Mais interessante, então, seria a utilização de algum método que, como o CLS, use o espectro inteiro para análise, e como o MLR, requeira somente a concentração do analito de interesse no conjunto de calibração.

Dois métodos que preenchem estes requisitos são PCR (Principal Component Regression) e PLS (Partial Least Squares)[6,7], daí a escolha dos mesmos neste tutorial. Estes dois métodos são consideravelmente mais eficientes para lidar com ruídos experimentais, colinearidades e não linearidades. Todas as variáveis relevantes são incluídas nos modelos via PCR ou PLS, o que implica que a calibração pode ser realizada eficientemente mesmo na presença de interferentes, não havendo necessidade do conhecimento do número e natureza dos mesmos. Os métodos PCR e PLS são robustos, isto é, seus parâmetros praticamente não se alteram com a inclusão de novas amostras no conjunto de calibração[6,7]. Em especial o método PLS tem se tornado uma ferramenta extremamente útil e importante em muitos campos da química, como a físico-química, a química analítica, a química medicinal, ambiental e ainda no controle de inúmeros processos industriais.

A base fundamental da maioria dos métodos modernos para tratamento de dados multivariados é o PCA[5,6] (Principal Component Analysis), que consiste numa manipulação da matriz de dados com objetivo de representar as variações presentes em muitas variáveis, através de um número menor de "fatores". Constrói-se um novo sistema de eixos (denominados rotineiramente de fatores, componentes principais, variáveis latentes ou ainda autovetores) para representar as amostras, no qual a natureza multivariada dos dados pode ser visualizada em poucas dimensões.

Com o intuito de entender como funciona o método PCA, será usado um exemplo pedagógico simples com duas varáveis. A figura 1 mostra o gráfico bidimensional de um conjunto de 30 amostras (n = 30). A matriz de dados consiste, neste caso, de duas colunas (m = 2) representando as medidas de intensidades registradas para dois comprimentos de onda λ_1 e λ_2 nas 30 amostras. Cada linha da matriz de dados é representada por um ponto no gráfico. Em termos geométricos a função das componentes principais é descrever a variação ou espalhamento entre os pontos usando o menor número possível de eixos. Isto é feito definindo novos eixos, componentes principais, que se alinham com os dados. Note que na figura 1, nem λ_1 nem λ_2 descrevem a maior variação nos dados. No entanto, a primeira componente principal PC1, tem uma direção tal que descreve o máximo espalhamento das amostras, mais que qualquer uma das duas variáveis originais. Além disto, a percentagem da variação total nos dados descrita por qualquer componente principal pode ser previamente calculada. Neste exemplo, PC1 descreve 92,5% da variação e PC2, ortogonal a PC1, é estimada para descrever a máxima variação restante, isto é, 7,5%. As novas coordenadas das amostras, no novo sistema de eixos das componentes principais mostradas pela linha cheia na figura 1 são denominadas de "scores". Cada componente principal é construída pela combinação linear das variáveis originais. Os coeficientes da combinação linear (o peso, ou quanto cada variável antiga contribui) são denominados de "loadings" e representados pela linha tracejada na Figura 1. Note que eles são, na realidade, os cossenos dos ângulos entre os eixos originais e o novo eixo (PC).

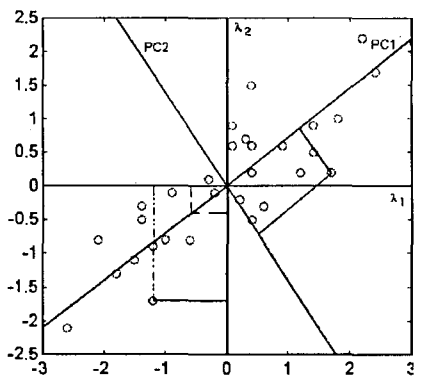


Figura 1. *Gráfico de um conjunto de dados bidimensionais (λ_1,λ_2), mostrando os eixos das componentes principais (PC_1,PC_2)*.*

[a] Para definição de média e desvio padrão, o leitor deve consultar qualquer livro básico de estatística, por exemplo, J. E. Freund & G. A. Simon, "Modern Elementaru Statistics", 9ª Ed. Prentice Hall, 1997.

* Os "scores" de uma amostra são representados por (———) e os "loadings" por (----) As linhas tracejadas (—..—..—) respresentam as coordenadas de uma amostra em relação aos eixos originais.

Há uma variedade de algoritmos usados para calcular os "loadings" e" "scores". Um, comumente empregado, é o de decomposição de valores singulares (SVD)[8] onde a matriz de dados é escrita como:

$$\mathbf{X=USV^t}$$

Na linguagem do MATLAB (ou SciLab), basta usar o comando

```
[U S V]=svd(X);
```

para obtermos as matrizes **U**, **S** e **V**. Mais uma vez, este comando executa uma operação fundamental da álgebra matricial por algoritmos altamente eficientes e longe de serem triviais, mesmo para um analista numérico experiente. As colunas de **U** e **V** são ortonormais[b]. A matriz **V** é a matriz dos "loadings", onde a primeira coluna contém os "loadings" de PC1 e assim por diante. **U*S** corresponde à matriz **T** dos "scores". **S** é uma matriz diagonal cujos elementos (valores singulares) contêm informação sobre a quantidade de variância que cada componente principal descreve[c]. A matriz **S** é importante na determinação da dimensionalidade intrínseca da matriz de dados **X** (posto de **X**). Os autovalores que forem pequenos serão excluídos, e as informações relevantes podem de alguma maneira ser separadas, eliminando-se assim os ruídos experimentais. A expressão abaixo mostra a relação entre o valor singular e a variância contida na jésima componente principal

$$\text{VAR\%PC}j = \frac{s_{jj}^2}{\sum_{j=1}^{p} s_{jj}^2} \text{x100}$$

onde o denominador dá o valor da variância total e p é o número total de valores singulares do conjunto de dados (o menor valor dentre n e m). Os comandos para o cálculo de variância percentual em cada componente principal são

```
s=diag(S).^2;
vart=sum(s);
VAR%PC=(s/vart)*100;
```

A etapa de construção do modelo de calibração começa com a seleção de um conjunto de amostras cuidadosamente escolhidas para que sejam representativas de toda a região a ser modelada. Estes serão os padrões (conjunto de calibração) utilizados na construção de um modelo apropriado para relacionar as respostas instrumentais com a concentração, na forma

$$\mathbf{y = X}\,\beta$$

onde β é o vetor de regressão, que se deseja calcular. Isto se torna trivial por meio da decomposição de valores singulares, devido às propriedades das matrizes dos "loadings" e "scores". Neste caso temos

$$\mathbf{y = USV^t}\beta \Rightarrow \beta = \mathbf{VS^{-1}Uy}$$

o que é obtido pelo simples comando na linguagem do MATLAB (ou SciLab):

```
beta=V(:, 1:k)*inv(S(1:k, 1:k)*U(:, 1:k)'*y
```

onde V(:, 1:k) significa tomar todas as linhas e as colunas de 1 até k da matriz **V**; inv(S(1:k, 1:k), significa calcular a inversa da matriz quadrada formada pelas linhas e as colunas de 1 a k da matriz **S** e finalmente U(:, 1:k)', calcula a transposta daquela matriz obtida tomando-se todas as linhas e as colunas de 1 a k da matriz **U**.

A decisão sobre o número k de componentes principais a ser utilizado no modelo será discutida mais adiante.

Um aspecto característico do método PCR é a construção das componentes principais utilizando unicamente as respostas instrumentais (**X**) sem levar em consideração informações provenientes das concentrações (**y**). Isto pode se constituir numa fragilidade do método no caso em que o analito de interesse tem um sinal muito fraco e portanto não influencia fortemente nas primeiras componentes principais, fazendo com que um número maior delas seja necessário para a construção do modelo.

Outro método de regressão que utiliza a modelagem de componentes principais é o PLS, que contorna a dificuldade característica do PCR descrita acima usando a informação das concentrações na obtenção dos fatores, o que só é justificável se tais concentrações tiverem valores confiáveis. O primeiro fator, neste caso chamado de variável latente, descreve a direção de máxima variância que também se correlaciona com a concentração. Estas variáveis latentes são na realidade combinações lineares das componentes principais calculadas pelo método PCR. Há vários algoritmos para calcular a decomposição usada em PLS. Os dois mais populares são NIPALS[7,9] e SVD. Kowalski e Seasholtz[10] apresentam um algoritmo simples para PLS que usa a decomposição de valores singulares. Como no caso anterior, o método PLS assume um modelo inverso. A matriz de dados **X** é decomposta em três matrizes, como no método PCR, embora por outro processo[11].

A preferência de um dentre estes dois métodos não pode ser aconselhada de uma forma genérica uma vez que ambos são em geral igualmente eficientes e as pequenas variações dependem de caso para caso.

Antes da aplicação do modelo construído, o mesmo deve ser validado com o objetivo de testar a sua capacidade preditiva; sem esta etapa não há sentido em prosseguir. A validação consiste em testar o modelo prevendo concentrações de amostras (de preferência não usadas na sua construção), para estabelecer se ele de fato irá refletir o comportamento do analito de interesse. Durante a etapa de validação dois fatores devem ser considerados:

1 - O número de fatores k a ser utilizado no modelo (número de componentes principais ou número de variáveis latentes).
2 - Detecção de "outliers" (amostras anômalas).

Para a determinação das componentes principais que serão empregadas na modelagem, o ideal seria a utilização de um conjunto-teste de dados. Entretanto, na maioria das vezes isto não é possível pois pode ser um processo demorado e caro. Uma alternativa prática, e que funciona bem, é o método de validação cruzada[12].

A **validação cruzada** é uma metodologia utilizada para a escolha do número de componentes principais baseada na avaliação da magnitude dos erros de previsão de um dado modelo de calibração. Esta avaliação é feita pela comparação das previsões das concentrações previamente conhecidas (c_i i = 1:n), e em resumo consiste do seguinte:

1 - Remove-se uma ou mais amostras i do conjunto de calibração e constrói-se o modelo como anteriormente.
2 - Usa-se o novo modelo para prever os dados removidos $\hat{c}_i$.

---

[b] $v_i^t = \begin{cases} 0 \text{ para } i \neq j \\ 1 \text{ para } i = j \end{cases}$

[c] O método PCA está baseado na correlação entre variáveis e, na realidade, agrupa aquelas variáveis que estão altamente correlacionadas. As colunas da matriz de "loadings", **V**, correspondem aos autovetores da matriz de variância - covariância, $\mathbf{X^tX}$, enquanto que os elementos da matriz **S** são as raízes quadradas dos autovalores da mesma, ordenados em ordem decrescente. Fica assim claro, que cada valor singular, s_{ij}, representa uma fração da variância total.

3 - Calcula-se o erro de previsão (c_i-$\hat{c}_i$).

4 - Calcula-se a soma dos quadrados dos erros de previsão: PRESS $\sum_i (c_i - \hat{c}_i)^2$ ou a raiz quadrada RMSEP, que é na realidade um desvio padrão $\sqrt{\frac{\sum_i (c_i - \hat{c}_i)^2}{n}}$ onde n é o número de amostras do conjunto de calibração.

Na linguagem do MATLAB (ou SciLab) para a j-ésima componente principal temos:

```
X=[X(1:i-1,:);X(i+1:n,:)]; (i-ésima amostra excluída do conjunto de dados)
x=X(i,:); (dados referentes à i-ésima amostra)
        (calcula beta para ser usado na etapa seguinte)
cest=x*beta(:,j); (concentração estimada para a i-ésima amostra)
EP(i,j)=(c-cest); (erro de previsão calculado para a i-ésima amostra)
```

O processo é repetido para i=1:n

```
PRESS=sum(EP(:,j).^2);
RMSEP=(sum(EP(:,j))/n).^0.5;
```

O processo é repetido para modelos com uma, duas e assim por diante, componentes principais. Para cada sistema em estudo, o número mais adequado de fatores, k, será o correspondente ao menor valor de PRESS[3,10].

A detecção de "outliers" é tão importante quanto a determinação do número de componentes principais que serão empregadas no modelo. Ao verificar a qualidade do conjunto de calibração, deve-se assegurar de que as amostras formam um conjunto homogêneo, removendo-se aquelas amostras que são solitárias. Para a detecção de "outliers", usa-se duas grandezas complementares: "leverage" e resíduos de Student.

A "leverage" é uma medida da influência de uma amostra no modelo de regressão. Um valor de "leverage" pequeno indica que a amostra em questão influencia pouco na construção do modelo de calibração. Por outro lado, se as medidas experimentais de uma amostra são diferentes das outras do conjunto de calibração, ela provavelmente terá uma alta influência no modelo, que pode ser negativa. Em geral, estas amostras solitárias estão visíveis no gráfico de "scores". A "leverage" pode ser interpretada geometricamente como a distância de uma amostra ao centróide do conjunto de dados e é calculada segundo a equação

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})^t (\mathbf{X}^t\mathbf{X})^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

onde $\mathbf{x}_i$ é o espectro da $i$ésima amostra, $\bar{\mathbf{x}}$ é o espectro médio e $\mathbf{X}^t\mathbf{X}$ é a matriz de variância-covariância. Pela expressão acima, vê-se claramente que a "leverage" é uma função da distância da amostra à média e das correlações entre as variáveis.

Outra maneira bastante simples de se calcular "leverage" é medindo a distância de uma amostra ao centro do conjunto através do cálculo da distância Euclideana no espaço das componentes principais

$$\mathbf{H} = \mathbf{T}\mathbf{T}^t$$

onde T é a matriz dos "scores". Os elementos da diagonal de H, h_{ii}, estão diretamente relacionados com os valores da "leverage".

Existe uma regra prática que nos permite distinguir amostras anômalas: as amostras com $h_{ii} > h_{crit}$, onde

$$h_{crit} = \frac{3k}{n},$$

são consideradas suspeitas e devem ser analisadas caso a caso. Aqui, n é o número de amostras do conjunto de calibração e k é o número de componentes principais ou variáveis latentes.

É interessante, ainda, analisar os resíduos das concentrações[13] que são calculados, por exemplo, por validação cruzada. Amostras mal modeladas têm resíduos altos. Para obter a influência de cada amostra em particular, temos o resíduo de Student, que, para a amostra i, é dado como

$$Lresc_i = \sqrt{\frac{(c - c_i)^2}{(n-1)(1-h_i)}}$$

$$\text{Resíduo de Student}_i = \frac{(c_i - c_i)}{Lresc\sqrt{1-h_i}}$$

onde $Lresc_i$ é o resíduo da concentração da amostra i corrigido pela "leverage".

A seguir estão os comandos usados no cálculo da "leverage" e resíduos de Student para o q-ésimo analito.

```
lev = zeros(n,1);
h = V(:,1:k)*inv(T(:,1:k)'*T(:,1:k))*T(:,1:k)';
for i = 1:n (cálculo da "leverage" para as n amostras)
   lev(i) = X(i,:)*h(:,i);
end

res=X*beta-Y; (conc. Estimadas - conc. Experimentais)
lresc(q)=sqrt((1/(n-1)*((ones(n,1)-lev).^(-2))'.*res(:,q)')*res(:,q))
res_st(:,q) = res(:,q)./(lresc(q)*sqrt(ones(n,1)-lev)); (resíduos de Student para o q-ésimo analito)
```

Supondo-se que os resíduos de Student são normalmente distribuídos pode-se aplicar um teste t como indicativo, para verificar se a amostra está ou não dentro da distribuição com um nível de confiança de 95%. Como os resíduos de Student são definidos em unidades de desvio padrão do valor médio, os valores além de ±2,5 são considerados altos sob as condições usuais da estatística.

A análise do gráfico dos resíduos de Student versus "leverage" para cada amostra é a melhor maneira de se determinar as amostras anômalas. Amostras com altos resíduos mas com pequena "leverage" provavelmente têm algum erro no valor da concentração, que deve, de preferência, ser medida novamente. Outra opção será a exclusão de tal amostra do conjunto de calibração. Amostras com resíduo e "leverage" altos devem sempre ser excluídas e o modelo de calibração reconstruído.

Uma vez validado e otimizado o modelo está pronto, isto é, o número de fatores k está definido e as amostras anômalas foram detectadas e excluídas. Como resultado, obtém-se o vetor de regressão "beta" (β), que será então usado para a previsão da concentração, c_{prev}, de novas amostras (do conjunto de previsão):

$$\mathbf{c}_{prev} = \mathbf{x}^t_{prev}\beta$$

onde $\mathbf{x}_{prev}$ contém o espectro de uma nova amostra.

Entretanto, se o objetivo é prever concentrações de novas amostras, para que os resultados sejam confiáveis, é necessário que todas estas novas amostras estejam na mesma faixa de concentrações daquelas usadas na etapa de calibração.

Um ponto ainda a ser comentado é o vetor de regressão (β). É interessante notar que se a resposta instrumental do analito de interesse é ortogonal às outras respostas, isto é, se não há superposição de bandas no espectro, o vetor de regressão é igual à resposta do analito, a menos de um escalar. Por outro lado, se há alguma superposição o vetor de regressão perde as características individuais do analito, e neste caso uma interpretação qualitativa do vetor de regressão deve ser feita com muita cautela.

EXPERIMENTAL

Os dados usados neste tutorial são provenientes de misturas de três sais de potássio - 2,4,6-trinitrofenolato (picrato),

2,4-dinitrofenolato (2,4-dnf) e 2,5-dinitrofenolato (2,5-dnf). Na Figura 2 são mostrados os espectros destes três compostos puros, obtidos a partir de soluções $1{,}0.10^{-5}$ mol L^{-1}. É visível que os dois dinitrofenolatos têm espectros muito semelhantes; no entanto, os métodos multivariados podem ser empregados mesmo que os três compostos tenham espectros superpostos em uma larga faixa de comprimentos de onda.

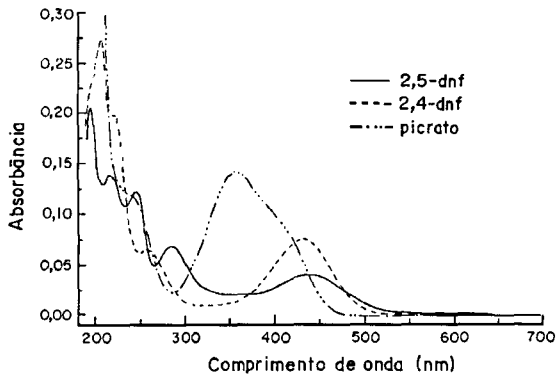


Figura 2. Espectros UV ñ Vis dos três compostos utilizados no experimento de difusão.

Para este trabalho foram preparadas quinze misturas (vide Tabela 1) com concentrações diferentes de cada um dos compostos, e seus espectros foram registrados num espectrofotômetro com arranjo de diodos HP 8452 A. Os espectros foram coletados entre 220 e 530 nm, com intervalos de 2 nm, produzindo 156 variáveis.

No experimento final, envolvendo o transporte dos três derivados fenólicos através de uma membrana líquida hidrofóbica, utiliza-se um sistema de tubo em U esquematizado na Figura 3, acoplado ao espectrofotômetro. Como fase fonte foi usada uma mistura $1{,}0 \times 10^{-2}$ mol L^{-1} em cada composto. A membrana empregada foi uma solução $3{,}0 \times 10^{-2}$ mol L^{-1} de éter coroa 18C6 (o carregador) em clorofórmio.

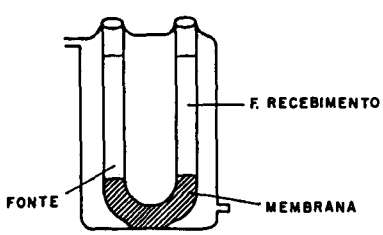


Figura 3. Sistema de tubo em U utilizado nos experimentos de transporte.

Os espectros foram registrados na fase de recebimento em intervalos regulares de cinco minutos; a partir destes espectros e do modelo recém construído, foi possível prever a concentração de cada composto em cada tempo e calcular a taxa de transporte dos mesmos.

A Figura 4 mostra o perfil de um conjunto de espectros registrados.

RESULTADOS E DISCUSSÃO

O modelo de calibração foi construído utilizando as onze primeiras amostras da Tabela 1, X=(n=11 x m=156), e o preprocessamento utilizado constituiu em centrar os dados na média. Convém salientar que neste exemplo há três analitos de interesse e, portanto, o bloco **Y** pode ser representado como uma matriz **Y**=(11 x 3).

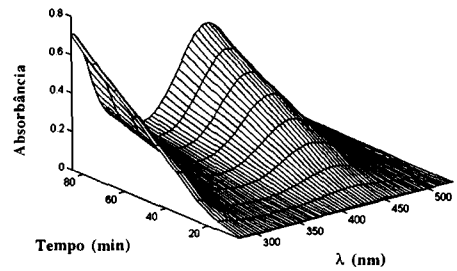


Figura 4. Gráfico da superfície de resposta para um experimento de transporte.

Inicialmente foi feita a Análise de Componentes Principais (PCA) com o objetivo de se ter uma idéia do número de componentes principais necessários para descrever o conjunto de dados. Pela análise PCA se observa que três componentes principais são suficientes e os resultados da porcentagem de variância explicada em cada componente principal estão na Tabela 2. A Figura 5 contém os "loadings" para as três primeiras componentes principais. É interessante comparar estes resultados com os espectros puros (Figura 2). Está claro que toda a região espectral utilizada é importante na análise.

Tabela 2. Porcentagem de variância obtida com a análise de componentes principais (PCA).

Porcentagem de variância explicada pelo modelo PCA		
# Componente principal	% Variância desta PC	% Variância total
1	97,52	97,52
2	1,41	98,94
3	1,05	99,99
4	0,00086	100,00
5	0,0008	100,00
6	0,00	100,00

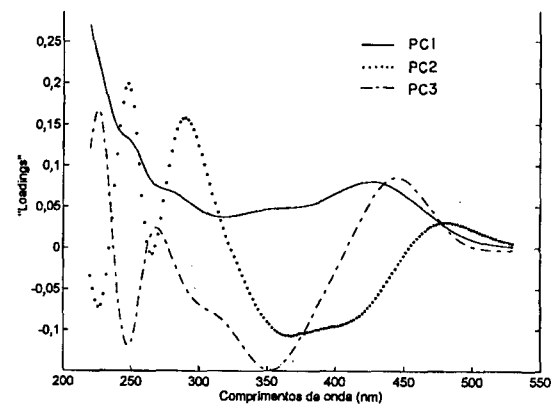


Figura 5. Gráfico de "loadings" das três primeiras componentes principais.

Neste tutorial, a título de ilustração, é feita uma calibração usando o método PLS. Na etapa de validação, é feita a escolha do número *k* de variáveis latentes (que corresponde às componentes principais neste caso) e a identificação de possíveis "outliers". A escolha de *k*, tal como discutido anteriormente, foi feita através do gráfico de PRESS versus número de variáveis latentes, mostrado na Figura 6. Conforme se pode observar, a partir de 3 variáveis latentes o valor do PRESS é próximo de zero. Este resultado de $k = 3$ já era esperado uma vez que temos no sistema três compostos químicos.

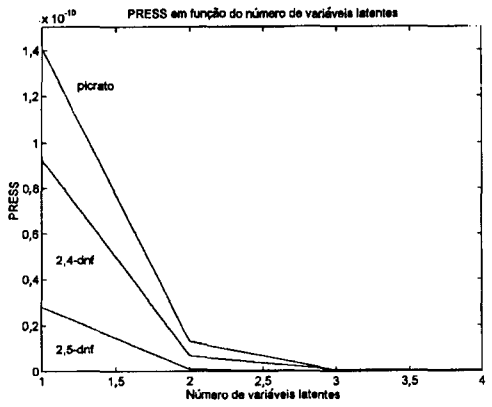


Figura 6. Gráfico de PRESS versus número de variáveis latentes obtidos pelo método PLS para as quatro primeiras componentes principais.

A Tabela 3 mostra as porcentagens da variância descrita pelo modelo para as três primeiras variáveis latentes. Os resultados são bem semelhantes àqueles obtidos com PCA, indicando que os resultados obtidos pela calibração com o método PLS serão bem semelhantes aos obtidos pelo método PCR.

Tabela 3. A variância descrita pelo método PLS com três variáveis latentes (VL).

Porcentagem de Variância descrita pelo método PLS				
	——Bloco - X——		——Bloco - Y——	
VL #	Esta VL	Total	Esta VL	Total
1	97,52	97,52	71,43	71,43
2	1,37	98,90	23,78	95,21
3	1.09	99.99	4,70	99.91

Observando-se, agora, o gráfico de resíduos versus "leverage" (Figura 7), é evidente que as misturas do conjunto de calibração possuem valores de "leverage" dentro da faixa considerada normal, ou seja, menor que h_{crit}, que neste caso tem o valor 0,81 (= 3*3/11). Pode-se ver também que os valores de Resíduo de Student para as concentrações, estão dentro da faixa prevista (-2,5 a +2,5). É necessário salientar que neste único gráfico estão incluídos os resultados obtidos para os três analitos. Assim, aparecem três resultados para uma mesma amostra, cada um se referindo a um dos analitos.

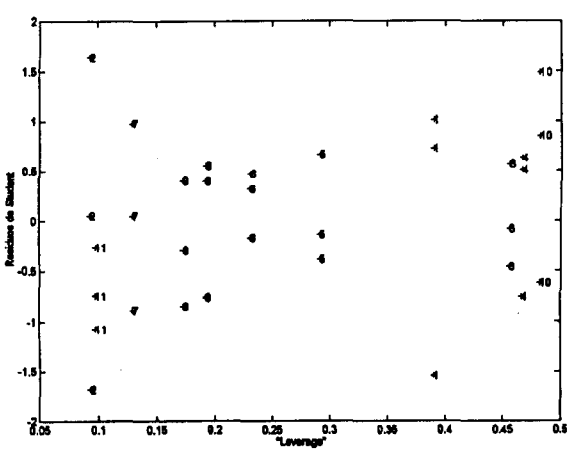


Figura 7. Gráfico de resíduos de Student versus "leverage" obtidos pelo método PLS com três variáveis latentes.

Foram calculados também os erros percentuais para cada amostra usada na etapa de calibração, usando a equação

$$E(\%) = \left(\frac{c_{est} - c}{c}\right) x100$$

Os erros percentuais de previsão obtidos na validação cruzada são apresentados na Tabela 4, para os três analitos usando um modelo com três variáveis latentes. Observa-se que os resultados são bastante satisfatórios, pois de 33 resultados, apenas 5 estão acima de 1,5%.

Tabela 4. Erros percentuais de previsão (conjunto de calibração), obtidos pela validação cruzada com três variáveis latentes.

Mistura	2,4-dnf	picrato	2,5-dnf
	Erro (%)	Erro (%)	Erro (%)
01	1,31	-0,52	-0,87
02	-2,25	-0,05	2,36
03	-0,91	0,39	0,06
04	-0,50	0,64	-0,81
05	0,25	-1,04	1,44
06	0,64	-0,77	-0,61
07	-0,13	0,96	-2,52
08	3,69	-0,56	-1,37
09	-0,98	0,42	4,25
10	0,65	-0,84	-1,35
11	0,40	0,96	1,29

O gráfico dos três vetores de regressão (um para cada analito) no modelo validado com três variáveis latentes (Figura 8), ressalta mais uma vez que toda essa faixa dos espectros entre 220 e 530 nm é importante para a modelagem, visto que para cada intervalo dessa faixa o coeficiente de regressão é mais significativo para um dos compostos. É útil comparar os dados originais (Figura 2) com o dos "loadings" (Figura 5) e os vetores de regressão (Figura 8). Esta comparação pode facilitar a interpretação das forças que estão por trás do modelo de regressão.

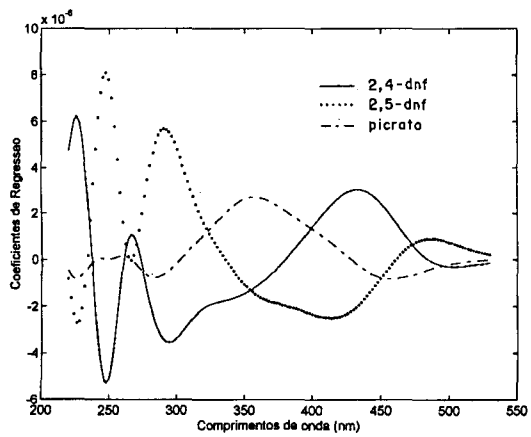


Figura 8. Vetores de regressão obtidos pelo método PLS, com três variáveis latentes.

Para terminar o processo de validação, as quatro últimas amostras da Tabela 1 serão utilizadas como conjunto-teste. A Tabela 5 mostra os erros percentuais de previsão obtidos para as concentrações dos três compostos neste conjunto. Conforme se observa, os valores previstos estão bem próximos dos reais, indicando que o modelo está bem ajustado e com boa capacidade preditiva.

Tabela 5. Erros percentuais de previsão do conjunto de previsão.

Mistura	2,4-dnf	picrato	2,5-dnf
	Erro (%)	Erro (%)	Erro (%)
12	-0,30	0,60	-1,60
13	1,75	-0,14	0,63
14	0,93	1,13	0,50
15	3,10	0,29	0,92

As etapas de construção e validação do modelo estão concluídas estando o mesmo pronto para ser utilizado em previsões. Este conjunto de dados foi escolhido para ilustração, pelo seu interesse do ponto de vista químico.

Para finalizar a ilustração, o modelo de calibração recém construído pode ser utilizado para determinar as taxas de transporte das três espécies químicas. Para tal, foram registrados 19 espectros na fase de recebimento em intervalos regulares de 5 minutos (Figura 4). As concentrações de cada espécie em cada tempo foram previstas pelo modelo e estão representadas na Figura 9. A inclinação das respectivas retas correspondem às taxas de transporte e podem ser determinadas por uma simples regressão das concentrações com os respectivos tempos. Estes valores são apresentados na Tabela 6.

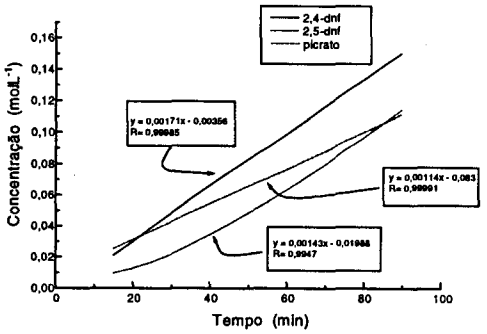


Figura 9. Gráfico das concentrações previstas versus tempo para um experimento de transporte em membrana.

Tabela 6. Taxas de transporte obtidas para cada composto.

Composto	Taxa de Transporte (mol min^{-1})
2, 4-dnf	$1{,}71 \times 10^{-7}$
2, 5-dnf	$1{,}43 \times 10^{-7}$
Picrato	$1{,}14 \times 10^{-7}$

CONCLUSÕES

O objetivo deste trabalho é apresentar um tutorial em Quimiometria, como um guia prático para se fazer uma calibração multivariada eficiente, bem argumentada e com boa capacidade preditiva.

Os métodos teóricos foram introduzidos, inclusive com os comandos apropriados para a linguagem do MATLAB. Abaixo, é apresentado um roteiro geral que deve ser seguido nestes estudos.

O processo geral de construção de modelos de regressão consiste de diversas etapas:

1 - Escolha apropriada do preprocessamento nos dados originais.
2 - Calibração, isto é, construção do modelo de regressão para o conjunto de calibração.
3 - Validação do modelo, escolhendo-se o número de componentes principais a serem utilizados e detectando-se amostras anômalas, para otimizar a capacidade preditiva do modelo.
4 - Interpretação dos "loadings", "scores" e vetor de regressão do modelo validado.
5 - Previsão de novos dados.

Deve-se observar que a utilização de um ambiente computacional de alto nível é indispensável para que operações matemáticas sejam feitas com maior rapidez, eficiência e precisão. Com isto, o químico tem à mão ferramentas matemáticas de última geração compatíveis com seus instrumentos de laboratório e dispõe melhor de seu esforço para atuar naquilo que é sua especialidade.

Os dados experimentais podem ser obtidos através de contato com os autores. marcia@iqm.unicamp.br

AGRADECIMENTOS

Os autores agradecem o apoio das agências financiadoras, FAPESP (MMCF) e CNPq (AMA).

REFERÊNCIAS

1. Vandeginste, B. G. M.; *Top. Curr. Chem.* **1987**, *141*, 1.
2. Araki, T.; Tsukube, H.; *"Liquid Membranes: Chemical Applications"*; CRC Press, Boca Raton 1990.
3. Martens, H.; Naes, T.; em *"Multivariate Calibration"*; John Wiley & Sons, New York 1989.
4. Haaland, D. M.; Thomas, E. V.; *Anal. Chem.* **1988**, *60*, 1193.
5. Malinowski, E. R.; *"Factor Analysis in Chemistry"*; John Wiley & Sons, New York 1991.
6. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B.; *"Chemometrics: A Practical Guide"*; Wiley, New York 1998.
7. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**, *185*, 1.
8. Golub, G. H.; Loan, C. F. van; *"Matrix Computation"*, The John Hopkins University Press, London 1989.
9. Wold, H.; *"Research Papers in Statistics"*; Daid, F.; Ed.; Wiley, New York 1966, 411.
10. Kowalski, B. R.; Seasholtz, M. B.; *J. Chemom.* **1991**, *5*, 129.
11. Lorber A.; Wangen, L.; Kowalski, B. R.; *J. Chemom.* **1987**, *1*, 19.
12. Wold, S.; *Technometrics* **1978**, *20*, 397.
13. Welsch, R. E.; *Technometrics* **1983**, *25*, 245.