5.
Quimiometria

Quimiometria é o nome que se dá ao uso de métodos matemáticos e estatísticos para manipular, interpretar e predizer dados químicos.

O termo só apareceu na década de 1970, na química analítica, como decorrência dos grandes progressos na instrumentação (que produz dados, cada vez mais) e nos microcomputadores (que permitem processar esses dados).

Hoje, a quimiometria pode ser vista como um conjunto de métodos para formular, ajustar, validar e usar modelos empíricos, em geral, com aplicação em todos ramos da química. Uma de suas contribuições foi apresentar aos químicos o modo de pensar multivariado. Isso se aplica aos dados da química analítica instrumental, que são por natureza multivariados, mas também ao planejamento de experimentos, tradicionalmente dominado pela mentalidade de variar um fator de cada vez, Neto et. alli (20).

As análises quantitativas que eram realizadas, na maioria das vezes, por via úmida como titulação, precipitação e reações específicas, demoradas e muitas vezes pouco precisas, estão cada vez mais sendo substituídas por técnicas instrumentais como: Ressonância Magnética Nuclear, Espectroscopias no Infravermelho, no Visível/ Ultra Violeta, Espectroscopia de Massa, Cromatrografia, Polarografia, Análise por Injeção de Fluxo etc., que aliam a velocidade de análise a uma boa qualidade de resultados. Nestas técnicas instrumentais não é obtida uma informação direta do resultado, mas sim, uma grande quantidade de sinais (curvas, picos) que podem ser tratados para uma possível determinação quantitativa das várias espécies químicas presentes, (5).

Muita ênfase tem sido dada aos sistemas multivariados, nos quais se pode medir muitas variáveis, simultaneamente, ao se analisar uma amostra qualquer. Nesses sistemas, a conversão da resposta instrumental no dado químico de interesse, requer a utilização de técnicas de estatística multivariada, álgebra linear e análise numérica. Praticamente todas as técnicas quimiométricas podem ser formuladas em termos de matrizes. Dependendo da representação matricial é

possível classificá-la em dois grandes grupos: a Análise Exploratória e o Reconhecimento de Padrões quando é conveniente representar os dados em uma única matriz, e os problemas de Classificação e de Calibração, quando se busca relacionar duas ou mais matrizes, Adams, M. J.(21).

5.1. Análise exploratória de dados

De um modo geral, um conjunto de dados químicos consiste em um certo número de objetos, descrito por um certo número de variáveis.

Os objetos químicos típicos são amostras analíticas ou espécies químicas. As variáveis são, muitas vezes, derivadas das quantidades de constituintes químicos nos objetos, por exemplo, concentrações de elementos mais importantes, altura de picos em perfis cromatrográficos, comprimentos de onda em perfis espectroscópicos. As variáveis medidas devem ser as mesmas para todos os objetos,(21).

O objetivo principal de uma análise exploratória é extrair informações dos dados, estabelecendo relações entre objetos e variáveis. A análise exploratória não estabelece modelos à priori, mas permite, que a partir das relações observadas nos dados, sejam levantadas hipóteses e propostos modelos, (21).

Por não estabelecer modelos, aprioristicamente, este estudo multivariado é dito de aprendizagem não supervisionada, pois não faz uso de nenhum conjunto de treino para classificar as amostras em pesquisa, em contraste com os estudos multivariados supervisionados, que usam conjuntos de amostras conhecidas (conjunto de treino), (21).

O esquema geral ou algoritmo seguido para desenvolver um estudo de aprendizagem não supervisionada, e empreender o agrupamento de objetos é desenvolvido da seguinte maneira: o conjunto de dados originais ou devidamente processado, que caracteriza as amostras em estudo, é primeiro tansformado em algum outro conjunto que permita medir as similaridades ou dissimilaridades entre as amostras. A seguir é garantir que os objetos similares estão reunidos em grupos com separação mínima, ao mesmo tempo que mantêm um afastamento máximo entre grupos diferentes, (21).

A medida da similaridade ou coeficiente de associação pode ser feita de diferentes formas, sendo o coeficiente de correlação o mais comumente usado. A Análise Hierárquica de Agrupamentos, a Análise de Componentes Principais, o Agrupamento de Fuzzy e outras técnicas são usadas em análise exploratórias de dados, (21).

Tendo feito o agrupamento das amostras pela análise exploratória de dados torna-se possível enfrentar um novo problema, que é o de classificar uma amostra desconhecida nos diferentes agrupamentos obtidos. Esta nova tarefa exige o conhecimento de amostras cujas atribuições já tenham sido feitas. Este novo conjunto de amostras é denominado de conjunto de treino, de aprendizagem ou de calibração. A situação demanda um estudo multivariado denominado, na literatura, de supervisionado, (21). Os estudos anteriores nos conduzem ao agrupamento de amostras ou à classificação de amostras desconhecidas em grupos pré- estabelecidos. Contudo se o objetivo é a determinação quantitativa dos componentes da amostras, ter-se-á de desenvolver um modelo matemático descritivo ou preditivo, ou seja, realizar uma análise de regressão, (21).

5.2. Calibração multivariada

A associação de uma variável dependente e uma única variável independente, diz respeito a uma calibração univariada.

Nas análises espectroscópicas é comum relacionar as intensidades de diversas linhas espectrais com a concentração de um dado analito, assim a variável resposta (concentração do analito) será função de diversas variáveis independentes, exigindo uma calibração multivariada.

A calibração multivariada é levada a cabo por técnicas de regressão multivariada. Diversas técnicas podem ser usadas: regressão linear múltipla, regressão em componentes principais, regressão por mínimos quadrados parciais Adams, (21).

Uma etapa importante ao empreender qualquer técnica de regressão é a seleção de variáveis.

Não se conhecendo previamente as variáveis, sua seleção poderá ser feita por algumas regras teóricas ou, como no caso da espectroscopia, uma inspeção

visual talvez permita verificar a relevância de umas variáveis em relação às outras. No caso particular dos espectros Raman, a seleção visual não é uma empreitada difícil pois os picos observados são pouco sobrepostos. Nas análises quantitativas, as intensidades são medidas em um intervalo de freqüências visando eliminar ou compensar interferências. Havendo dificuldades para a escolha de um intervalo preferencial de freqüência, o espectro todo poderá ser usado. A variável dependente, y_i, pode então ser representada pelo modelo linear, (21)

$$y_t = a + \sum_{i=1}^{n} b_i x_i + \varepsilon_i ,$$

onde y_i é, por exemplo, a concentração de um analito, x_i é intensidade na freqüência i, a é o coeficiente linear e b_i é o coeficiente ou peso associado a cada variável e ε_i o erro aleatório. Para um espectro completo, por exemplo, no intervalo de 4000 – 100 cm^{-1}, i assumirá um valor da ordem das centenas e a resolução destas centenas de equações simultâneas necessárias para determinar todos os coeficientes para prever o valor de y_t exigirá um tempo considerável de computação, (21). Por isso ao preparar tais modelos de calibração é razoável atentar-se para dois pontos: *i)* quais destas variáveis contribuem de forma mais significativa para o modelo de predição e quais poderiam ser abandonadas sem prejudicar o modelo *ii) se h*á algum outro problema, além do tempo de computação, em se ter mais variáveis do que o estritamente necessário.

5.2.1.
Regressão por mínimos quadrados parciais (PLSR)

É uma técnica de análise multivariada de dados utilizada para relacionar uma ou mais variáveis respostas $\mathbf{Y}$, com diversas variáveis independentes $\mathbf{X}$, baseada no uso de fatores. Neste trabalho a matriz $\mathbf{X}$ é formada pelos valores das intensidades Raman, adquiridas em diversos valores de freqüência do deslocamento Raman e a matriz $\mathbf{Y}$ formada por valores de concentração ou propriedades das amostras sintéticas ou de gasolinas comerciais analisadas, por um método de referência, (16)

O PLS permite identificar fatores (combinações lineares das variáveis $\mathbf{X}$) que melhor modelam as variáveis dependentes $\mathbf{Y}$. Além disto, admite, com

eficiência, trabalhar com conjuntos de dados onde haja variáveis altamente correlacionadas e que apresentam ruído aleatório considerável, (16) e (21).

ORGANIZAÇÃO DE DADOS PARA O PLS

Os dados multivariados são, em geral, organizados em matrizes através de vetores em linha ou coluna, de acordo com a convenção adotada. Além disso, os valores relativos às variáveis independentes (espectros das amostras) e às variáveis dependentes (composição ou propriedades das amostras), são organizados separadamente nas chamadas matrizes de intensidades espectrais e matrizes de concentrações e propriedades, respectivamente, (16).

Na matriz intensidade cada espectro é representado como um vetor linha:

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1w} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2w} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3w} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{s1} & A_{s2} & A_{s3} & \cdots & A_{sw} \end{vmatrix}$$

A_{sw} representa a intensidade da amostra s na freqüência w, resultando numa matriz com o número de linhas correspondendo ao número de amostras e o número de colunas correspondendo ao número de freqüências.

Já na matriz concentração, os valores de concentração ou propriedades dos componentes são representados como vetores coluna. Desta forma, cada amostra ocupa uma linha da matriz.

$$\begin{vmatrix} C_{11} & C_{12} & C_{13} & \cdots & C_{1c} \\ C_{21} & C_{22} & C_{23} & \cdots & C_{2c} \\ C_{31} & C_{32} & C_{33} & \cdots & C_{3c} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ C_{s1} & C_{s2} & C_{s3} & \cdots & C_{sc} \end{vmatrix}$$

C_{sc} representa a concentração ou propriedade do componente c na amostra s, resultando numa matriz com o número de linhas correspondendo ao número de amostras e o de colunas ao de componentes, (16).

Essas matrizes de dados são organizadas em pares de modo que cada matriz intensidade possua uma matriz concentração correspondente. Um par de matrizes forma um conjunto de dados, que pode ter diferentes nomes de acordo com sua origem e utilidade.

O conjunto treinamento ou calibração, é o conjunto de dados que contém medidas de amostras conhecidas e utilizadas para desenvolver a calibração. Consiste de uma matriz intensidade contendo os espectros obtidos e de uma matriz concentração contendo valores determinados por um método de referência confiável e independente, (16).

Para que uma calibração seja válida, o conjunto de treinamento utilizado deve conter dados que sejam representativos das amostras reais a serem analisadas. Além disso, como o PLS é uma técnica multivariada, é muito importante que as amostras no conjunto treinamento sejam mutuamente independentes.

Em termos práticos, isso significa que um conjunto treinamento deve, (16):

i) conter todos os componentes esperados;

ii) abranger a faixa de concentração de interesse;

iii) abranger as condições de interesse;

iv) conter amostras mutuamente independentes.

De todos os pré-requisitos, a independência mútua costuma ser a mais difícil de avaliar, principalmente porque a técnica de diluições ou adições sucessivas não pode ser utilizada para o preparo das amostras. Apesar de padrões assim obtidos serem perfeitamente aplicáveis a calibrações univariadas, eles não se aplicam a técnicas multivariadas (forma um sistema inconsistente de equações lineares) O problema é que as concentrações relativas dos vários componentes não variam, e consequentemente, os erros relativos entre as concentrações dos vários componentes também não. As únicas fontes de variação do erro seriam os erros de diluição e o ruído instrumental. A diluição de amostras conduz a valores múltiplos de concentração, o que conduz a indeterminação na resolução das equações simultâneas, (16).

Outro conjunto de dados muito importante é o conjunto validação, utilizadas para avaliar o desempenho da calibração que contém medidas de amostras conhecidas que sejam independentes das amostras usadas no conjunto

treinamento. Trata-se as amostras de validação como se seus valores de concentração fossem desconhecidos, e utiliza-se a calibração com o conjunto treinamento para estimá-los. Compara-se, então, os valores estimados com os valores teóricos (aqueles determinados pelo método de referência) para avaliar o desempenho da calibração em amostras realmente desconhecidas.

Há, finalmente, o conjunto de amostras desconhecidas que contém apenas a matriz intensidade. A calibração construída é utilizada para calcular a matriz resultado que contém os valores de concentrações preditos, (16).

PRINCÍPIOS BÁSICOS DE UMA TÉCNICA MULTIVARIADA

Quando se trabalha com muitas variáveis, alguns fatores devem ser levados em conta para a obtenção de dados com qualidade e sem redundância de informação entre eles, (16).

i) Número de amostras no conjunto treinamento: deve ser igual a pelo menos três vezes o número de componentes.

ii) Exatidão dos valores de concentração nos conjuntos treinamentos: em geral, calibrações satisfatórias são obtidas a partir de valores de concentração determinados por métodos de referência com erro relativo em relação à média inferior a ± 5%.

iii) Número de amostras no conjunto validação (quando houver): em geral utiliza-se um número igual a 30% do total de amostras de calibração e validação.

iv) Exatidão dos valores de concentração no conjunto validação

v) Nível de ruído no espectro.

Com os espectros e valores de referências obtidos, a etapa de construção do modelo de calibração é, geralmente, a mais rápida de todo o processo quando se tem a disposição programas computacionais adequados. É, nessa etapa que algumas escolhas, quanto ao pré-tratamento de dados e aos parâmetros utilizados na construção do modelo PLS, devem ser feitas, (10, 21).

O modelo obtido é, então, testado na etapa de validação, calculando-se o erro entre os valores de concentração teórico e estimado para as amostras de

validação. Esse cálculo indica o erro que se pode esperar ao utilizar-se a calibração para estimar a concentração de amostras reais desconhecidas, (16, 21).

A aplicabilidade do modelo obtido deve ser avaliada à medida que novas amostras são analisadas, pois a representatividade inicial das amostras treinamento pode não estar sendo mantida. Uma proteção quanto a isso é analisar, a intervalos de tempo apropriados, uma amostra de referência. Em geral, como os instrumentos e sistemas envelhecem e os processos mudam, verifica-se uma deterioração gradual no desempenho da calibração inicial, a qual pode ser prevenida com uma atualização periódica do conjunto de treinamento, (16).

PRÉ-TRATAMENTO OPCIONAL DE DADOS

Há diversas maneiras possíveis de tratar os dados antes de encontrar os componentes principais e realizar a regressão. Esses pré-tratamentos são aplicáveis, tanto ao CLS (mínimos quadrados clássico) quanto ao ILS (mínimos quadrados inverso), PCR (regressão em componentes principais) e PLS (mínimos quadrados parciais). Eles se enquadram em três grupos pincipais, (16)

i) remoção de artefatos e linearização

ii) centralização dos dados em torno da média

iii) escalonamento e ponderação

Os pré-tratamentos opcionais podem ser aplicados a, apenas, um espaço de dados ou a ambos espaços de dados espectrais e dados de concentração.

A forma comum de remoção de artefato é a correção da linha base de um espectro. Em um espectro, por exemplo, a conversão da transmitância em absorvância é uma forma comum de linearização. Deve-se ter o cuidado ao emprender essas ações, cuja remoção incorreta pode acarretar a introdução de outros artefatos que são piores para o modelo do que aqueles que se está tentando remover. Contudo, para cada artefato removido corretamente dos dados, ganha-se graus de liberdade que o modelo pode usar para melhor ajustar as relações entre os espaços de dados em estudo e, em conseqüência, gerar uma calibração mais precisa e robusta, (16, 21).

A centralização dos dados em torno da média é, simplesmente, a subtração da intensidade média de cada freqüência espectral da intensidade de cada espectro, ou, em outras palavras: computa-se o espectro médio do conjunto de dados, e o subtrai do espectro de cada amostra. Esta operação desloca a origem do sistema de coordenadas para o centro do conjunto de dados. Do ponto de vista estatístico, a centralização tem como objetivo prevenir que os pontos mais distantes do centro dos dados tenham maior influência que os mais próximos. Dependendo do tipo de dados e da sua aplicação, a centralização pode ter efeito positivo, negativo ou neutro no desempenho da calibração, (16, 21).

O escalonamento, ou ponderação dos dados, implica multiplicar todos os espectros por um fator de escala diferente para cada freqüência, de modo a aumentar ou diminuir a influência sobre a calibração de cada freqüência do espectro. A forma mais comum de ponderação é a seleção das freqüências espectrais que serão incluídas ou excluídas da calibração – as freqüências incluídas são escalonadas por um fator 1, enquanto as excluídas por um fator 0.

O escalonamento de variância significa ajustar o conjunto de dados de modo a igualar a variância de cada variável, ou seja, igualar a influência de cada variável sobre o conjunto de dados. Por isso, a influência de variáveis que contém informações analíticas úteis pode diminuir, enquanto que a de variáveis que contenham principalmente ruído pode aumentar. Fica,então, claro que o escalonamento de variância não deve ser feito a menos que haja uma razão específica para tal, (16).

A regressão por mínimos quadrados parciais envolve encontrar um conjunto de vetores- base (componentes principais) para os dados espectrais e outro de vetores-base para os dados de concentração e, em seguida, relacioná-los. A equação, (16).

$$Y_f = B_f X_f$$

onde, Y_f é a projeção dos dados de concentração de um componente sobre o f-ésimo fator de concentração, X_f é a projeção dos dados espectrais correspondentes àqueles das concentrações sobre o f-ésimo fator e B_f é a constante de proporcionalidade para o f-ésimo par de fatores de dados de concentração e espectral, (16).

As projeções dos dados é um novo sistema de coordenadas que permite representar os dados de concentração ou espectrais reduzindo-lhes a dimensionalidade (compressão de dados) e são denominadas de “scores”. Os “scores” de dados espectrais são diretamente proporcionais aos “scores” dos autovetores dos dados de concentração. Entretanto esta relação linear perfeita é uma idealização que a presença de ruídos nos dados afeta. Os ruídos dos dados espectrais e de concentração são independentes e os deslocamentos dos dados espectrais são diferentes daqueles observados nos dados de concentração, tanto em grandeza quanto em sentido.

A idéia geral do PLS, (16) é tentar alcançar, tanto quanto possível, a congruência ótima entre cada fator espectral e seu fator de concentração correspondente. O PLS tenta restaurar a congruência ótima entre cada fator espectral e seu fator de concentração correspondente girando um contra outro até que o ângulo entre eles seja zero, ou seja, o procedimento matemático dos mínimos quadrados tenta definir um único vetor, $\mathbf{W}$, que represente o melhor compromisso entre os fatores espectral e de concentração,ou seja, que maximize a relação linear entre as projeções dos dados espectrais sobre o fator e as projeções dos dados de concentração correspondentes sobre o mesmo fator. Cada vetor, $\mathbf{W}$, terá tantos elementos quanto forem as freqüências nos espectro e, embora $\mathbf{W}$ seja um fator abstrato, seus elementos são chamados de pesos.

Os fatores $\mathbf{W}$ são obtidos um a um. Após o primeiro fator $\mathbf{W}_1$ ser encontrado, a porção da variância dos dados espectrais capturada por ele é removida do espectro. Da mesma forma, a porção da variância dos dados de concentração capturada por este primeiro fator é removida. Logo, o segundo fator, $\mathbf{W}_2$ é encontrado para os resíduos espectrais e de concentração que não foram capturados pelo primeiro fator. Esse processo continua até que todos possíveis fatores sejam encontrados.

Em geral, devido ao fato dos ruídos dos dados de concentração serem independentes daqueles dos dados espectrais, cada fator $\mathbf{W}$ localiza-se em um plano diferente daquele que contém os dados de concentração e espectrais, devendo, assim, serem projetados para os planos de dados espectrais. Essas projeções são chamadas de fatores de carga e são representadas por $\mathbf{P}$ e $\mathbf{Q}$ para os fatores de carga espectral e de concentração, respectivamente. Cada fator $\mathbf{P}$ e

Q é organizado nas matrizes sob a forma de vetores- linha. Os fatores **W**, **P** e **Q** podem ser mais ou menos semelhantes, dependendo da maior ou menor quantidade de variância não correlacionada com as variâncias dos dados de concentração e espectral. Os fatores **W** do PLS e suas correspondentes cargas espectrais **P**, no caso de a variância espectral ser linearmente correlacionada com a variância dos dados de concentração, serão muito semelhantes entre si, e também, tenderão a ser muito semelhantes aos componentes principais.

O método dos mínimos quadrados se propõe a dar soluções com aproximações ótimas para sistemas lineares do tipo $Ax = b$, de n incógnitas e m equações, encontrando um vetor x, se possível, que minimize a diferença $\|Ax - b\|$ relativa ao produto interno Euclidiano no espaço R^m. Esse vetor é a chamada solução de mínimo quadrado de $Ax = b$, Anton(22).

De acordo com o procedimento, minímos quadrados clássicos, (16) e contextualizando a abordagem matemática à espectroscopia, apresenta-se a seguir diferentes formas de representar a lei de Lambert-Beer.

A absorvância de um componente, em uma determinada freqüência é dada pela equação

$$A = KC$$

onde A é a absorvância, K é o coeficiente de absorção e C é a concentração do componente. Esclareça-se que os significados dos termos desta equação podem ser ampliados. A é uma quantidade a ser medida que é proporcional a uma propriedade C.

Generalizando para uma situação de múltiplos componentes e múltiplas freqüências, a lei pode ser representada por:

$$A_w = \sum K_{wc} C_c$$

onde A_w é a absorvância na w-ésima freqüência, K_{wc} é o coeficiente de absorção na w-ésima freqüência para o c-ésimo componente, C_c é a concentração do c-ésimo componente.

A equação da somatória pode ser expandida, como a seguir:

$$
\begin{aligned}
A_1 &= K_{11}C_1 + K_{12}C_2 + \cdots + K_{1c}C_c \\
A_2 &= K_{21}C_1 + K_{22}C_2 + \cdots + K_{2c}C_c \\
A_3 &= K_{31}C_1 + K_{32}C_2 + \cdots + K_{3c}C_c \\
\vdots \quad & \quad \vdots \quad \quad \vdots \\
A_w &= K_{w1}C_1 + K_{w2}C_2 + \cdots + K_{wc}C_c
\end{aligned}
$$

As equações expandidas indicam que a absorvância observada em uma determinada freqüência é igual a soma das absorvâncias, nessa freqüência, devida a cada componente presente.

Alternativamente, pode-se adotar uma representação matricial:

$$\mathbf{A} = \mathbf{KC}$$

Onde **A** é uma matriz em que cada coluna representa uma amostra e a informação espectral correspondente. Por exemplo $A_{11,}$representaria a absorção da amostra 1 na freqüência 1.

C é uma matriz em que cada coluna representa uma amostra e a informação da concentração do componente correspondente. Por exemplo, C_{11} representaria a concentração do componente 1 na amostra 1.

K é uma matriz em que cada coluna representa os coeficientes de absorção de cada componente puro em uma determinada freqüência. Por exemplo K_{11} é a absorção do componente 1 puro na freqüência 1.

A equação **A** = **KC,** considerando a existência de ruídos nos dados de **A** e **C**, não será passível de solução exata. Assim necessita-se encontrar a solução ótima em mínimos quadrados. Os erros são as diferenças entre o valor de A medido no espectro (experimental) e aquele calculado pelo produto **KC**, donde, erros = **KC - A**.

Para resolver a equação para **K**, inicialmente multiplica-se ambos os membros da mesma pela transposta de **C,** representada por ($\mathbf{C^T}$)

$$\mathbf{AC^T} = \mathbf{K\ C\ C^T}$$

A seguir, multiplica-se ambos os membros da equação obtida por [$\mathbf{C\ C^T}$]$^{-1}$, obtendo-se

$$\mathbf{AC^T\ [\ C\ C^T\ ]^{-1}} = \mathbf{K\ C\ C^T}[\ \mathbf{C\ C^T}\ ]^{-1}$$

O produto $\mathbf{C\ C^T}[\ \mathbf{C\ C^T}\ ]^{-1}$ corresponde à matriz identidade que é igual a 1. A equação de cálculo de **K** fica:

$$\mathbf{AC^T\ [\ C\ C^T\ ]^{-1}} = \mathbf{K}$$

Para que o inverso de [$\mathbf{CC}^{\mathbf{T}}$] exista é necessário que **C** possua, no mínimo, o mesmo número de colunas e de linhas. Desde que **C** possui uma linha para cada componente e uma coluna para cada amostra, isso acarreta que no mínimo deve-se ter um número de amostras igual ao de componentes para se habilitar a resolver a equação. Por outro lado, havendo qualquer dependência linear entre as colunas da matriz **C**, ela é classificada de singular e seu inverso não existirá. Uma das formas mais comuns de se introduzir dependência linear é a construção de um conjunto de amostras por diluição sucessiva, (16).

O procedimento mínimos quadrados clássicos é inadequado para misturas complexas pois exige que se conheça todos os componentes da mesma. Já o mínimo quadrados parciais permite modelar cada componente separadamente.

DESENVOLVIMENTO DO MODELO PLS

A regressão por mínimos quadrados parciais envolve encontrar um conjunto de vetores- base (componentes principais) para os dados espectrais e um conjunto de vetores- base para os dados de propriedades resposta, e, em seguida, relacioná-los. Pode-se calcular todos os fatores (componentes principais) para ambas matrizes usando, por exemplo, o algoritimo NIPALS (Non linear iterative parcial least square), (17)

A equação $\mathbf{Y}_f = \mathbf{B}_f \mathbf{X}$, (16)

Onde, $\mathbf{Y}_f$ é a projeção dos dados de concentração sobre o f-ésimo fator de concentração.

X é a projeção dos dados espectrais correspondentes sobre o f-ésimo fator espectral.

$\mathbf{B}_f$ é a constante de proporcionalidade para o f-ésimo par de fatores de concentração e espectral.

As matrizes **X** e **Y** são decompostas simultaneamente em uma soma de h fatores (variáveis latentes), de acordo com as equações a seguir:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \Sigma\, t_h\, p'_h + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F} = \Sigma u_h q'_h + \mathbf{F}$$

Onde, **T** e **U** são as matrizes de escores das matrizes **X** e **Y**, respectivamente.

P e **Q** são as matrizes dos pesos das matrizes **X** e Y, respectivamente.

E e **F** são os resíduos.

A correlação entre os os dois blocos **X** e **Y** é simplesmente uma relação linear obtida pelo coeficiente de regressão linear, como descrito pela equação a seguir, $u_h = b_h t_h$ para h fatores, sendo que os valores de b são agrupados na matriz diagonal B, que contém os coeficientes de regressão entre a matriz de scores **U** de **Y** e a matriz de escores **T** de **X**. A melhor relação linear possível entre os escores desses dois blocos é obtida através de pequenas rotações dos fatores dos blocos de **X** e **Y**.

A matriz **Y** pode ser calculada de u_h, através da equação: $\mathbf{Y} = \mathrm{TBQ}' + \mathbf{F}$ e a concentração de novas amostras preditas a partir dos novos escores, $\mathbf{T}^*$, substituidos na equação anterior, ou seja, $\mathbf{Y} = \mathbf{T}^* \mathbf{BQ}'$.

Nesse processo, é necessário achar o melhor número de fatores, o que normalmente é feito por validação cruzada ou por uma outra estatística de validação.

Note que não é necessário calcular todos fatores, apenas os primeiros h fatores onde h é grande o suficiente para que se possa decidir quantos fatores devem ser incluídos no espaço vetorial. O programa de cálculo usado foi o Unscrambler 6.11 da CAMO.

O NÚMERO ÓTIMO DE FATORES

O número ótimo de fatores pode ser estabelecido examinando-se a soma dos quadrados do erro residual de predição(PRESS-predict residual error sum of squares) se o planejamento previr um conjunto independente de validação, (16).

O número de fatores que conduzir a um valor mínimo para o PRESS em princípio é o número ótimo de fatores a ser usado na calibração, validação e predição das respostas de novas amostras.

Não se possuindo um conjunto independente de amostras de validação, a escolha do número ótimos de fatores para o PLS ajustar o modelo deve ser feita, para evitar que ruídos sejam incorporados no modelo ajustado, pelo

procedimento denominado validação cruzada. Na validação cruzada, as mesmas amostras são usadas para a calibração e validação das variáveis resposta. O método consiste em deixar fora algumas amostras do conjunto, e então, calibrar o modelo com as demais e, daí, predizer os valores para as amostras que foram mantidas a parte e os resíduos de predição. O processo é repetido com outro segmento do conjunto de calibração e assim prosseguindo até que cada objeto tenha sido deixado a parte uma vez. Então, todos os resíduos de predição são combinados para computar a variância residual de validação e a raiz quadrada do erro médio de predição (RMSEP-root mean standard erro of prediction), (16).

A validação cruzada pode ser realizada por diversos procedimentos:

- divisão do conjunto de calibração em diversos segmentos (subconjuntos);
- divisão em dois subconjuntos, que alternativamente, são usados como calibração ou teste;
- retirando uma amostra do conjunto, efetuando-se a calibração com as demais e, com isso, predizer a amostra retirada. (Este procedimento é denominado na literatura de leave- one- out).

O procedimento PLS pode gerar modelos de calibração para todas as propriedades- resposta simultaneamente ou encontrar o melhor fator para cada propriedade- resposta, ignorando as relações necessárias para acomodar as demais respostas. Quando o procedimento modela as propriedades individualmente é, por vezes, denominado de PLS-1 e na outra situação de PLS-2, (16).

MODELOS PLS DE AJUSTE

No estudo realizado, trabalhou-se com dois conjuntos de amostras. O conjunto de misturas sintéticas (MS) formado por 52 amostras e o conjunto de 68 gasolinas comerciais (GC).

No conjunto MS, 36 amostras foram usadas para estabelecer o modelo de calibração/validação e 16 como amostras desconhecidas para a predição. No

conjunto GC, 48 amostras foram usadas para estabelecer o modelo de calibração/validação e 20 como amostras desconhecidas para a predição.

As propriedades destes conjuntos são as variáveis resposta (y). As intensidades Raman são as variáveis preditoras (x). Os espectros foram adquiridos num intervalo de números de onda de 98 a 3600 cm^{-1}.

OBTENÇÃO DOS RESULTADOS

Uma vez preparadas as matrizes de dados espectrais e de propriedades rodou-se o modo PLS-1, usando o programa UNSCRAMBLER 6.11, objetivando verificar o erro mínimo (RMSP) por validação cruzada, procedimento leave-one-out, para cada propriedade considerando todas as variáveis preditoras (1817 pontos) gerando-se, assim, um primeiro modelo PLS para a propriedade. Analisando-se diagramas do tipo coeficiente de regressão ou variáveis importantes para a predição versus variáveis preditoras, procurou-se verificar quais intervalos de freqüência contribuíram mais para o ajuste do modelo e quais intervalos que, em princípio, poderiam ser eliminadas da regressão para otimizar o ajuste do modelo, (16).

Finalmente, usando-se o modelo contruido procedeu-se à predição das respostas para amostras que não foram usadas no modelo de calibração/validação. A seguir é apresentado o detalhamento de alguns resultados obtidos.

A Figura 5.2.1.1.1 representa os resultados obtidos no estabelecimento do número ótimo de componentes principais para a modelagem PLS do benzeno em gasolinas comerciais. Observe que a modelagem PLS estabelece 5 componentes principais para a obtenção de um RMSEP mínimo para o ajuste do benzeno nas gasolinas comerciais estudadas.

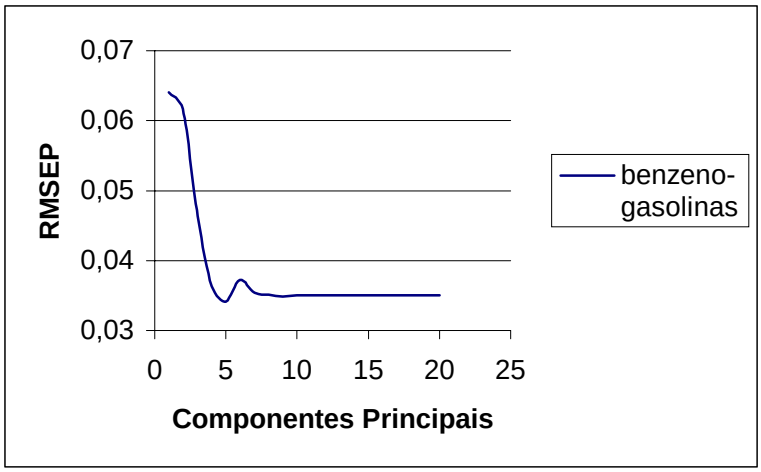

Figura 5.2.1.1.1 – RMSEP e números de componentes principais para o benzeno em gasolinas comerciais

	RMSEP
PC01	0,064
PC02	0,062
PC03	0,046
PC04	0,036
PC05	0,034
PC06	0,037
PC07	0,035
PC08	0,035
PC09	0,035
PC10	0,035
PC11	0,035
PC12	0,035
PC13	0,035
PC14	0,035
PC15	0,035
PC16	0,035
PC17	0,035
PC18	0,035
PC19	0,035
PC20	0,035