Capítulo 2

Quimiometria Aplicada aos Dados Espectrais no Infravermelho Próximo

Marcelo Martins de Sena
Mariana Ramos de Almeida

Introdução

A absorção de radiação na região do infravermelho próximo (NIR, *near infrared*) por moléculas dá origem a espectros bastante complexos, os quais no passado já foram considerados impossíveis de utilizar (Davies, 1998). Os sinais observados nessa região, referentes a bandas de combinações e sobretons de ligações C-H, O-H, N-H e S-H, são relativamente pouco intensos e bastante sobrepostos. Como tais ligações estão presentes na maior parte das moléculas, a utilização da espectroscopia NIR para comparações com padrões espectrais torna-se difícil, ao contrário do que acontece tradicionalmente com a espectroscopia no infravermelho médio. Além disso, a grande sobreposição espectral torna rara a possibilidade de se encontrar um comprimento de onda seletivo, mesmo para matrizes pouco complexas. Por isso, a utilização da calibração univariada em conjunto com espectros NIR é praticamente inviável. Em função destes motivos, a espectroscopia NIR ficou por muito tempo "adormecida" dentre as técnicas espectroscópicas (Wetzel, 1983). Na verdade, a espectroscopia NIR só começou a se difundir a partir dos anos 1970, com a chegada dos microprocessadores e microcomputadores aos laboratórios químicos, e o consequente aumento da capacidade de geração de dados. Com a necessidade de tratamento dos dados gerados pelos químicos, a aplicação de estatística multivariada em química ganhou maturidade e deu origem a uma nova disciplina, batizada de quimiometria (Bruns; Faigle, 1985).

A difusão do uso da espectroscopia NIR sempre esteve estritamente ligada ao desenvolvimento da quimiometria. De maneira recíproca, pode-se afirmar também que a evolução da quimiometria recebeu grande contribuição das demandas geradas pelas aplicações desta técnica. Na verdade, existe historicamente uma "simbiose" entre espectroscopia NIR e quimiometria, que tem contribuído sinergicamente para a evolução de ambas. Desde os primeiros trabalhos publicados pelo pioneiro Karl Norris nos anos 1960, já se notava a necessidade de analisar mais de uma variável para se

extrair informação útil de espectros NIR. A “maneira de pensar multivariada” já estava presente nos primeiros modelos NIR quantitativos desenvolvidos por Norris, o qual buscava empiricamente encontrar alguns comprimentos de onda seletivos que permitissem a utilização de regressão linear múltipla para, por exemplo, determinar o teor de umidade em grãos e sementes (Norris; Hart, 1996; Pasquini, 2003).

Enquanto num período de cinco décadas, entre 1930 e 1980, apenas 255 artigos foram publicados citando espectroscopia NIR (Pasquini, 2003), a década de 1990-1999 marcou uma verdadeira explosão do uso desta técnica (McClure, 1994), com mais de 13 mil artigos publicados. Foi nesta década que surgiu um periódico específico, o *Journal of Near Infrared Spectroscopy*, não por acaso, apenas alguns anos depois do surgimento das duas revistas especializadas em quimiometria fundadas na década anterior, *Chemometrics and Intelligent Laboratory Systems* e *Journal of Chemometrics*. Nos períodos seguintes, o número de publicações envolvendo NIR continuaria a crescer incrivelmente (mais de 35 mil artigos entre 2000-2009 e mais de 62 mil entre 2010-2017, de acordo com pesquisa no *Web of Science*). Seguramente, pode-se afirmar que este crescimento só foi possível graças à evolução simultânea da quimiometria. Atualmente, o desenvolvimento de modelos multivariados qualitativos e quantitativos encontra grande aplicação nas mais diversas áreas, tais como na análise de alimentos, produtos farmacêuticos, combustíveis e materiais. Uma das áreas de aplicação de maior relevância envolve a análise de produtos agrícolas, particularmente grãos e sementes.

A quimiometria trata da aplicação de métodos estatísticos multivariados em química e pode ser definida como a disciplina química que usa métodos matemáticos e estatísticos para planejar ou selecionar condições ótimas de medidas e experimentos e extrair o máximo de informações de dados químicos (Bruns; Faigle, 1985; Otto, 1999). Uma definição mais atual, recomendada pela International Union of Pure and Applied Chemistry - IUPAC, apresenta a quimiometria como a ciência que relaciona medidas feitas em um sistema químico ou processo com o estado do sistema através da aplicação de métodos matemáticos ou estatísticos (Hibbert, 2016).

A quimiometria, como disciplina e linha de pesquisa, foi formalizada no começo dos anos 1970, a partir da colaboração entre seus pioneiros, o químico sueco Svante Wold e o matemático e químico analítico estadunidense Bruce Kowalski, os quais fundaram a Sociedade Internacional de Quimiometria, em 1974 (Geladi; Esbensen, 1990). Desde o começo, as aplicações de quimiometria concentraram-se na área de química analítica. Com a evolução dos equipamentos relacionados às modernas técnicas de espectroscopia molecular, os quais são cada vez mais acessíveis a um menor custo, modelos multivariados tornaram-se mais frequentes e importantes na

literatura. Uma das características mais vantajosas dos métodos quimiométricos é a possibilidade de quantificação ou classificação sem a necessidade de resolução do sinal analítico. Desta forma, a separação física ou química dos interferentes pode ser substituída pela separação matemática dos seus sinais. Isto tem contribuído para o desenvolvimento de inúmeros métodos analíticos mais simples, rápidos e relativamente baratos. Esses métodos demandam um mínimo (ou mesmo nenhum) pré-tratamento das amostras, sendo ambientalmente amigáveis por não gerarem resíduos nem consumirem reagentes ou solventes.

Os métodos quimiométricos podem ser divididos em diferentes tipos de acordo com seus objetivos, tais como planejamento e otimização de experimentos, análise exploratória dos dados, classificação supervisionada, calibração multivariada e resolução de curvas. Com o objetivo de cobrir os principais aspectos teóricos relacionados à construção de modelos quimiométricos utilizando espectros NIR, serão apresentadas neste capítulo quatro seções focadas, respectivamente, no pré-processamento dos dados, na análise exploratória de dados, com destaque para a análise de componentes principais, na calibração multivariada e na classificação supervisionada.

Os dados gerados por espectroscopia NIR são considerados de primeira ordem, pois para cada amostra é gerado um vetor (espectro). Na linguagem da álgebra matricial, vetores são tensores de primeira ordem. Como esses vetores/espectros são coletados em uma matriz de dados, modelos bilineares são usados no seu tratamento. Embora modelos não lineares possam também ser utilizados, tais como redes neurais artificiais e máquina de vetores de suporte (SVM, *support vector machines*), eles não serão abordados aqui. O uso de métodos não lineares é reduzido, pois só se justifica quando métodos lineares não são capazes de descrever matematicamente os dados (princípio da parcimônia), o que normalmente está associado à modelagem de faixas analíticas muito extensas.

Os arranjos de dados a serem processados serão organizados em forma de matrizes, de modo que os espectros de cada amostra estarão dispostos em suas linhas, e os valores de absorbância, dispostos em suas colunas. Para a representação formal dos modelos matemáticos, este capítulo adota a notação algébrica padrão. Matrizes são representadas por letras maiúsculas em negrito, enquanto que vetores são representados por letras minúsculas em negrito e escalares por letras minúsculas normais. A letra T maiúscula sobrescrita será usada para indicar matrizes transpostas, e “-1” sobrescrito, para indicar matrizes inversas.

Pré-processamento dos dados

Na análise multivariada é importante adotar uma sequência de etapas que, dentre outros objetivos, visa a detectar possíveis erros grosseiros e amostras anômalas, além de extrair o máximo de informação dos dados. Invariavelmente, o pré-processamento dos dados é a etapa inicial deste processo. O seu objetivo geral é eliminar ou reduzir a variância aleatória, além de fontes de variação sistemáticas não desejadas. Dessa maneira, a extração das informações poderá focar-se na variância que realmente interessa aos objetivos da análise.

Na análise de dados espectroscópicos, particularmente de espectros NIR, é comum a presença de variação sistemática indesejável, a qual pode mascarar a variância de interesse. Por exemplo, se o objetivo é determinar a composição química de grãos, as variações de densidade de amostra para amostra, ou as diferenças no tamanho das partículas no caso de amostras trituradas, gera variação física que pode prejudicar a habilidade preditiva do modelo. Essa variância deve, portanto, ser eliminada. Mas a escolha do pré-processamento deve ser feita com cautela, baseada no conhecimento sobre os princípios de cada método e sobre a natureza das fontes químicas e físicas de variação nas amostras. O uso de métodos de pré-processamento inadequados pode eliminar informação relevante, prejudicando a qualidade do modelo quimiométrico.

Os métodos de pré-processamento podem ser aplicados tanto nas linhas (amostras) quanto nas colunas (variáveis) da matriz de dados (Ferreira, 2015; Sena et al., 2017). O pré-processamento nas linhas é aplicado em uma amostra de cada vez, considerando todas as variáveis. Reciprocamente, o pré-processamento nas colunas é aplicado a cada variável, considerando todas as amostras. As fontes mais comuns de variância indesejável em espectros NIR são ruídos, espalhamentos e desvios de linha-base (Engel et al., 2013).

Espectros NIR sempre terão certo nível de ruído aleatório, o qual dependerá da amostra, das condições de medida e do detector do espectrofotômetro. Os métodos de alisamento (*smoothing*) são usados para reduzir matematicamente o ruído, aumentando a razão sinal/ruído dos espectros. Assume-se que os ruídos têm alta frequência em relação ao sinal de interesse. O método mais usado para esse fim é o alisamento Savitzky-Golay (Savtzky; Golay, 1964). Este método utiliza um filtro de média móvel, o qual sucessivamente ajusta um polinômio a cada parte/janela do espectro. A escolha do número de variáveis dessa janela é crucial. Um número muito pequeno de variáveis pode não filtrar o ruído de maneira satisfatória, enquanto um número muito grande pode eliminar informação relevante e distorcer o espectro. O usuário deverá, portanto, ajustar o tamanho da janela, que costuma variar entre 7 e 15 pontos, e esco-

lher o grau do polinômio, sendo que os de 2º ou 3º grau são os mais usados (Rinnan et al., 2009).

Outro artefato comum em espectros NIR é a presença de desvios de linha-base, os quais podem ser lineares, causados pela presença de um sinal de fundo (*offset*) que precisa ser corrigido por causa de variações instrumentais ou compensação inadequada do branco, ou não-lineares, causados por efeitos multiplicativos. Desvios de linha-base podem ser corrigidos pelo uso da simples compensação do *offset*, derivadas ou métodos específicos para a correção de espalhamentos multiplicativos. As derivadas são ferramentas matemáticas úteis para melhorar a resolução de sinais analíticos, embora apresentem, como efeito colateral, o aumento do ruído. Por isso, o uso de derivadas deve ser feito conjuntamente com métodos de alisamento (frequentemente, usam-se derivadas de Savitzky-Golay). A primeira derivada elimina desvios de linha-base aditivos, enquanto que a segunda derivada elimina também efeitos multiplicativos.

A eliminação de desvios causados pelo espalhamento de luz é particularmente relevante no processamento quimiométrico de espectros NIR obtidos para sólidos através de medidas de reflectância difusa. O uso de acessórios de reflectância foi muito importante na difusão da espectroscopia NIR, permitindo medidas diretas e aumentando a versatilidade desta técnica. Esse tipo de desvio de linha-base não linear (*drift*) é causado pelas diferenças de caminho ótico oriundas da falta de homogeneidade no tamanho das partículas ou das diferentes densidades dos grãos. Os métodos mais usados para mitigar esta fonte de variação são a correção do espalhamento multiplicativo (MSC, *multiplicative scatter correction*) e a variação normal padrão (SNV, *standard normal variate*), além da normalização vetorial, empregada menos frequentemente (Rinnan et al., 2009). O MSC corrige o espalhamento multiplicativo através de uma regressão em relação a um espectro de referência, normalmente o espectro médio do conjunto de calibração. Dessa forma, interceptos e inclinações são corrigidos, eliminando variação causada pelo espalhamento. O SNV é um tipo de normalização feito ao longo das linhas da matriz. Nele, o valor médio de cada linha é subtraído de todos os valores da respectiva linha, os quais são, em seguida, divididos pelos respectivos desvios-padrão. Portanto, o SNV corresponde ao autoescalamento da matriz de dados transposta (Dhanoa et al., 1994). Pelo fato de não utilizar um espectro médio de referência, o SNV é considerado mais robusto à presença de amostras anômalas (*outliers*) do que o MSC.

Nos três parágrafos anteriores desta seção, foram descritos métodos de pré-processamento aplicados ao longo das linhas da matriz, cuja utilização é mais focada em dados espectrais. Por outro lado, métodos aplicados nas variáveis possuem um uso

mais geral. Dentre estes, os mais importantes são centrar os dados na média e no autoescalamento (Bro; Smilde, 2003). Centrar na média é o pré-processamento mais utilizado, aplicado a todo o tipo de dados, incluindo espectros de um modo geral. Neste método, a média de cada variável é calculada e subtraída do valor de cada elemento da respectiva coluna. A distribuição de cada variável passa a ter média zero, ou seja, o centro das coordenadas do sistema é movido da origem natural para a média multivariada dos dados (espectro médio). O autoescalamento é mais utilizado em dados obtidos para variáveis discretas (medidas físico-químicas, concentrações de elementos, etc.), em modelos de análise exploratória. Ele inclui a "centragem" na média, mas, após esta etapa, cada elemento é dividido pelo desvio-padrão da respectiva coluna. Como consequência, todas as variáveis terão média zero e variância um, o que dará a elas pesos iguais, independentes de suas escalas naturais. É importante frisar que espectros de um modo geral, e de NIR em particular, não devem ser autoescalados, pois os picos de maior intensidade normalmente contribuem com mais informação para o modelo. Além disso, o autoescalamento de espectros aumenta a contribuição de variáveis ruidosas, em regiões espectrais de pouca absorbância, para o modelo.

Após o pré-processamento, o próximo passo da análise quimiométrica é a construção dos modelos. Nas três seções seguintes, serão apresentados os principais tipos de modelos utilizados para a análise de espectros NIR.

Análise de componentes principais (PCA)

Os métodos de análise exploratória de dados são também conhecidos como métodos de classificação não supervisionada. O termo "não supervisionada" indica que não serão fornecidas ao modelo quaisquer informações sobre a origem das amostras. Os objetivos desse tipo de método são verificar a presença de agrupamentos naturais de amostras, reduzir as dimensões dos dados e extrair os padrões latentes de informações mais relevantes. Por isso, às vezes usa-se também o termo "reconhecimento de padrões" para se referir a esses métodos. Modelos exploratórios extraem informação latente, que pode ser visualizada em gráficos de dispersão simples ou em dendrogramas. Dessa forma, o analista pode identificar mais facilmente as tendências e os padrões subjacentes em grandes conjuntos de dados, os quais não seriam perceptíveis observando-se os valores de uma variável de cada vez. Os mais importantes métodos quimiométricos exploratórios são a análise de componentes principais (PCA, *principal component analysis*) e a análise hierárquica de agrupamentos (HCA, *hierarchical cluster analysis*).

A HCA (Otto, 1999; Ferreira, 2015) busca observar agrupamentos de amostras (*clusters*) baseando-se em medidas de similaridade, normalmente na distância euclidiana, sem envolver uma etapa de redução da dimensionalidade dos dados. Como resultados, são obtidos dendrogramas que permitem analisar a formação de agrupamentos em função da distância entre as amostras no espaço multidimensional da variância dos dados. Por não envolver uma etapa de projeção dos dados e não permitir observar simultaneamente as relações entre amostras e variáveis, a HCA possui menor capacidade de extrair informações do sistema do que a PCA (Sena et al., 2017) e, por isso, não será discutida em profundidade aqui.

A PCA pode ser considerada o mais importante método quimiométrico. Além de ter ampla aplicação em diversas áreas da ciência, a sua compreensão teórica é um pré-requisito para o entendimento da maioria dos métodos quimiométricos de classificação e calibração multivariada. Historicamente, pode-se afirmar que a PCA foi reinventada por diversas vezes (Bro; Smilde, 2014), datando do início do século XX a sua formulação original (Pearson, 1901).

A PCA é um método que permite a redução da dimensionalidade de um conjunto de dados através de sua projeção em um subespaço vetorial, um novo sistema de eixos. Os novos eixos são estimados através de combinações lineares das variáveis originais e são denominados componentes principais (CPs). Os CPs condensam a informação dos dados em um número bem menor de fatores, que são mais facilmente visualizados. Por exemplo, espectros NIR típicos têm centenas de variáveis, que contêm muita variância redundante. Na verdade, um número muito menor de fatores, os componentes químicos das amostras, é responsável pela variação sistemática de interesse. Nesses casos, a PCA reduz centenas de variáveis a apenas uns poucos CPs, que retêm a maior parte da variância. A PCA pode ainda ser vista como um método que "filtra" as correlações presentes nos dados, o que é essencial na análise de espectros, que em geral contêm muitas variáveis correlacionadas. Na literatura, podem ser encontrados bons artigos tutoriais recentes descrevendo a aplicação da PCA tanto a dados discretos (variáveis físico-químicas) (Bro; Smilde, 2014) quanto contínuos (espectros) (Souza; Poppi, 2012).

Uma matriz de dados (espectros) $\mathbf{X}_{(n,m)}$, formada por n linhas (amostras) e m colunas (comprimentos de onda), é decomposta em um modelo PCA como o produto de uma matriz de escores $\mathbf{T}_{(n,A)}$ e uma matriz de pesos (loadings) $\mathbf{P}_{(m,A)}$, de acordo com a seguinte equação:

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \ldots + \mathbf{t}_A \mathbf{p}_A^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (1)$$

sendo *A* o número de CPs no modelo PCA, $\mathbf{t}_1$ e $\mathbf{p}_1$ os vetores de escores e pesos do primeiro CP, respectivamente, e **E** a matriz de resíduos contendo a variância não descrita pelos CPs incluídos no modelo. Desta forma, a estimativa da matriz de dados experimentais pelo modelo PCA é dada por $\hat{\mathbf{X}} = \mathbf{TP}^T$. Deve-se ressaltar que os CPs são definidos de modo a explicar a máxima variância dos dados. O vetor $\mathbf{p}_1$ descreve a direção de maior variação no espaço multidimensional. Se os dados não estiverem centrados na média, $\mathbf{p}_1$ descreverá o espectro médio das amostras. Na sequência, o vetor de pesos $\mathbf{p}_2$ descreve a direção de maior variância restante nos dados, mas submetido à restrição de ser ortogonal ao primeira CP, e assim por diante. Por isto, os CPs são sempre modelados em ordem decrescente da variância explicada. A restrição de ortogonalidade entre os CPs indica que eles são independentes, ou seja, totalmente não correlacionados.

Os escores e pesos são os mais importantes parâmetros de modelos PCA. Algebricamente, os elementos de cada vetor de pesos (**p**) representam os cossenos dos ângulos entre o respectivo CP e os eixos definidos pelas variáveis originais. Ou seja, os elementos de **p** mostram o quanto cada variável original contribui para o respectivo CP. Por outro lado, os escores de cada vetor **t** são as projeções ortogonais das amostras no eixo de cada CP. Para a interpretação de modelos PCA, é essencial entender que os escores fornecem a composição de cada CP relacionada aos objetos/amostras, enquanto os pesos relacionam essa mesma composição com as variáveis mais importantes (Sena et al., 2002). Após a escolha do número de CPs a ser incluído no modelo, os dados originais são projetados no novo espaço de dimensões reduzidos. As inter-relações e os possíveis agrupamentos de amostras podem ser observados através de gráficos de dispersão bi ou tridimensionais, nos quais os escores dos principais CPs são plotados uns contra os outros. Esta é a principal ferramenta de interpretação de modelos PCA. As projeções ortogonais dos escores em cada eixo separadamente indicam contrastes, que discriminam amostras mais negativas de amostras mais positivas. Amostras com escores próximos de zero em um CP indicam sua falta de importância na interpretação desta componente. É importante ressaltar que bons modelos PCA nem sempre mostram a formação de agrupamentos de amostras claramente distintos. Se o fenômeno que está sendo modelado dá origem, por exemplo, a amostras de composição contínua entre dois extremos, não é esperada a observação de agrupamentos nitidamente separados.

A observação dos escores deve sempre ser complementada pela observação simultânea dos pesos, buscando relacionar quais variáveis contribuem mais para caracterizar o contraste de amostras observado em cada CP. O analista deve sempre interpretar o significado de cada um dos CPs em termos químicos, biológicos, etc. Ele obrigatoriamente deve ser capaz de responder à pergunta “O que significa CP1?”.

A resposta pode ser, por exemplo, um contraste entre sementes de duas origens diferentes, ou entre grãos sadios e infestados por pragas. A observação conjunta de gráficos de escores e de pesos de CP1 x CP2 pode ser condensada em gráficos bivariados (*biplots*). Os biplots mostram escores e pesos, após normalização específica, no mesmo gráfico (Gabriel, 1971) e estão disponíveis como recurso visual em todos os softwares quimiométricos. No entanto, biplots e gráficos de dispersão de pesos de CP1 x CP2 são úteis apenas na análise de dados discretos, constituídos de teores de elementos e/ou variáveis físico-químicas. Quando os dados são contínuos, como no caso de espectros, os pesos devem ser observados individualmente em função das variáveis, tais como comprimento ou número de onda (Souza; Poppi, 2012). Nesse caso, biplots não terão utilidade.

Uma das etapas mais importantes na PCA é a escolha do número de CPs. Idealmente, o número de CPs significativos é igual ao número de fontes de variação linearmente independentes presentes nos dados. Por exemplo, na análise de espectros de misturas de compostos, esse número deve ser igual ao número de componentes químicos que contribuem de maneira não correlacionada para o sinal espectral. O número de CPs de um modelo depende da natureza dos dados em particular. Um dos modos mais intuitivos utilizados nessa estimativa é baseado na quantidade de variância capturada pelo modelo PCA (Vexp). Essa quantidade pode ser estimada pela razão entre as somas dos quadrados dos elementos de $\hat{\mathbf{X}}$ e de **X**, expressa pela equação:

$$V_{exp} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{x}_{i,j})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}(\boldsymbol{x}_{i,j})^2}\ 100 \qquad (2)$$

na qual $x_{i,j}$ e $\hat{x}_{i,j}$ são os valores experimentais e estimados obtidos para a amostra *i* medida na variável *j* em **X** e $\hat{\mathbf{X}}$, respectivamente. Com base nos valores de V_{exp} obtidos de 1 a *A* CPs, a quantidade de variância explicada para cada CP é calculada, em conjunto com a variância acumulada pelo modelo PCA. Na modelagem de espectros em geral e NIR em particular, os primeiros CPs costumam explicar a maior parte da variação nos dados. Então, uma regra útil sugere a rejeição do CP cuja variância for menor que uma porcentagem predefinida (por exemplo, 5%). Em muitos casos, apenas os primeiros dois ou três CPs são suficientes para modelar o fenômeno de interesse. Em outras situações, o interesse pode estar em alguma informação que representa uma pequena parte da variação de dados (por exemplo, 2%). Nestes casos, uma regra mais conservadora é inspecionar os vetores peso e rejeitar os CPs que descrevam apenas ruído, ou seja, que apresentem comportamento aleatório.

Um último aspecto importante a ser ressaltado é a detecção de amostras anômalas em modelos PCA. Esta detecção é feita através da observação de gráficos que mostram dois tipos de parâmetros, os resíduos Q e os valores de T^2 de Hotelling. Estes parâmetros estão relacionados, respectivamente, à variância residual não modelada e à influência da variância incluída no modelo para cada amostra (Sena et al., 2014). Amostras que possuam simultaneamente altos valores para esses dois parâmetros são candidatas a serem detectadas como outliers.

Calibração multivariada

Em um processo de calibração, a concentração de um componente ou outra propriedade de interesse é quantificada indiretamente através de medidas realizadas em um sistema químico. Quando existir uma variável seletiva que possa ser medida para esse sistema, a calibração univariada é usada. No entanto, espectros NIR raramente apresentam variáveis seletivas, mesmo para matrizes menos complexas (Pasquini, 2003). Nesse caso, a alternativa é o uso da calibração multivariada (Martens; Naes, 1996; Brereton, 2000), na qual muitas variáveis, centenas ou até milhares no caso de espectros de infravermelho em geral, são usadas para construir o modelo. Como na calibração multivariada predomina a calibração inversa, define-se um bloco de variáveis independentes (espectros), contidas na matriz **X**, e uma (ou mais) variável dependente a ser prevista, contida no vetor **y**. Assim, o modelo é definido matematicamente como $\mathbf{y} = \mathbf{Xb}$, sendo **b** o vetor dos coeficientes de regressão, um para cada variável (comprimento de onda) em **X**.

Dentre as vantagens da calibração multivariada, pode-se mencionar a possibilidade de determinações diretas, na ausência de resolução do sinal analítico, na presença de interferentes desde que presentes no conjunto de calibração, e também a possibilidade de quantificações simultâneas a partir do mesmo conjunto de espectros, quando houver mais de um analito ou propriedade a ser analisado. Pelo menos dois conjuntos de amostras são necessários para o desenvolvimento do modelo. O primeiro é o conjunto de calibração, para o qual os valores das concentrações ou propriedades a ser determinadas são conhecidos, seja pela preparação de amostras de composição planejada ou pelo uso de métodos de referência. Estes valores em conjunto com os espectros são utilizados para estimar os coeficientes de regressão. O outro conjunto de dados é o de validação, no qual o modelo desenvolvido é testado para prever a concentração de amostras independentes, não incluídas na calibração. As amostras costumam ser divididas em cerca de dois terços para o conjunto de calibração e um terço para o conjunto de validação. As amostras de calibração devem ser represen-

tativas de toda a variância sistemática a ser modelada e estar distribuídas homogeneamente. O número de amostras de validação deve ser suficiente para testar uma generalização representativa das previsões. Quando não for possível obter amostras de composição planejada, situação usual na análise de grãos e sementes, deve-se usar algum método para selecionar as amostras de calibração mais representativas a partir do conjunto total disponível inicialmente. Os métodos mais usados para isto são os algoritmos de Kennard-Stone e Duplex (Westad; Marini, 2015).

Enquanto que a regressão clássica ou direta (na qual a concentração é a variável independente/bloco X e a absorbância espectral a variável dependente/bloco Y) predomina na calibração univariada, o seu uso é muito restrito na calibração multivariada. O método dos mínimos quadrados clássicos (CLS, *classical least squares*) tem como limitação a exigência de que as concentrações de todas as espécies absorventes no sistema sejam previamente conhecidas. Obviamente, esta condição não está presente na imensa maioria das situações que envolvem análises espectrais de produtos agrícolas, tornando impossível o uso do CLS, o qual não fornece previsões aceitáveis na presença de interferentes. Por isso, o uso de calibração multivariada inversa é largamente predominante. As diferentes maneiras de estimar o vetor de regressão (**b**) dão origem aos principais métodos de calibração multivariada: regressão linear múltipla (RLM), regressão em componentes principais (PCR, *principal component regression*) e mínimos quadrados parciais (PLS, *partial least squares*).

A maneira mais simples e direta de estimar **b** na equação $\mathbf{y} = \mathbf{Xb}$ é usando a RLM. A solução dos mínimos quadrados para esta equação é $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. No entanto, a matriz $\mathbf{X}^T\mathbf{X}$ só será inversível na ausência de colinearidade entre as variáveis independentes e se existirem mais linhas do que colunas na matriz **X**, ou seja, mais amostras que variáveis. Mesmo que o número de amostras disponível para construir o modelo seja maior do que o número de comprimentos de onda modelados, espectros NIR sempre apresentam alto grau de colinearidade. Como resultado, a solução dessa matriz inversa representará um sistema instável (mal condicionado). De outra maneira, pode-se dizer que a RLM é limitada por incluir toda a variância espectral no modelo. Com isso, quantidade significativa de informação não relevante será incorporada, tornando o modelo pouco robusto. Por isso, a RLM só funcionará satisfatoriamente para sistemas simples e bem-comportados, que apresentem respostas lineares, sem interferentes, com baixo ruído e nenhuma colinearidade. Uma alternativa é combinar a RLM com uma prévia seleção de variáveis, que encontre um pequeno número de comprimentos de onda não correlacionados.

Porém, outra alternativa de uso mais amplo foi imaginada. O que acontece se a RLM for combinada com a PCA? E se, ao invés de a regressão ser feita nos dados

espectrais brutos, ela fosse feita nos escores (**T**) de um modelo PCA? O resultado é exatamente a PCR. A estimativa de **b** se torna igual a $(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$, que não representa mais um sistema instável, pois os escores em **T** não são correlacionados e estão em número bem mais reduzido. A variância redundante é descartada nos CPs não aproveitadas no modelo.

A decomposição dos dados espectrais num modelo PCR é feita de maneira idêntica ao modelo PCA, ignorando nesta etapa a informação contida nas variáveis dependentes. Isto deu origem a críticas à PCR e culminou na formulação de um novo modelo que simultaneamente decompõe ao blocos X e Y (espectros e concentrações), maximizando a correlação entre eles. Esse modelo é o PLS, proposto inicialmente pelo econometrista Herman Wold nos anos 1960 (Sanchez, 2015) e introduzido na literatura química na década de 1980 (Geladi; Kowalski, 1986). Nos anos seguintes, o PLS tornou-se o mais importante método de calibração multivariada. Assim como o PCR, o PLS é capaz de quantificar analitos na presença de interferentes, desde que eles tenham sido incluídos no conjunto de calibração. Entretanto, os fatores/componentes decompostos no PLS estão sob a restrição de explicar simultaneamente a variância dos espectros e das concentrações, o que provoca perda de ortogonalidade. Por isto, os fatores estimados em um modelo PLS não são chamados CPs como na PCR, mas variáveis latentes (VLs), um termo mais geral. Como consequência e de maneira diferente do PCR, as VLs em um modelo PLS não são modeladas necessariamente em ordem decrescente da variância explicada.

Uma das etapas mais importantes na construção de um modelo PLS/PCR é a escolha do número de VLs/CPs. Ela é feita através da validação cruzada, na qual uma amostra (ou grupo) do conjunto de calibração é retirada de cada vez e prevista pelo modelo construído com as demais. Este processo é repetido até que todas as amostras tenham sido previstas. A escolha do número de VLs corresponde ao menor erro de validação cruzada (RMSECV, *root mean square error of cross validation*) ou ao último decréscimo significativo desse valor. O RMSECV é calculado como:

$$RMSECV = \sqrt{\frac{\sum_{1=1}^{n}(y_i - \hat{y}_i)^2}{n_c}} \qquad (3)$$

sendo n_c o número de amostras no conjunto de calibração, $\hat{y}_i$ o valor previsto pelo modelo e y_i o valor de referência da i-ésima amostra. A aplicação do modelo construído às amostras de calibração e validação origina outros dois parâmetros importantes, RMSEC (*RMSE of calibration*) e RMSEP (*RMSE of prediction*). Existem vários tipos de validação cruzada, dependendo dos critérios utilizados na reamostragem. Um dos mais usados é o *leave-one-out* (LOO), no qual uma amostra é removida por vez. No entanto, o LOO é recomendado apenas para pequenos conjuntos de calibração, com

não muito mais do que 20 amostras. Quando o conjunto de calibração for maior, as alternativas são baseadas na remoção de grupos de amostras, tais como blocos contíguos, subconjuntos aleatórios ou venezianas (*venetian blinds*).

A determinação do número correto de VLs é fundamental. Se esse número for menor do que um valor ideal, variância relevante fica fora do modelo, que se torna subajustado e produz resultados inexatos. Por outro lado, se um número maior do que o ideal de VLs é usado, variância redundante é incluída no modelo, que se torna sobreajustado (Faber; Rajkó, 2007). Com isso, este modelo produzirá previsões exatas apenas para o conjunto de dados usado na sua própria construção. Muitas práticas inadequadas, incluindo o uso de softwares como "caixas pretas", geram modelos com sobreajuste, tornando este tipo de erro um dos mais comuns na literatura quimiométrica. Um sinal típico de sobreajuste ocorre quando o RMSEP é muito superior ao RMSEC. Todos os modelos construídos devem ser rigorosamente validados. Práticas como a escolha tendenciosa das amostras de calibração, ou o aumento do conjunto de dados à custa do uso de replicatas como se fossem amostras verdadeiras, devem ser evitadas.

Um bom modelo de calibração multivariada deve explicar a maior parte da variância, geralmente mais de 90%, nos dois blocos, X e Y. A validação externa deve ser realizada através da previsão do conjunto de validação independente e da estimativa do RMSEP. Outro aspecto importante é a interpretação espectral dos modelos construídos, muitas vezes ausente em artigos publicados na literatura. Para isto, vetores informativos gerados no modelo PLS devem ser interpretados criticamente e associados às respectivas bandas espectrais. Os vetores informativos mais importantes são o vetor dos coeficientes de regressão e os VIP (*variable importance in projection*) escores.

Finalmente, dois últimos aspectos importantes devem ser mencionados. A detecção de *outliers* em calibração multivariada (Martens; Naes, 1996; Valderrama et al., 2007) é mais importante do que em análise exploratória, pois a presença deles prejudica a habilidade preditiva do modelo. Além da detecção de amostras com altos resíduos espectrais (resíduos Q) e alta influência (alto *leverage*), deve-se testar também a presença de *outliers* no bloco Y (valores previstos). Após a construção do modelo de calibração multivariada, é necessário ainda checar se o método possui um desempenho adequado para sua aplicação específica. Este processo de validação analítica é realizado através da estimativa de figuras de mérito, tais como linearidade, exatidão (veracidade + precisão), sensibilidade analítica, viés (*bias*) e relação de desempenho do desvio (RPD, *residual prediction deviation*). Atualmente, a grande maioria das normas regulatórias oficiais é concebida de acordo com um raciocínio univariado, o que gera a necessidade de harmonização para métodos multivariados. Nesse sentido, foi importante a criação de uma norma pioneira voltada para calibração multivariada

usando dados de infravermelho, no ano de 2000 (ASTM, 2012). Nos últimos anos, esse assunto tem sido bastante discutido na literatura (Valderrama et al., 2009; Botelho et al., 2013; Sena et al., 2017).

Classificação supervisionada

O desenvolvimento de métodos qualitativos, de classificação supervisionada, na literatura quimiométrica, cresceu muito nos últimos anos, levando alguns autores a qualificar esse crescimento como “vibrante” (Szymanska et al., 2015). O objetivo desses métodos é atribuir amostras desconhecidas a classes previamente definidas. Conforme já mencionado, nos métodos não supervisionados as amostras podem ser atribuídas a grupos pela observação de gráficos de dispersão de escores. Porém, nesses métodos, informações sobre as classes não são fornecidas ao modelo. Por isso, a classificação supervisionada é mais objetiva, proporcionando uma maneira sistemática de classificar novas amostras. No entanto, recomenda-se sempre a construção de modelos PCA previamente aos modelos de classificação, pois eles fornecem um bom quadro dos padrões e diferenças naturalmente presentes nas amostras. De outra maneira, pode-se definir o objetivo da classificação supervisionada como a delimitação de regiões específicas do hiperespaço das variáveis associadas a cada classe, sendo que a definição de classe dependerá de cada problema em particular (Marini, 2010; Sena et al., 2017).

A combinação de classificação supervisionada com técnicas espectroscópicas é bastante apropriada para desenvolver métodos de triagem capazes de fornecer ferramentas rápidas para a solução de problemas que envolvam comprovação de autenticidade, atribuição de origem, detecção de contaminações, fraudes e falsificações, etc. Particularmente, publicações recentes têm combinado espectroscopia NIR e métodos de classificação na análise de grãos e sementes, em problemas que envolvem atribuição de origem geográfica ou genotípica, autenticidade e rastreabilidade de cereais, etc. (Vitale et al., 2013; Cozzolino, 2014; Marquetti et al., 2016).

As estratégias de classificação e a escolha do método mais apropriado podem variar bastante, dependo de cada situação em particular. Em muitas situações, apenas duas classes são definidas, quando existem somente duas classificações possíveis. Exemplos disso são métodos que buscam detectar a ausência ou presença de um contaminante, se uma molécula é ativa ou não, se um produto está ou não adulterado, etc. Por outro lado, quando modelos para discriminar mais de duas classes são construídos, pode ou não fazer sentido que uma amostra seja atribuída a mais de uma

classe simultaneamente, ou pode-se permitir que o modelo rejeite a amostra como não pertencente a nenhuma das classes e seja classificada como um *outlier*. Assim como na calibração multivariada, os dados devem ser divididos em dois conjuntos e a validação cruzada deve ser usada. Em classificação supervisionada, os conjuntos usados para construir e validar o modelo são mais comumente denominados conjuntos de treinamento e teste, respectivamente. A divisão das amostras pode ser feita usando os já citados algoritmos de Kennard-Stone ou Duplex, mas em modelos de classificação é mais apropriado que eles sejam aplicados às amostras de cada classe em separado. Da mesma forma que muitos modelos de calibração sobre-ajustados são publicados em artigos, modelos de classificação enviezadamente otimistas também são encontrados. Dentre as principais práticas incorretas que devem ser evitadas, podem-se citar ausência de validação apropriada, uso de um número pequeno de amostras não representativas da variância que deveria ser associada a determinada classe, e o uso de repetições como amostras independentes, aumentando artificialmente o tamanho de uma ou mais classes.

Métodos de classificação podem ser subdivididos de várias maneiras, como lineares ou não lineares, paramétricos ou não paramétricos, etc. De particular relevância é a subdivisão desses métodos em discriminantes ou de modelagem de classe. Os primeiros definem um ou mais delimitadores entre duas ou mais classes, dividindo o hiperespaço das variáveis em um número correspondente de regiões. Os do segundo tipo modelam cada classe individualmente, sem levar em conta informações relativas às demais classes. Recentemente, alguns autores têm criticado o predomínio de publicações envolvendo métodos discriminantes quando aplicados a problemas de autenticação, pois frequentemente não é possível obter um conjunto de amostras representativo de todas as possíveis situações de não autenticidade (Rodionova et al., 2016).

O mais representativo método quimiométrico de modelagem de classes é o SIMCA (*soft independent modeling of class analogy*), que pode ser traduzido como “modelagem independente por analogia de classes”, desenvolvido por Wold nos anos 1970 (Wold, 1976). O SIMCA consiste numa coleção de modelos PCA, construídos individualmente para cada classe. Os números de CPs são selecionados por validação cruzada e o SIMCA pode ser usado tanto em modelagem multiclasses quanto na modelagem de apenas uma classe, situação indicada em problemas que envolvem a comprovação de autenticidade ou a detecção de fraudes.

Atualmente, o modelo discriminante mais comum na literatura quimiométrica é o PLS-DA (*PLS discriminant analysis*), uma extensão do PLS para classificação (Brereton; Lloyd, 2014). Com tal, o PLS-DA partilha das propriedades do PLS, sendo que o bloco

Y é composto por variáveis categóricas, que assumem os valores 1 ou 0 dependendo se a amostra pertence ou não à respectiva classe. Outra diferença em relação ao PLS está na validação cruzada, na qual a escolha do número de VLs é feita não em função de um parâmetro quantitativo (RMSECV), mas do menor número de amostras incorretamente classificadas, um parâmetro qualitativo. As críticas ao frequente uso do PLS-DA em problemas de autenticação deram origem ao desenvolvimento recente de novos métodos de modelagem de classe, baseados no PLS ou no SIMCA, tais como PLS-DM (*PLS density modeling*) (Oliveri et al., 2014), OC-PLS (*one-class PLS*) (Xu et al., 2013) e DD-SIMCA (*data-driven SIMCA*) (Pomerantsev; Rodionova, 2014). Finalmente, um último tópico que merece destaque é a validação analítica de métodos qualitativos (Isabel López et al., 2015), que se baseia na estimativa de figuras de mérito específicas para a análise qualitativa, tais como a sensibilidade (taxa de verdadeiro positivo) e especificidade ou seletividade (taxa de verdadeiro negativo).

Conclusão e perspectivas

Este capítulo descreveu os mais importantes tipos de métodos quimiométricos que podem ser combinados com espectros NIR para gerar modelos úteis na extração de informações de grandes conjuntos de dados. A difusão de métodos analíticos baseados em espectroscopia NIR propicia análises cada vez mais simples, rápidas e ambientalmente amigáveis. Essa difusão é constatada através do grande aumento de publicações científicas. Mas existe um potencial também para o crescimento do uso prático de tecnologias NIR em laboratórios de controle de qualidade de indústrias e órgãos reguladores. Espera-se que, no futuro, este tipo de tecnologia se torne mais acessível em toda a cadeia produtiva, do produtor ao consumidor. No setor agrícola, em particular, duas novas tendências são importantes. Primeiro, o desenvolvimento de espectrofotômetros NIR portáteis, cada vez menores, que permitem a obtenção de medidas no campo, tornando as análises mais simples e versáteis (Santos et al., 2013). O uso de equipamentos portáteis aumenta a demanda pelo desenvolvimento de modelos de transferência de calibração (Honorato et al., 2007), os quais permitem generalizar o uso de um método quantitativo também para espectrofotômetros de bancada, que ainda costumam fornecer resultados mais precisos. Outra tendência importante é o uso de imagens hiperespectrais, as quais permitem, por exemplo, obter mapas de concentração para analitos distribuídos ao longo da superfície de amostras heterogêneas (Burger; Gowen, 2011).

Referências

ASTM. E1655-05: standards practices for infrared multivariate quantitative analysis. West Conshohocken, 2012. Annual Book of ASTM Standards.

BOTELHO, B. G.; MENDES, B. A. P.; SENA, M. M. Implementação de um método robusto para o controle fiscal de umidade em queijo Minas artesanal: abordagem metrológica multivariada. **Química Nova**, v. 36, n. 9, p. 1416-1422, 2013.

BRERETON, R. G. Introduction to multivariate calibration in analytical chemistry. **Analyst**, v. 125, n. 11, p. 2125-2154, 2000.

BRERETON, R. G.; LLOYD, G. R. Partial least squares discriminant analysis: taking the magic way. **Journal of Chemometrics**, v. 28, n. 4, p. 213-225, 2014.

BRO, R.; SMILDE, A. K. Centering and scaling in component analysis. **Journal of Chemometrics**, v. 17, n. 1, p.16-33, 2003.

BRO, R.; SMILDE, A. K. Principal component analysis. **Analytical Methods**, v. 6, n. 9, p. 2812-2831, 2014.

BRUNS, R. E.; FAIGLE, J. F. G. Quimiometria. **Química Nova**, v. 8, n. 2, p. 84-99, 1985.

BURGER, J.; GOWEN, A. Data handling in hyperspectral image analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 108, n. 1, p. 13-22, 2011.

COZZOLINO, D. An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals. **Food Research International**, v. 60, p. 262-265, June 2014.

DAVIES, T. The history of near infrared spectroscopic analysis: past, present and future – “From sleeping technique to the morning star of spectroscopy”. **Anulusis**, v. 26, n. 4, p. M17-M19, 1998.

DHANOA, M. S.; LISTER, S. J.; SANDERSON, R.; BARNES, R. J. The link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) transformations of NIR spectra. **Journal of Near Infrared Spectroscopy**, v. 2, n. 1, p. 43-47, 1994.

ENGEL, J.; GERRETZEN, J.; SZYMANSKA, E.; JANSEN, J. J.; DOWNEY, G.; BLANCHET, L.; BUYDENS, L. M. C. Breaking with trends in pre-processing? **TrAC – Trends in Analytical Chemistry**, v. 50, p. 96-106, Oct. 2013.

FABER, N. M.; RAJKÓ, R. How to avoid over-fitting in multivariate calibration – the conventional validation approach and an alternative. **Analytica Chimica Acta**, v. 595, n. 1/2, p. 98-106, 2007.

FERREIRA, M. M. C. **Quimiometria:** conceitos, métodos e aplicações. Campinas: Ed. UNICAMP, 2015. 833 p.

GABRIEL, K. R. The biplot display of matrices with application to principal component analysis. **Biometrika**, v. 58, n. 3, p. 453-467, 1971.

GELADI, P.; KOWALSKI, B. R. Partial least squares regression: a tutorial. **Analytica Chimica Acta**, v. 185, p. 1-17, 1986.

GELADI, P.; ESBENSEN, K. The start and early history of chemometrics: selected interviews. Part 1. **Journal of Chemometrics**, v. 4, n. 5, p. 337-354, 1990.

HIBBERT, D. B. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). **Pure and Applied Chemistry**, v. 88, n. 4, p. 407-443, 2016.

HONORATO, F. A.; BARROS NETO, B.; MARTINS, M. N.; GALVÃO, R. K. H.; PIMENTEL, M. F. Transferência de calibração em métodos multivariados, **Química Nova**, v. 30, n. 5, p. 1301-1312, 2007.

ISABEL LÓPEZ, M.; PILAR CALLAO, M.; RUISÁNCHEZ, I. A tutorial on the validation of qualitative methods: from the univariate to the multivariate approach. **Analytica Chimica Acta**, v. 891, p. 62-72, Sep. 2015.

MARINI, F. Classification methods in chemometrics. **Current Analytical Chemistry**, v. 6, n. 1, p. 72–79, 2010.

MARQUETTI, I.; LINK, J. D.; LEMES, A. L. G.; SCHOLZ, M. B. S.; VALDERRAMA, P.; BONA, E. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. **Computers and Electronics in Agriculture**, v. 121, p. 313-319, Feb. 2016.

MARTENS, H.; NAES, T. **Multivariate calibration**. Chichester: Wiley, 1989. 419 p.

MCCLURE, W. F. Near-infrared spectroscopy. The giant is running strong. **Analytical Chemistry**, v. 66, n. 1, p. 43A-53A, 1994.

NORRIS, K. H.; HART, J. H. Direct spectrophotometric determination of moisture content of grain and seeds. **Journal of Near Infrared Spectroscopy**, v. 4, n. 1, p. 23-30, 1996.

OLIVERI, P.; ISABEL LÓPEZ, M.; CASOLINO, M. C.; RUISÁNCHEZ, I.; PILAR CALLAO, M.; MEDINI, L.; LANTERI. S. Partial least squares density modeling (PLS-DM) - A new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy. **Analytica Chimica Acta**, v. 851, p. 30-36, Dec. 2014.

OTTO, M. **Chemometrics**: statistics and computer application in analytical chemistry. Weinheim: Wiley-VCH, 1999. 343 p.

PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 2, p. 198-219, 2003.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, n. 11, p. 559-572, 1901.

POMERANTSEV, A. L.; RODIONOVA, O. Y. On the type II error in SIMCA method. **Journal of Chemometrics**, v. 28, n. 6, p. 518-522, 2014.

RINNAN, A.; VAN DEN BERG,F.; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC - Trends Analytical Chemistry**, v. 28, p. 1201–22, Nov. 2009

RODIONOVA, O. Y.; OLIVERI, P.; POMERANTSEV, A. L. Rigorous and compliant approaches to one-class classification. **Chemometrics and Intelligent Laboratory Systems**, v. 159, p. 89-96, Dec. 2016.

SANCHEZ, G. **The saga of PLS**. 2015. Disponível em: <http://sagaofpls.github.io>. Acesso em: 9 fev. 2018.

SANTOS, C. A. T.; LOPO, M.; PÁSCOA, R. N. M. J.; LOPES, J. A. A review on the applications of portable near-infrared spectrometers in the agro-food industry. **Applied Spectroscopy**, v. 67, n. 11, p. 1215-1233, 2013.

SAVITZKY, A.; GOLAY, M. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627-1639, 1964.

SENA, M. M.; FRIGHETTO, R. T. S.; VALARINI, P. J.; TOKESHI, H.; POPPI, R. J. Discrimination of management effects on soil parameters by using principal component analysis: a multivariate analysis case study. **Soil & Tillage Research**, v. 67, n. 2, p. 171-181, 2002.

SENA, M. M.; ALMEIDA, M. R.; BRAGA, J. W. B.; POPPI, R. J. Multivariate statistical analysis and chemometrics. In: FRANCA, A. S.; NOLLET, L. M. L. (Org.). **Spectroscopic methods in food analysis**. Boca Raton: CRC Press, 2017. p. 273-314.

SOUZA, A. M.; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I. **Química Nova**, v. 35, n. 1, p. 223-229, 2012.

SZYMANSKA, E.; GERRETAEN, J.; ENGEL, J.; GEURTS, B.; BLANCHET, L.; BUYDENS. L. M. C. Chemometrics and qualitative analysis have a vibrant relationship. **TrAC – Trends in Analytical Chemistry**, v. 69, p. 34-51, June 2015.

VALDERRAMA, P.; BRAGA, J. W. B.; POPPI, R. J. Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy. **Journal of Agricultural and Food Chemistry**, v. 55, n. 21, p. 8331-8338, 2007.

VALDERRAMA, P.; BRAGA, J. W. B.; POPPI, R. J. Estado da arte de figuras de mérito em calibração multivariada. **Química Nova**, v. 32, n. 5, p. 1278-1287, 2009.

VITALE, R.; BEVILACQUA, M.; BUCCI, R.; MAGRÌ, A. D.; MARINI, F. A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometric. **Chemometrics and Intelligent Laboratory Systems**, v. 121, p. 90-99, Feb. 2013.

WESTAD, F.; MARINI. F. Validation of chemometric models – a tutorial. **Analytica Chimica Acta**, v. 893, p. 14-24, Set. 2015.

WETZEL, D. L. Near-infrared reflectance analysis - sleeper among spectroscopic techniques. **Analytical Chemistry**, v. 55, n. 12, p. 1165A-1176A, 1983.

WOLD, S. Pattern recognition by means of disjoint principal component analysis. **Pattern Recognition**, v. 8, n. 3, p. 127-139, 1976.

XU, L.; YAN, S. M.; CAI, C. B.; YU, X. P. One-class partial least squares (OCPLS) classifier. **Chemometrics and Intelligent Laboratory Systems**, v. 126, p. 1-5, 2013.