5
Análise Multivariada dos Dados

5.1.
Quimiometria

A quimiometria é uma área que se refere à aplicação de métodos estatísticos e matemáticos a problemas de origem química.

Com a sofisticação crescente das técnicas instrumentais, impulsionada pela invasão de microprocessadores e microcomputadores no laboratório químico, tornaram-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).

As análises quantitativas que eram realizadas na maioria das vezes por "via úmida" como titulação, precipitação e reações específicas, que são demoradas e algumas vezes pouco precisas, estão cada vez mais sendo substituídas por técnicas instrumentais como: Espectroscopia no Infravermelho, Espectroscopia no Visível/Ultravioleta, Espectroscopia de Massa, Cromatografia, Ressonância Magnética Nuclear, Polarografia, Análise por Injeção em Fluxo, etc., que aliam a velocidade da análise com uma boa qualidade de resultados. Nessas técnicas instrumentais, geralmente, não é obtida uma informação direta do resultado, mas sim uma grande quantidade de sinais (curvas, picos) que podem ser tratados para uma possível quantificação das várias espécies presentes.

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 matricial e análise numérica. Essas técnicas são apontadas, atualmente, como a melhor alternativa para a interpretação de dados e para a aquisição do máximo de informação sobre o sistema (27).

A quimiometria, cada vez mais difundida, não pode ser considerada como uma simples ferramenta, pois envolve diferentes métodos, tais como (3):

1. otimização de experimentos;
2. otimização e validação de metodologias analíticas;
3. planejamento de experiências;
4. ajuste de curva;
5. processamento de sinal;
6. análise de fatores;
7. calibração multivariada.

Desde o início da década de noventa, o número de publicações em química analítica envolvendo calibração multivariada vem aumentando acentuadamente e, neste contexto, as aplicações mais freqüentes foram as análises utilizando espectroscopia. Isso ocorreu, provavelmente, em decorrência de sua versatilidade e do fato de permitir análises não destrutivas (3).

A regressão por Mínimos Quadrados Parciais (PLS, do inglês “Partial Least Squares”), que foi proposta inicialmente por H. Wold (3), é uma técnica de análise de dados multivariados utilizada para relacionar uma ou mais variáveis resposta (Y) com diversas variáveis independentes (X), baseada no uso de fatores. Usando como exemplo o presente trabalho, a matriz X seria formada por valores de absorvância em diversos comprimentos de onda na região do infravermelho e a matriz Y formada por valores de concentração ou propriedades físicas das amostras de gasolina.

O PLS permite identificar fatores (combinações lineares das variáveis X) que melhor modelam as variáveis dependentes Y. Além disso, admite, com eficiência, trabalhar com conjuntos de dados onde haja variáveis altamente correlacionadas e que apresentam ruído aleatório considerável.

Diversos modelos relacionando propriedades de gasolinas com espectros infravermelho já foram desenvolvidos utilizando quimiometria, sendo a regressão PLS uma das ferramentas mais utilizadas.

No texto a seguir são apresentadas as bases para o entendimento de como funciona o PLS e a sua aplicação para estimar propriedades de gasolinas a partir de espectros no infravermelho. É importante ressaltar que quando se menciona aqui “concentração” se está referindo a qualquer propriedade física ou química da amostra que se deseje estimar.

5.2.
Organização dos Dados

Em problemas como os de calibração multivariada, onde o número de objetos e de variáveis é muito grande, torna-se absolutamente indispensável a disposição ordenada dos dados em forma de matriz para tornar mais fácil a sua manipulação.

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 de gasolina) e às variáveis dependentes (composição ou propriedades das amostras de gasolina) são organizados separadamente nas chamadas matriz absorvância e matriz concentração, respectivamente.

Na matriz absorvância cada espectro é representado como um vetor linha.

$$\begin{matrix} A_{11} & A_{12} & A_{13} & A_{14} & \dots & A_{1w} \\ A_{21} & A_{22} & A_{23} & A_{24} & \dots & A_{2w} \\ A_{31} & A_{32} & A_{33} & A_{34} & \dots & A_{3w} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ A_{s1} & A_{s2} & A_{s3} & A_{s4} & \dots & A_{sw} \end{matrix}$$

A_{sw} representa a absorvância da amostra **s** no comprimento de onda **w**, resultando numa matriz com o número de linhas correspondente ao número de amostras e o de colunas ao número de comprimentos de onda.

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

$$\begin{matrix} C_{11} & C_{12} & \dots & C_{1c} \\ C_{21} & C_{22} & \dots & C_{2c} \\ C_{31} & C_{32} & \dots & C_{3c} \\ \dots & \dots & \dots & \dots \\ C_{s1} & C_{s2} & \dots & C_{sc} \end{matrix}$$

C_{sc} representa a concentração do componente **c** na amostra **s**, resultando numa matriz com o número de linhas correspondente ao número de amostras e o de colunas ao de componentes.

Essas matrizes de dados são organizadas em pares de modo que cada matriz absorvância possua uma matriz concentração correspondente. Um par de matrizes forma um conjunto de dados, que pode receber 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 absorvância contendo os espectros obtidos e de uma matriz concentração contendo valores determinados por um método de referência confiável e independente.

Para que uma calibração seja válida o conjunto treinamento utilizado para construí-la 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 (28).

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

- conter todos os componentes esperados;
- abranger a faixa de concentração de interesse;
- abranger as condições de interesse (temperatura, pH, umidade, etc.);
- 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. O problema é que as concentrações relativas dos vários componentes na amostra não variam e, conseqüentemente, 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.

Outro conjunto de dados muito importante é o *conjunto validação*, que contém medidas de amostras conhecidas que sejam independentes das amostras usadas no conjunto treinamento, utilizado para avaliar o desempenho

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

Há, finalmente, o *conjunto de amostras desconhecidas* ou *amostras teste* que contém apenas a matriz absorvância. A calibração construída é utilizada para calcular a matriz resultado que contém os valores de concentração preditos.

5.3. Princípios Básicos de uma Técnica Multivariada

A calibração multivariada tem como princípio básico a utilização simultânea de muitas variáveis independentes x_1, x_2, ... x_n (por exemplo, valores de absorvância a vários comprimentos de onda), para quantificar alguma variável dependente y (por exemplo, concentração).

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 (28), entre eles:

1. O número de amostras no conjunto treinamento deve ser igual a pelo menos 3 vezes o número de componentes presentes na amostra. Ou, no mínimo, igual a 3 vezes o número de componentes que se deseja estimar.

2. Calibrações satisfatórias são obtidas, em geral, a partir de valores de concentração determinados por métodos de referência com erro relativo em relação à média inferior a $\pm$5%.

3. Para o 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.

4. O nível de ruído nos espectros deve ser sempre avaliado para não interferir nos resultados da análise.

Com os espectros e os valores de referência obtidos, a etapa de construção do modelo de calibração é geralmente a mais rápida de todo o processo quando se tem a disposição “softwares” adequados. É nessa etapa que algumas escolhas quanto ao pré-tratamento dos dados e aos parâmetros utilizados na construção de modelo PLS, discutidos mais tarde, devem ser feitas.

O modelo obtido é então testado na etapa de validação, calculando-se o erro entre os valores de concentração teóricos (fornecidos pelo método de referência) e estimados 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.

A aplicabilidade do modelo obtido deve sempre 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 os sistemas de amostras envelhecem e os processos mudam, verifica-se uma deterioração gradual no desempenho da calibração inicial. Uma atualização periódica no conjunto treinamento pode prevenir essa deterioração.

5.4. Análise dos Componentes Principais (PCA)

5.4.1. Posto de uma Matriz

Posto de uma matriz é o número de vetores linearmente independentes que compõem uma matriz, ou seja, são os vetores que não podem ser escritos como uma combinação linear de outros vetores que pertençam ao mesmo espaço vetorial. A interpretação química para o posto é o número de espécies distintas contidas nas amostras químicas, desprezando os ruídos aleatórios inerentes às medidas (3).

5.4.2. Autovetores e Autovalores

Quando um operador, representado por uma matriz, é aplicado a um espaço vetorial e o produto dessa operação retorna o próprio espaço vetorial

multiplicado por uma constante, tem-se uma equação de autovetores e autovalores (29).

$$\mathbf{M}\Psi = \Lambda\Psi \qquad \text{equação 6}$$

Onde a matriz **M** é o operador que é aplicado no espaço vetorial formado pelos vetores da matriz Ψ resultado em uma constante, Λ, multiplicada pelo próprio espaço vetorial Ψ.

5.4.3. O Espaço de Fatores

O espaço de fatores nada mais é que um sistema de coordenadas particular que oferece certas vantagens para técnicas multivariadas. Quando se trabalha em um espaço de fatores, ao invés do espaço formado pelos dados originais, faz-se simplesmente uma troca do sistema de coordenadas empregado, sem qualquer modificação nos dados em si.

Há várias razões para o uso de um sistema de coordenadas formada por um espaço de fatores apropriado, ao invés das coordenadas originais (28):

1. Eliminação de problemas causados por dados altamente colineares como um conjunto de espectros muito semelhantes.

2. Remoção de ruído dos dados de forma mais eficiente.

3. O espaço de fatores pode elucidar quais variáveis x apresentam maior correlação com as variáveis y, quantos componentes estão realmente presentes, ou quais amostras são semelhantes ou diferentes entre si.

4. Redução da dimensionalidade dos dados.

A utilização dos componentes principais (autovetores) para definir um espaço de fatores que englobe os dados, não modifica os dados em si, mas simplesmente encontra um sistema de coordenadas mais conveniente, capaz de

remover ruído dos dados sem distorcê-los e de diminuir sua dimensionalidade sem comprometer seu conteúdo de informações.

Cada componente principal tem um autovalor associado a ele. Esse autovalor é igual a soma dos quadrados das projeções (*scores*) dos dados sobre o fator correspondente, que nada mais é que a medida da variância total capturada pelo autovetor.

Como cada fator captura o máximo de variância possível, ao fator seguinte resta a variância residual, que se torna cada vez menor a cada fator sucessivo. Conseqüentemente, cada autovalor terá um valor menor que o de seu antecessor.

5.4.4. Descrição Matemática da PCA

Para descrever matematicamente a análise dos componentes principais (30) vamos supor que *n* amostras tiveram seus espectros no infravermelho adquiridos em *m* comprimentos de onda. Essas informações podem ser arranjadas na forma de um matriz absorvância **X** de dimensões *n* x *m*. A PCA é um método de decomposição de uma matriz **X** de posto *r* em um somatório de r matrizes de posto 1, onde posto é o número que expressa a dimensão de uma matriz.

As novas matrizes de posto 1 podem ser escritas como produtos dos vetores chamados "*scores*" (t_h) e "*loadings*" (p'_h), calculados par a par, como na equação 7.

$$X = t_1 p'_1 + t_2 p'_2 + ...+ t_h p'_h \qquad \text{equação 7}$$

A figura 6 apresenta a matriz **X** decomposta em produtos de matrizes "*scores*" e "*loadings*".

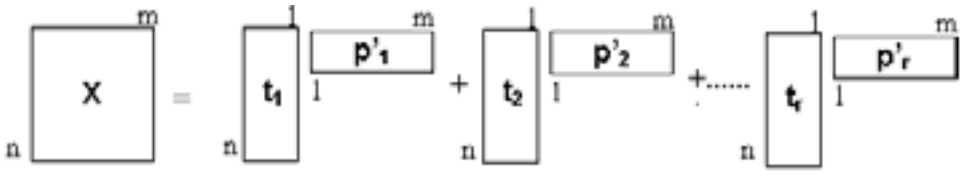

Figura 6- Representação da matriz de dados X decomposta em produtos de matrizes de posto igual a um (30).

Para ilustrar o significado de t_h e p'_h, a figura 7 mostra, no plano bidimensional, duas variáveis x_1 e x_2. A figura 7A mostra um componente principal que é a reta que aponta para a direção de maior variabilidade das amostras da Figura 7B. Os "*scores*" t_h são as projeções das amostras na direção do componente principal e os "*loadings*" p'_h são os cosenos dos ângulos formados entre a componente principal e cada variável.

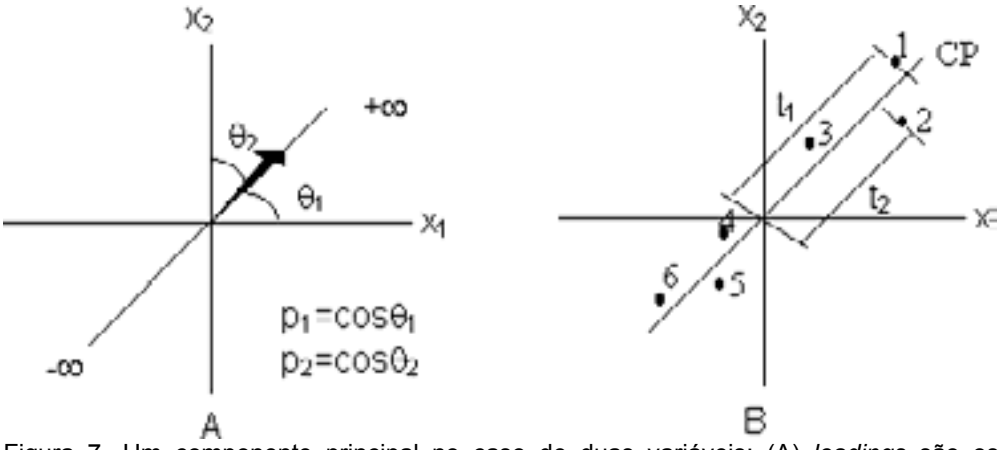

Figura 7- Um componente principal no caso de duas variáveis: (A) *loadings* são os cosenos dos ângulos do vetor direção; (B) *scores* são as projeções das amostras 1 a 6 na direção do componente principal (30).

Em síntese, a análise dos componentes principais é um método que tem por finalidade básica, a redução de dados a partir de combinações lineares das variáveis originais.

5.5.
A Regressão por Mínimos Quadrados Parciais

Toda calibração multivariada utiliza modelos matemáticos para estabelecer uma relação entre uma propriedade que possa ser monitorada com alguma outra propriedade de interesse. O método dos mínimos quadrados parciais é um modelo baseado em variáveis latentes (fatores), onde cada fator é definido como uma combinação linear das variáveis originais das matrizes **X** (variáveis independentes) ou **Y** (variáveis dependentes) (30).

O primeiro componente principal correspondente ao maior autovalor é, por definição, a direção no espaço de **X** que descreve a máxima quantidade de

variância das amostras. Quando toda a variância de um conjunto de amostras não puder ser explicada por apenas um componente principal, um segundo componente principal perpendicular ou ortogonal ao primeiro será utilizado, e assim por diante. Após a modelagem, teoricamente, a matriz dos quadrados dos resíduos deverá conter apenas a variância não explicada associada ao ruído.

A importância da ortogonalidade dos componentes principais se dá pelo fato de que somente desta forma pode-se garantir que a nova base formada resulta de uma combinação de vetores linearmente independentes e, portanto, constituindo um novo espaço vetorial.

A regressão por mínimos quadrados parciais implica em encontrar um conjunto de vetores base (componentes principais) para os dados espectrais e um conjunto separado de vetores base para os dados de concentração e, em seguida, relacioná-los um com o outro. A relação básica entre esses dois conjuntos de vetores é apresentada na equação 8 (28),

$$Y_f = B_f * X_f \qquad \text{equação 8}$$

onde, Y_f é a projeção dos dados de concentração sobre o f-ésimo fator de concentração.

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

B_f é a constante de proporcionalidade para o f-ésimo par de fatores concentração e espectral.

A idéia geral do PLS é tentar alcançar, tanto quanto possível, a congruência ótima entre cada fator espectral e seu fator concentração correspondente, ou seja, encontrar uma relação perfeitamente linear entre as projeções (*scores*) dos dados espectrais e de concentração sobre os seus respectivos fatores.

No entanto, como o ruído dos dados espectrais é independente do ruído dos dados de concentração, aquela relação perfeitamente linear não é possível. A melhor maneira, então, de alcançar uma congruência ótima é utilizar o conceito dos mínimos quadrados. Para isso, os fatores espectral e de concentração correspondentes sofrem uma rotação até que o ângulo entre eles seja zero (28). Em outras palavras, o PLS procura por um único vetor, **W**, que

represente o melhor compromisso entre os fatores espectral e concentração, ou seja, que maximize a relação linear entre as projeções dos dados espectrais sobre o fator **W** e as projeções dos dados de concentração correspondentes sobre o mesmo fator. Cada vetor **W** terá tantos elementos quantos forem os comprimentos de onda nos espectros e, embora **W** seja de fato um fator abstrato, normalmente seus elementos são chamados de pesos (*loading weights*).

Os fatores **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 dos espectros. Do mesmo modo, a porção da variância dos dados de concentração capturada por $\mathbf{W_1}$ é removida. Logo, o próximo fator, $\mathbf{W_2}$, é encontrado para os resíduos espectrais e de concentração que não foram capturados por $\mathbf{W_1}$. Esse processo continua até que todos os possíveis fatores sejam encontrados.

As projeções dos vetores **W** sobre o plano contendo os dados espectrais são chamadas de cargas espectrais (*spectral loadings*), geralmente designados como variável **P**. Do mesmo modo, as projeções dos vetores **W** sobre o plano contendo os dados de concentração são chamadas de cargas de concentração (*concentration loadings*), designados como variável **Q**.

No caso de a variância espectral ser linearmente correlacionada com a variância dos dados de concentração, os fatores **W** do PLS, e suas correspondentes cargas espectrais, **P**, serão muito semelhantes entre si e também tenderão a ser muito semelhantes aos componentes principais.

Assim sendo, no PLS as matrizes X e Y são decompostas simultaneamente em uma soma de h variáveis latentes (30), como nas equações 9 e 10:

$$X = TP' + E = \Sigma\, t_h p'_h + E \qquad \text{equação 9}$$

$$Y = UQ' + F = \Sigma\, u_h q'_h + F \qquad \text{equação 10}$$

onde T e U são as matrizes de "*scores*" das matrizes X e Y, respectivamente; P' e Q' são as matrizes dos "*loadings*" das matrizes X e Y, respectivamente; e E e F são os resíduos. A correlação entre os dois blocos X e Y é simplesmente uma relação linear obtida pelo coeficiente de regressão linear, tal como descrito na equação 11,

$$u_h = b_h t_h \qquad \text{equação 11}$$

para h variáveis latentes, sendo que os valores de b_h são agrupados na matriz diagonal B, que contém os coeficientes de regressão entre a matriz de "*scores*" U de Y e a matriz de "*scores*" T de X. Como já foi mencionado, a melhor relação linear possível entre os "*scores*" desses dois blocos é obtida através de pequenas rotações das variáveis latentes dos blocos de X e Y.

A matriz Y pode ser calculada de u_h, através da equação 12,

$$Y = TBQ' + F \qquad \text{equação 12}$$

e a concentração de novas amostras prevista a partir dos novos "*scores*", T*, substituídos na equação anterior:

$$Y = T^*BQ' \qquad \text{equação 13}$$

Nesse processo, é um passo crítico estabelecer o número correto de componentes principais a serem utilizados nos modelos de calibração, já que os valores preditos para as propriedades dos combustíveis, calculados a partir desses modelos, dependem diretamente do número de componentes principais utilizados. Poucos fatores podem não ser suficientes para modelar adequadamente o sistema, enquanto muitos fatores podem introduzir ruído à calibração, o que resulta num baixo poder de predição para amostras fora do conjunto calibração (18).

A maioria dos programas PLS disponíveis fornece dados para a seleção do número ótimo de componentes principais, construindo o gráfico do erro médio quadrático da predição (RMSEP, do inglês "*root mean square error of prediction*") *versus* o número de componentes principais utilizado. O RMSEP é calculado segundo a equação 14, onde n é o número de amostras. O número de componentes selecionado é, em geral, aquele que fornece um erro de predição mínimo.

$$RMSEP = [\Sigma(y_{predito} - y_{referência})^2 / n]^{1/2} \qquad \text{equação 14}$$

O cálculo do erro da predição pode ser feito através de um conjunto de amostras independente da calibração, o conjunto validação, ou através de validação cruzada. Na validação cruzada, as mesmas amostras são usadas

tanto para construir o modelo quanto para testá-lo. Esse método de validação consiste em deixar algumas amostras de calibração de fora da construção do modelo e então utilizá-las para predição e cálculo dos resíduos. O processo é repetido com um outro subconjunto de amostras de calibração até que todas as amostras tenham sido utilizadas para predição. No passo seguinte, todos os resíduos são combinados para computar a variância residual da validação e o valor do RMSEP e uma calibração final é então calculada com todas as amostras. A validação cruzada completa ("full cross validation", FCV) deixa de fora uma única amostra de cada vez.

5.6. Pré-tratamentos Opcionais dos Dados

O pré-tratamento dos dados é análogo às operações de pesagem, extração, secagem, centrifugação, precipitação, de uma análise química e, freqüentemente, é a fase mais difícil, mais demorada e que determina a qualidade do método de calibração multivariada empregado (3).

Antes de desenvolver qualquer modelo de calibração multivariada recomenda-se um pré-tratamento dos dados para eliminar amostras anômalas, minimizar ruídos e informações superpostas de espécies de interesse, bem como de interferentes.

Há diversas possíveis maneiras de tratar os dados antes de encontrar os componentes principais e realizar a regressão. As mais utilizadas são:

1. *Remoção de artefatos e/ou linearização*. Em espectroscopia, a forma mais comum de remover artefatos é a correção de linha base. Apesar disso, esse tipo de correção deve ser feito com critério para que outros artefatos não sejam introduzidos nos dados ao invés da remoção dos já existentes.

   Para minimizar o ruído (aplicar "*smooth*") pode-se utilizar um filtro com média móvel, transformada de Fourier, transformada de Wavelet, Savitsky-Golay, etc. (3).

   A derivada é freqüentemente utilizada para melhorar a definição de picos que se encontram sobrepostos em uma mesma região e para correção de linha base.

2. A *centralização dos dados em torno da média* é simplesmente a subtração da absorvância média em cada comprimento de onda, de cada espectro no conjunto de dados (28). 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 dos dados e da sua aplicação, a centralização pode ter efeito positivo, negativo ou neutro no desempenho da calibração.

   Após centrar os dados, a matriz terá valores positivos e negativos, e o novo valor de absorvância média para todas as amostras em cada comprimento de onda será igual a zero. Da mesma maneira, a matriz concentração também pode ser centralizada em torno da média, componente a componente. A decisão de centrar ou não os dados da absorvância é independente da mesma decisão para a matriz concentração.

   Para ilustrar a centralização em torno da média, a figura 8A mostra os dados antes da centralização e a figura 8B mostra os dados após a centralização. Podemos verificar que as posições relativas dos dados não foram alteradas, havendo apenas o deslocamento da origem do novo sistema de coordenadas para o centróide dos dados.

3. *Escalonar ou ponderar os dados* implica em multiplicar todos os espectros por um diferente fator de escala para cada comprimento de onda, de modo a aumentar ou diminuir a influência sobre a calibração de cada comprimento de onda particular (28). Um dos tipos mais comuns de escalonamento é o de variância (*variance scaling*), muitas vezes chamado de **padronização**.

   Aplicar a padronização 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. Para isso, a influência de variáveis que contenham informações analíticas úteis pode diminuir enquanto que a de variáveis que contenham principalmente ruído pode aumentar. Fica claro, portanto, que a padronização não deve ser feita a menos que haja uma razão específica para isto.

   No caso da matriz concentração, por exemplo, a variância de cada componente é ajustada para a unidade dividindo-se o valor de concentração de cada amostra pelo desvio padrão para este componente.

   Ao contrário de centrar os dados, padronizá-los modifica as posições relativas dos dados, mas não modifica a posição do centróide do conjunto de dados, como mostrado na figura 8C. Por outro lado, a decisão de

padronizar ou não a matriz concentração também é independente da decisão de padronizar ou não a matriz absorvância. Uma ilustração do efeito da padronização aplicada junto com a centralização sobre os dados é mostrado na figura 8D.

4. *Seleção de variáveis*. Esse pré-tratamento permite eliminar os termos que não são relevantes na modelagem, gerando uma submatriz com apenas as variáveis que possuem informação. No presente trabalho, a seleção de variáveis importantes foi feita com auxílio do programa estatístico utilizado ou simplesmente através do conhecimento das posições das bandas relativas aos respectivos analitos.

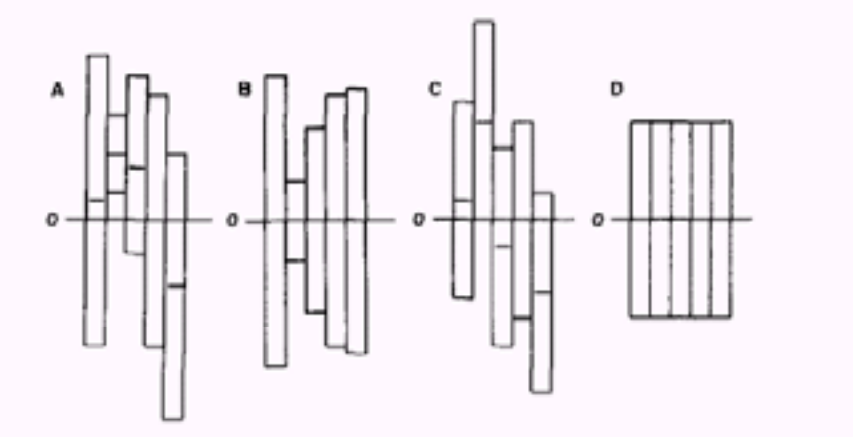

Figura 8- Pré-processamento dos dados.
Os dados para cada variável estão representados por uma barra de variância e seu centro. (A) A maioria dos dados sem tratamento apresentam esse tipo de variação. (B) O resultado após somente a centralização em torno da média. (C) O resultado após somente a padronização. (D) O resultado após centrar e padronizar os dados (30).