Quim. Nova, Vol. 43, No. 3, 371-381, 2020 http://dx.doi.org/10.21577/0100-4042.20170480

Educação

EXPERIMENTO DIDÁTICO DE QUIMIOMETRIA PARA CLASSIFICAÇÃO DE ÓLEOS VEGETAIS COMESTÍVEIS POR ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO COMBINADO COM ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS: UM TUTORIAL, PARTE V

Felipe Bachion de Santana[a], André Marcelo de Souza[b], Mariana Ramos Almeida[c], Márcia Cristina Breitkreitz[a], Paulo Roberto Filgueiras[d], Marcelo Martins Sena[c] e Ronei Jesus Poppi[a,*]

[a]Instituto de Química, Universidade Estadual de Campinas, 13084-971 Campinas – SP, Brasil
[b]Embrapa Solos, Empresa Brasileira de Pesquisa Agropecuária, 22460-000 Rio de Janeiro – RJ Brasil
[c]Departamento de Química, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte – MG, Brasil
[d]Departamento de Química, Universidade Federal do Espírito Santo, 29075-910 Vitória – ES, Brasil

Recebido em 29/08/2019; aceito em 18/11/2019; publicado na web em 17/02/2020

DIDACTIC EXPERIMENT OF CHEMOMETRICS FOR THE CLASSIFICATION OF EDIBLE VEGETABLE OILS BY FOURIER TRANSFORM INFRARED SPECTROSCOPY AND PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS: A TUTORIAL, PART V. A teaching experiment on supervised pattern recognition in chemometrics was proposed in this tutorial to introduce partial least squares discriminant analysis (PLS-DA). A new approach of the experiment published in the first tutorial of this series was revisited and employed to the classification of edible vegetable oils. The spectra of olive, canola, soybean and corn oils were obtained using an attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectrometer in the range of 600 to 4000 cm^{-1}. The combination of ATR-FTIR and PLS-DA classification method was able to correctly classify 100% of the validation samples. The Matlab commands, routines and functions were presented, and a didactic explanation of the concepts and interpretation of the data was provided.

Keywords: didactic experiment; infrared spectroscopy; chemometrics; discriminant analysis; partial least squares.

INTRODUÇÃO

As bases para o estabelecimento e a consolidação da Quimiometria estão fundamentalmente pautadas nos avanços tecnológicos e computacionais que impactaram os instrumentos analíticos modernos a partir das últimas décadas. A Química Analítica moderna desfruta de um vasto e significativo arsenal de instrumentos avançados que possibilitam a obtenção de respostas instrumentais intrinsecamente multivariadas, produzindo uma grande quantidade de dados uni e multidimensionais de forma muito rápida e eficiente a partir de uma única medida. A consequência imediata desse avanço foi o aumento significativo da complexidade dos conjuntos de dados gerados por esses instrumentos. Devido à necessidade de extrair informações confiáveis e relevantes de forma rápida e eficiente, cada vez mais indústrias e pesquisadores estão utilizando métodos Quimiométricos de análise.[1]

Com esse cenário, torna-se urgente a formação de estudantes de graduação e pós-graduação e a capacitação e treinamento de profissionais da indústria em análise de dados multivariados de origem química. Objetivando colaborar para isso, desde 2012 está sendo publicada uma sequência de tutoriais no periódico nacional Química Nova para o ensino de conceitos básicos de Quimiometria, buscando alimentar a literatura científica nacional com guias práticos para cursos de graduação, pós-graduação e capacitação de profissionais. Foram publicados até o presente momento quatro tutoriais apresentando os principais métodos quimiométricos empregados em: análise exploratória,[2] calibração multivariada,[3] resolução de curvas[4] e planejamento e otimização de experimentos.[5] Dando prosseguimento às publicações, neste tutorial será apresentado um dos métodos supervisionados de reconhecimento de padrões, ou de classificação supervisionada multivariada, mais utilizados em química analítica: a análise discriminante por mínimos quadrados parciais (*Partial Least Squares Discriminant Analysis*, PLS-DA).[6,7]

Os métodos de reconhecimento de padrões podem ser classificados como métodos supervisionados e não-supervisionados. No tutorial I[2] foi apresentado o principal método não supervisionado em análise multivariada, a análise de componentes principais (*Principal Component Analysis*, PCA), aplicada na avaliação de óleos vegetais comestíveis por espectroscopia no infravermelho médio (*Mid Infrared Spectroscopy*, MIRS). Nos métodos supervisionados são construídos modelos de classificação utilizando amostras com características conhecidas, em seguida, o modelo é utilizado para prever a classe de amostras desconhecidas. Os principais métodos supervisionados são: análise discriminante linear (*Linear Discriminan tAnalysis*, LDA), *k*-vizinhos mais próximo (*k-Nearest Neighbor*, kNN), PLS-DA e modelagem independente e flexível por analogia de classe (*Soft Independent Modeling of Class Analogy*, SIMCA).

Neste tutorial, o método PLS-DA será empregado na classificação de óleos vegetais comestíveis de acordo com sua oleaginosa utilizando seus espectros na região do infravermelho médio, resgatando o experimento do tutorial I. Os espectros foram obtidos empregando um espectrômetro no infravermelho médio com transformada de Fourier (FTIR, do inglês *Fourier transform infrared*). Na literatura, são encontrados diversos exemplos dessa abordagem, seja na classificação dos óleos vegetais de acordo com sua região,[8] oleaginosa,[9] ou na discriminação entre óleos autênticos e adulterados com óleos vegetais de menor valor econômico.[10,11]

A grande quantidade de estudos envolvendo este tipo de aplicação pode ser atribuída à dificuldade em distinguir os óleos vegetais de acordo com a sua oleaginosa, uma vez que a maioria dos óleos vegetais comestíveis possuem propriedades físico-químicas semelhantes e a análise da composição de ácidos graxos e dos parâmetros físico-químicos podem não ser suficientes para identificá-los, requerendo métodos alternativos para classificar o óleo vegetal de acordo com a sua oleaginosa.[12]

*e-mail: ronei@iqm.unicamp.br

A facilidade de obtenção dos espectros FTIR empregando o dispositivo amostrador de reflectância total atenuada (ATR, do inglês *attenuated total reflectance*), combinada ao método de classificação PLS-DA, é apresentada como um grande atrativo, uma vez que, além da simplicidade e curto tempo de análise, essa combinação apresenta, de modo geral, excelentes resultados de classificação dos óleos vegetais empregando o perfil espectral como impressão digital da amostra.

A espectroscopia ATR-FTIR, região de 4000–400 cm^{-1}, é considerada uma técnica rápida (gera respostas em segundos), não destrutiva e sem ou quase nenhum preparo de amostra, além de apresentar menor custo quando comparada a técnicas de cromatografia ou ressonância magnética nuclear, entre outras empregadas nesse tipo de análise. Seus espectros fornecem informações das ligações moleculares em frequências especificas de cada grupo funcional. No entanto, essas frequências também são influenciadas (deslocadas) pela presença dos grupos funcionais próximos (acoplamentos), atuando como uma impressão digital de uma dada amostra quando utilizadas integralmente.[2]

Uma das principais regiões espectrais presentes nessa faixa é a região de impressão digital (1.200 a 600 cm^{-1}), na qual pequenas diferenças na estrutura e na constituição de uma molécula resultam em mudanças significativas na distribuição das bandas de absorção.[13] Mais informações a respeito da técnica podem ser encontradas em Skoog.[13] Devido a tais vantagens e por apresentar custo relativamente baixo, essa técnica é largamente empregada em indústrias e laboratórios de controle de qualidade para diversos fins, entre eles: caracterização de compostos orgânicos, quantificação de analitos, classificação e autenticação de diversas matrizes, incluindo os óleos vegetais.

Análise discriminante por mínimos quadrados parciais

O PLS-DA é um método de classificação supervisionado, ou seja, é um método que determina a qual classe pertence uma amostra desconhecida a partir das informações fornecidas ao sistema, neste caso, os espectros na região do infravermelho médio e a classe original das amostras. Os métodos supervisionados exigem o conhecimento inicial sobre as amostras e suas classes para definir as regras que serão utilizadas para a classificação das amostras.[7]

O método de classificação PLS-DA utiliza a técnica de regressão multivariada por mínimos quadrados parciais (PLS), a qual já foi discutida no tutorial II.[3] O PLS é um método de calibração inversa, no qual se busca uma relação direta entre a resposta instrumental (matriz **X**) e a propriedade de interesse (matriz **Y** ou vetor **y**). O procedimento utilizado para a construção do modelo de classificação é o mesmo utilizado pelo PLS, no entanto, a propriedade de interesse em modelos de classificação é uma variável categórica que descreve a atribuição de classe da amostra. Geralmente, o valor 1 é atribuído à classe de interesse e o valor 0 é atribuído à outra classe.

A Figura 1 ilustra como é feita a organização dos dados utilizados para a construção do modelo de classificação PLS-DA. A matriz **X** é composta pelos espectros, sendo que suas linhas representam as amostras e suas colunas representam as absorbâncias para cada número de onda. O vetor **y** é construído com valores 1 ou 0.

Existem duas variantes da PLS-DA, a PLS1-DA e a PLS2-DA. Na primeira, cada coluna de **Y** é modelada individualmente. Caso existam três classes "a, b, c", são construídos três modelos. O primeiro modelo será construído utilizando valores de **Y** igual a 1 para a classe "a" e 0 para as demais classes, o segundo modelo será construído utilizando valores 1 para a classe "b" e zero para as demais classes, o mesmo procedimento será empregado para a terceira classe. Já na PLS2-DA, é calculado um único conjunto de escores e pesos para todas as colunas da matriz **Y**, o que leva à restrição de usar o mesmo número de variáveis latentes para modelar todas as classes.

Neste tutorial, o método de classificação supervisionado PLS2-DA será utilizado para classificar amostras de óleos vegetais de 4 diferentes oleaginosas. Caso fosse utilizado o algoritmo PLS1-DA, seria necessário o desenvolvimento de 4 diferentes modelos de classificação, tornando o método classificação mais trabalhoso. De modo geral, o algoritmo PLS1-DA é utilizando em problemas binários de classificação, enquanto em problemas multiclasse geralmente é utilizado o algoritmo PLS2-DA, uma vez que um único modelo é desenvolvido para as todas as amostras. Devido a esse fato, neste tutorial será empregada somente a PLS2-DA e sempre que neste artigo for feita referência ao método PLS-DA, estaremos nos referindo ao método PLS2-DA.

O algoritmo dos quadrados mínimos parciais iterativos não lineares (NIPALS, do inglês, *non linear iterative partial leasts quares*) pode ser utilizado pelo método PLS2 para decompor as matrizes **X** e **Y**.[14]

Empregando esse algoritmo, a decomposição das matrizes **X** e **Y** são realizadas empregando as Equações 1 e 2, em que **T** e **U** são os escores, **P** e **Q** os pesos. Vale relembrar que os escores representam as coordenadas das amostras na projeção das variáveis latentes e os pesos são os coeficientes da combinação linear das variáveis originais e são interpretados como a contribuição de cada variável original em cada variável latente.[15,16] e são os resíduos de **X** e **Y** respectivamente, ou seja, são as matrizes que contêm a parte não modelada de **X** e **Y**.

$$\mathbf{X} = \mathbf{TP}^{T} + \mathbf{R}_{X} \quad (1)$$

$$\mathbf{Y} = \mathbf{UQ}^{T} + \mathbf{R}_{y} \quad (2)$$

Os pesos **W** (*weights*) do PLS são calculados proporcionalmente à covariância entre os blocos **X** e **Y**, conforme representado na Equação 3. Na Equação 3 o símbolo "‖ ‖" é referente ao cálculo da norma. A matriz de escores **T** é estimada pela combinação linear de

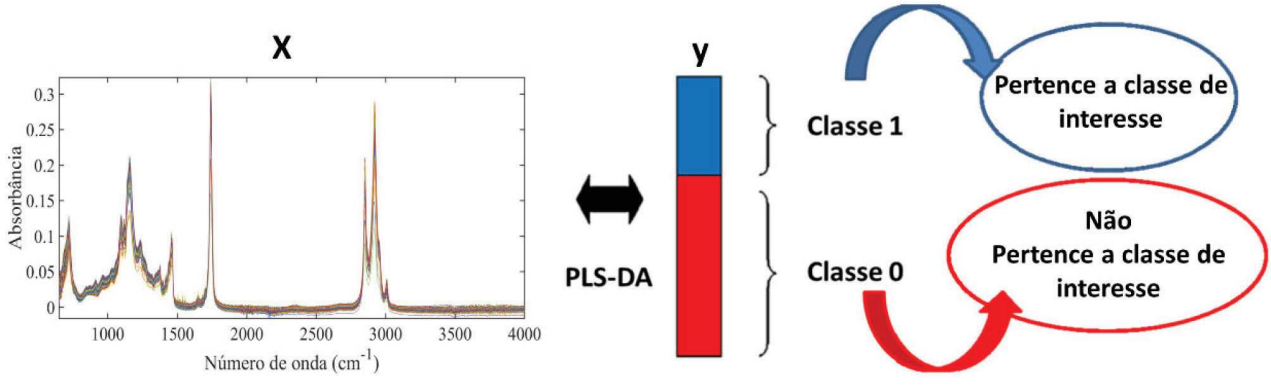

Figura 1. *Esquema da organização dos dados para a construção do modelo de classificação usando PLS-DA*

X com a matriz de pesos **W** do PLS**.** Os escores (Equação 4) são calculados simultaneamente pela decomposição das matrizes **X** e **Y**, reunindo os vetores coluna da matriz **W**, que denotam as direções das variáveis latentes.[14]

$$\mathbf{W} = \frac{\mathbf{X}^T\mathbf{U}}{\|\mathbf{X}^T\mathbf{U}\|} \quad (3)$$

$$\mathbf{T} = \mathbf{XW} \quad (4)$$

Após a estimativa dos pesos **W** e dos escores **T**, são calculados os pesos **Q** e **P** e os escores **U**, conforme representado nas Equações 5, 6 e 7, respectivamente.

$$\mathbf{Q} = \frac{\mathbf{Y}^T\mathbf{T}}{\|\mathbf{U}^T\mathbf{T}\|} \quad (5)$$

$$\mathbf{P} = \frac{\mathbf{X}^T\mathbf{T}}{\|\mathbf{T}^T\mathbf{T}\|} \quad (6)$$

$$\mathbf{U} = \mathbf{YQ} \quad (7)$$

Em seguida, são estimadas as matrizes de resíduos $\mathbf{R_X}$ e $\mathbf{R_y}$, e os valores de **X** e **Y** são atualizados para o cálculo da próxima variável latente. Ao final do cálculo, os coeficientes de regressão do modelo PLS são estimados através da Equação 8 e os valores obtidos de $\hat{\mathbf{Y}}$ são calculados através da Equação 9. Mais informações a respeito do algoritmo NIPALS para PLS2 podem ser encontradas em Geladi.[14] A descrição do algoritmo NIPALS para PLS1 está disponivel no Material Suplementar e mais informações a respeito deste algoritmo pode ser encontrada em Ferreira.[16]

$$\mathbf{B}_{PLS} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T \quad (8)$$

$$\hat{\mathbf{Y}} = \mathbf{X}.\mathbf{B}_{PLS} \quad (9)$$

Assim como no PLS, uma das etapas principais para construir o modelo PLS-DA é a escolha correta do número de variáveis latentes. Quando é escolhido um número insuficiente de variáveis latentes, há falta de ajuste no modelo (subajuste), ou seja, não é utilizada toda a informação útil para sua construção. Todavia, quando é escolhido um número excessivo de variáveis latentes, tem-se o sobreajuste, situação em que são incorporadas informações não relacionadas à propriedade de interesse, como por exemplo ruídos espectrais, na construção do modelo.[17]

A escolha do número de variáveis latentes é comumente feita através da validação cruzada, na qual uma parte das amostras do conjunto de calibração/treinamento é separada e utilizada para validação interna. Em seguida, é construído o modelo de treinamento utilizando diferentes números de variáveis latentes e são previstas as amostras de validação interna para cada um dos modelos construídos. Os erros obtidos são armazenados e o processo é repetido até que todas as amostras de treinamento sejam previstas.

Enquanto no PLS são analisados os valores dos erros médios quadráticos de validação cruzada (RMSECV, do inglês *root mean square errors of cross-validation*) para definir o número ideal de variáveis latentes, na PLS-DA analisam-se a porcentagem de amostras classificadas corretamente em cada classe na validação cruzada. A Figura 2 ilustra de maneira genérica a região ideal para a escolha do número de variáveis latentes.

Os resultados fornecidos pela matriz $\hat{\mathbf{Y}}$ (Equação 9) representam a previsão da classe à qual pertence cada amostra. No entanto, como a variável categórica é modelada de maneira contínua, na prática, os valores previstos não são exatamente 0 ou 1.[18] Dessa forma, é necessário estabelecer um valor limite (limiar/*threshold*) entre as classes. Esse valor pode ser estimado utilizando curvas ROC

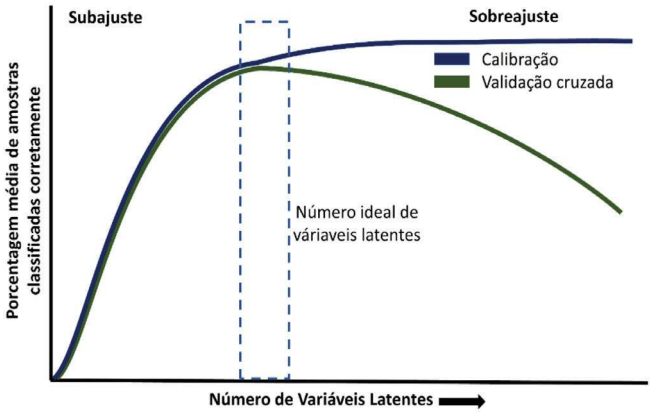

Figura 2. *Porcentagem média de amostras classificadas corretamente em função do número de variáveis latentes*

(*receiver operating characteristic*)[19] ou mais comumente empregando estatística Bayesiana.[18]

Para isso, é construído inicialmente um histograma dos valores previstos de $\hat{\mathbf{Y}}$ para as amostras da classe 0 e 1, conforme ilustrado na Figura 3, em que a classe 0 é representada pela cor azul e a classe 1 é representada pela cor verde. Em seguida, é ajustada uma distribuição normal para cada uma das classes. Após isso, utiliza-se essa distribuição para calcular qual a probabilidade da amostra pertencer à classe 0 ou 1 em função do valor previsto pelo modelo. A linha em azul representa a probabilidade da amostra pertencer à classe 0 em função do seu valor previsto e a linha verde representa a probabilidade da amostra pertencer à classe 1. No cruzamento destas linhas temos o ótimo limiar para a discriminação, minimizando o número de resultados falsos positivos e negativos.[18] Nota-se que a Bayesiana considera que os valores previstos de $\hat{\mathbf{Y}}$ para cada classe (classes 0 e 1) seguem uma distribuição normal.

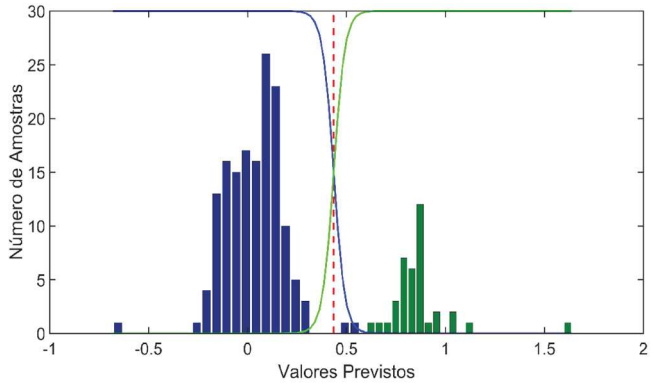

Figura 3. *Distribuição dos valores previstos de* ***Y*** *para determinação do valor limite entre as classes (verde/1 e azul/0) seguindo a estatística Bayesiana. A linha tracejada em vermelho representa o limiar* ***ótimo*** *para discriminação*

Outra etapa muito importante é a etapa de validação/teste do modelo de classificação empregando amostras externas, ou seja, que não foram empregadas na etapa de treinamento. É importante que se conheça previamente a classe dessas amostras para avaliar o desempenho do modelo. Nessa etapa, se o número de variáveis latentes escolhido na etapa de treinamento não for adequado, a classificação das novas amostras não será correta.[20]

Conforme já reportados nos tutoriais anteriores, durante a construção dos modelos multivariados o pré-processamento dos dados é extremamente importante e deve ser reportado.[2-4] Neste tutorial, os espectros dos óleos vegetais na região do infravermelho médio serão pré-processados empregando a primeira derivada com suavização por

meio do algoritmo Savitzky-Golay (janela de 9 pontos e polinômio de 2º grau).[21] A combinação da derivada com a suavização dos espectros é muito útil para atenuar desvios de linha de base e ruídos espectrais, além de ressaltar a variação das bandas espectrais. Após o derivar e suavizar os espectros eles serão centrados na média.

Podem ser encontrados na literatura outros tipos de pré-processamentos como normalização dos espectros, MSC (Multiplicative Scatter Correction), SNV (Standard Normal Variate), entre outros. A escolha do pré-processamento deve ser realizada em função da técnica analítica, tipo de amostra e método quimiométrico. Uma maior discussão a respeito dos pré-processamentos utilizados em espectroscopia vibracional pode ser encontrada em Rinnan, Åsmund.[22]

Avaliação de amostras anômalas

Antes de iniciar a construção do modelo multivariado de classificação é importante realizar a análise visual dos dados espectrais para identificar possíveis erros grosseiros dos dados. Esta visualização deve ser realizada antes e após o pré-processamento dos dados. Caso existam amostras com erros grosseiros de análise, estas devem ser removidas antes de iniciar a construção do modelo de classificação. Além deste processo, a identificação de amostras anômalas (*outliers*) também pode ser avaliada, no caso do PLS-DA, pelo uso da estatística Q (resíduos) e T^2 (*Hotelling*) nos conjuntos de treinamento e teste a nível de significância de 5%, ou seja, nível de confiança de 95%.[23]

Os resíduos da soma quadrática (Equação 10), também conhecidos como resíduos Q, representam a parte do bloco **X** não modelada pelo PLS-DA, sendo Q calculado, para cada amostra, utilizando a Equação 10.

$$Q = \mathbf{r}_i \mathbf{r}_i^T \tag{10}$$

em que $\mathbf{r}_i$ é a *i*-linha da matriz de resíduos $\mathbf{R_X}$ dada na Equação (1). Existem diversas maneiras de determinar os limites de confiança para os resíduos Q.[24] Neste tutorial, o valor crítico será calculado de acordo com o método proposto por Jackson e Mudholkar.[25]

A medida de T^2 de *Hotelling* está relacionada com a distância da amostra até o centro dos dados, e seu cálculo é realizado para cada amostra empregando a Equação 11.

$$T^2_{Hotelling} = \frac{\mathbf{t}_i^T \left(\mathbf{T}^T\mathbf{T}\right)^{-1} \mathbf{t}_i}{I-1} \tag{11}$$

em que **T** é a matriz de escores das amostras do conjunto de treinamento com I amostras e R variáveis latentes e $\mathbf{t}_i$ é a *i*-linha da matriz **T**. O valor crítico (Equação 12) é calculado com a suposição que os escores apresentam uma distribuição normal, onde representa o nível de significância adotado, neste caso 5%.[26]

$$T^2_{crítico} = \frac{R\left(I-1\right)}{I-R} F_{R,I-R,\alpha} \tag{12}$$

em que F é o valor tabelado da distribuição F com R e I-R graus de liberdade.

Amostras que apresentarem simultaneamente valores de resíduos Q e T^2 de Hotelling acima dos valores críticos serão consideradas anômalas pelo modelo e, consequentemente, excluídas.

Avaliação do modelo de classificação

O desempenho do modelo de classificação é avaliado através de tabelas de contingência, que mostram o número de amostras previstas em cada classe e seus respectivos valores de referência. Um exemplo desse tipo de tabela para um conjunto de duas classes é ilustrado na Tabela 1. Uma amostra verdadeiramente positiva (VP) é uma amostra da classe 1 (alvo) que foi corretamente classificada como positiva. Amostras falso positivas (FP) são amostras negativas (classe 0) que foram erroneamente previstas como positivas (classe 1). O mesmo raciocínio é aplicado às amostras verdadeiramente negativas (VN) e falso negativas (FN). Quando houver mais de duas classes a serem previstas (como será o caso deste artigo) a tabela de contingência é dada pelo número de amostras previstas em cada classe.

Tabela 1. Exemplo de uma tabela de contingência para duas classes

Tabela de Contingência 2x2		Previsto pelo modelo	
		Classe 1	Classe 0
Referência	Classe 1	VP	FP
	Classe 0	FN	VN

Da tabela de contingência é possível extrair as figuras de méritos utilizadas para avaliar os modelos de classificação, que são parâmetros como as taxas de sensibilidade, especificidade/seletividade e acurácia/eficiência.[15] Esses parâmetros, embora às vezes possuam o mesmo nome, diferem dos parâmetros empregados em modelos quantitativos. Em problemas de classificação, a sensibilidade é a habilidade do modelo classificar corretamente amostras positivas dado que elas são realmente positivas, e é calculada empregando a Equação 13.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{13}$$

A especificidade (também denominada seletividade) é a capacidade do modelo em identificar corretamente as amostras negativas dado que elas são negativas, e é calculada através da Equação 14.

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{14}$$

A acurácia (também denominada eficiência) é um parâmetro estatístico que fornece um único valor global para medir o desempenho do modelo de classificação. Seu cálculo é realizado através do número de amostras classificadas corretamente independentemente da classe dividido pelo número total de amostras, conforme mostra a Equação 15:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{15}$$

Uma discussão mais aprofundada sobre essas e outras figuras de mérito para validação analítica de métodos qualitativos, tanto univariados como multivariados, pode ser encontrada em um artigo tutorial recente de Isabel López.[27] Seguindo a linha da série de tutoriais anteriores, neste artigo será dada ênfase aos aspectos práticos de construção de um modelo multivariado de classificação.

PARTE EXPERIMENTAL

Amostras

As amostras de azeite de oliva e os óleos de canola, milho e soja foram adquiridas em comércios locais (Campinas, SP). Foram adquiridas amostras de diferentes marcas e lotes, totalizando 108 de azeite e 54 para cada um dos demais tipos de óleos vegetais.

Obtenção dos espectros ATR-FTIR

Os espectros de infravermelho foram obtidos utilizando o espectrômetro Cary 630 FTIR Agilent, equipado com acessório de reflectância total atenuada (ATR) composto pelo cristal de ZnSe. Os

espectros foram obtidos na faixa de 600 a 4000 cm^{-1}, com resolução de 4 cm^{-1} e 32 varreduras. Antes da leitura de cada amostra, o espectro do ar (ATR vazio) obtido e tomado como a medida do branco. Entre as aquisições dos espectros, o dispositivo amostrador ATR foi limpo com acetona e algodão, e essa limpeza foi monitorada através do *software* de aquisição dos espectros.

Softwares e algoritmos

Assim como nos tutoriais anteriores, o PLS-DA foi executado em ambiente computacional Matlab (MathWorks, Natick, MA, EUA). O conjunto de amostras, as rotinas e as funções necessárias para executar o tutorial estão disponíveis no *GitHub* dos autores[28] ou podem ser requisitadas pelo e-mail *rjpoppi@unicamp.br*. Recomendamos abrir a rotina "Tutorial_PLSDA" para acompanhar o passo-a-passo presente neste tutorial. As funções e as rotinas disponibilizadas foram testadas nas versões do Matlab 14b, 16a,16b e 17a. O *GitHub* é uma plataforma de hospedagem de código-fonte do tipo *Open Source*. Essa plataforma é muito utilizada por programadores para disponibilizar e colaborar com programas/rotinas de outros usuários.

RESULTADOS E DISCUSSÃO

Os espectros foram organizados em uma matriz (matriz **X**). Neste tutorial, a matriz de dados **X** e os comprimentos de onda dos espectros estão presentes no arquivo "amostras.mat" com os nomes "X" e "num", respectivamente.
(1) O primeiro passo consiste em carregar o conjunto de dados para o *Workspace* do Matlab:

```
>>load amostras
```

(2) Em seguida, é construído o gráfico contendo todos os espectros presentes na matriz **X** (Figura 4):

```
>>plot (num,X);
>>xlabel ('Número de Onda (cm^-^1)');
>>ylabel ('Absorbância');
```

As primeiras 108 amostras da matriz **X** são as amostras de azeite de oliva. Nas linhas de 109 a 162 (54 amostras) estão as amostras de óleo de canola, nas linhas 163 a 216 (54 amostras) estão as amostras de óleo de milho, e nas linhas 217 a 270 (54 amostras) estão as amostras de óleo de soja. Um vetor de classes (vetor **y**) deve ser construído para identificar a classe de cada amostra. As classes 1, 2, 3 e 4 irão representar as amostras de azeite de oliva, óleo de canola, óleo de milho e óleo de soja respectivamente.

(3) Criar o vetor de classes:

```
>>y = [ones(108,1);2*ones(54,1);3*ones(54,1);4*ones(54,1)];
```

Os espectros dos óleos vegetais contêm regiões que não apresentam informações químicas (linha de base) e regiões que apresentam pouca ou nenhuma variação. Tais regiões devem ser removidas para facilitar a construção do modelo de classificação. Neste tutorial, será utilizada uma região espectral um pouco maior do que a região da impressão digital (750 a 1580 cm^{-1}) para construir o modelo de classificação. Com o objetivo de separar amostras que representem toda a variabilidade dos dados para os conjuntos de treinamento e teste será utilizado o algoritmo Kennard-Stone.[29] Esse algoritmo seleciona o conjunto de amostras mais representativas e distribuídas homogeneamente baseado nas distâncias Euclidianas entre elas. As amostras mais representativas são utilizadas para o conjunto de treinamento e as amostras menos representativas são empregadas no conjunto teste. Em modelos de classificação, recomenda-se aplicar o algoritmo de Kennard-Stone separadamente à cada classe.

(4) Para realizar o corte dos espectros e selecionar os conjuntos de treinamento (70% das amostras de cada classe) e teste (30% das amostras de cada classe) será utilizado a função "caltestda":

```
>>[~,xcal,xval,ycal,yval] = caltestda (X(:,1300:1750),y,70,'k',[],[]);
```

Maiores informações a respeito da função "caltestda" podem ser obtidas digitando o comando:

```
>>help caltestda
```

Sempre que é realizado o corte dos espectros, também é necessário realizar o corte do vetor que contém a faixa espectral, denominado nesse artigo de vetor "num":

```
>>numc = num(1300:1750);
```

A matriz "xcal" é constituída dos espectros de treinamento de cada classe dos óleos vegetais já cortados na região de (750 a 1580 cm^{-1}), enquanto a matriz "xval" é constituída das amostras de teste. Os vetores "ycal" e "yval" contêm informações a respeito das classes das amostras de treinamento e teste, respectivamente.

(5) Podemos gerar os gráficos das amostras dos conjuntos de treinamento e teste (Figura 1S). Essa etapa é importante para

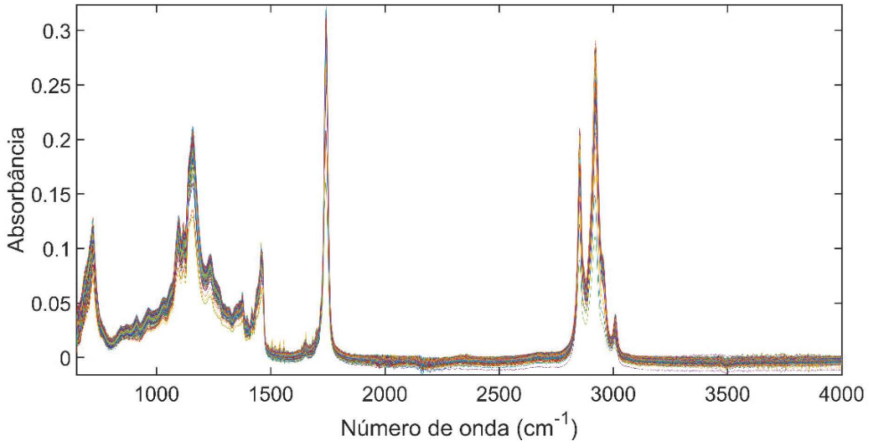

Figura 4. *Espectros de absorbância no infravermelho médio de todas as amostras de óleos vegetais*

identificar erros espectrais grosseiros nas amostras de treinamento e teste:

```
>>figure
>>subplot (2,1,1);
>>plot (numc,xcal)
>>xlabel ('Número de onda (cm^-^1)')
>>ylabel ('Absorbância')
>>axis tight
>>title ('Amostras de Treinamento','fontsize',14);
>>subplot (2,1,2);
>>plot (numc,xval)
>>axis tight
>>xlabel ('Número de onda (cm^-^1)')
>>ylabel ('Absorbância')
>>title ('Amostras de Valibração','fontsize',14);
```

(6) Como utilizaremos o PLS2-DA, é necessário transformar o vetor de classes **y**(1;1;2;2;3...) em uma matriz **Y**(1 0 0; 1 0 0; 0 1 0; 0 1 0;0 0 1...). Esse processo será realizado utilizando a função "vet_matrix":

```
>>ycal = vet_matrix (ycal);
>>yval = vet_matrix (yval);
```

(7) Aplicar o pré-processamento primeira derivada com suavização nas amostras do conjunto de treinamento e teste. Esse pré-processamento é aplicado para destacar as variações das bandas espectrais, corrigir as variações de linha de base e diminuir ruídos espectrais:

```
>>[xcal,xval] = pretrat (xcal,xval,{'deriv';[9,2,1]});
```

(8) Após aplicação do pré-processamento, devemos visualizar os espectros pré-processados para conferir se o procedimento foi feito de maneira adequada e se existem amostras com erros espectrais grosseiros (Figura 2S):

```
>>figure
>>subplot (2,1,1);
>>plot (numc,xcal)
>>title ('Amostras de treinamento')
>>xlabel ('Número de onda (cm^-^1)')
>>ylabel ('Absorbância')
>>subplot (2,1,2);
>>plot (numc,xval)
>>title ('Amostras de Teste')
>>xlabel ('Número de onda (cm^-^1)')
>>ylabel ('Absorbância')
```

Conforme já reportado anteriormente, o número ideal de variáveis latentes será escolhido através da análise da porcentagem de amostras classificadas corretamente nas amostras de validação cruzada. Neste tutorial, a validação cruzada será realizada através da seleção aleatória de 30% das amostras de cada classe das amostras de treinamento. Esse processo será repetido 10 vezes e a porcentagem de acerto médio das amostras de validação cruzada destas 10 repetições será utilizada para definir o número ótimo de variáveis latentes.

(9) O processo de validação cruzada será realizado utilizando a função "my_cross_validation" e o resultado obtido deverá ser um gráfico semelhante à Figura 5. Mais informações a respeito da função "my_cross_validation" podem ser obtidas digitando "help my_cross_validation".

```
>> cvvc = my_cross_validation (xcal, ycal, 10, 10, 4, 0.7);
>> figure;
>>plot (cvvc.porc_am_class_cor')
>>legend ('Classe 1', 'Classe 2', 'Classe 3', 'Classe 4')
>>xlabel ('Número de variáveis latentes');
>>ylabel ('Porcentagem de amostras classificadas corretamente');
```

Através da análise da porcentagem de amostras classificadas corretamente, serão escolhidas 4 variáveis latentes, pois este número corresponde à classificação correta de praticamente 98% das amostras de todas as classes. Definido o número de variáveis latentes, o próximo passo é construir o modelo PLS-DA. Construído o modelo de classificação, a próxima etapa a ser realizada é avaliar a presença de amostras anômalas (*outliers*) no conjunto de treinamento.

(10) A avaliação de amostras anômalas será realizada utilizando a função "my_calc_qt_limits_cal", o resultado obtido deverá ser um gráfico semelhante à Figura 6.

```
>> [model] = my_calc_qt_limits_cal (xcal,ycal,xval,4);
>> am_anomalasQ = find (model.Qres >= model.qlim);
>> am_anomalasT2 = find (model.Thot >= model.tlim);
>> a = ismember (am_anomalasQ, am_anomalasT2);
>> am_anomalas = am_anomalasQ(1,a);
```

É observado que existem 5 amostras que apresentam simultaneamente valores de Q e T^2 acima dos limites calculados. Essas amostras são consideradas anômalas e devem ser removidas do modelo de treinamento. Dentre os prováveis motivos dessas 5 amostras serem consideradas anômalas pelo modelo PLS-DA podemos destacar a limpeza incorreta do ATR durante a etapa de aquisição dos espectros MIR, gerando assim pequenas discrepâncias entre os espectros. Vale ressaltar que, idealmente, os espectros dessas amostras deveriam ser registrados novamente. Outro provável motivo é que estas amostras apresentam elevada discrepância entre as amostras da sua própria classe. Após a remoção das amostras anômalas será realizada novamente a validação cruzada para determinar o número ideal de variáveis latentes.

(11) Eliminando as amostras anômalas no conjunto de treinamento e selecionando novamente o número ideal de variáveis latentes (Figura 7).

```
>> xcal (am_anomalas,:) = [];
>> ycal (am_anomalas,:) = [];
>> cvvc = my_cross_validation (xcal,ycal,10,10,4,0.8);
>> figure;
>>plot (cvvc.porc_am_class_cor')
>>legend ('Classe 1', 'Classe 2', 'Classe 3', 'Classe 4')
>>xlabel ('Número de variáveis latentes');
>>ylabel ('Porcentagem de amostras classificadas corretamente');
```

"Serão selecionadas novamente 4 variáveis latentes, pois 100% das amostras de todas as classes são classificadas corretamente com este número de variáveis. Observe que o número de variáveis latentes não foi alterado após a remoção das amostras anômalas, ressaltando que este não é um comportamento geral, e sempre que forem removidas amostras anômalas na etapa de calibração deve-se escolher novamente o número de variáveis latentes. Definido o número de variáveis latentes, o próximo passo a ser realizado é construir o modelo PLS-DA e avaliar novamente a presença de amostras anômalas (outliers) no conjunto de calibração/treinamento."

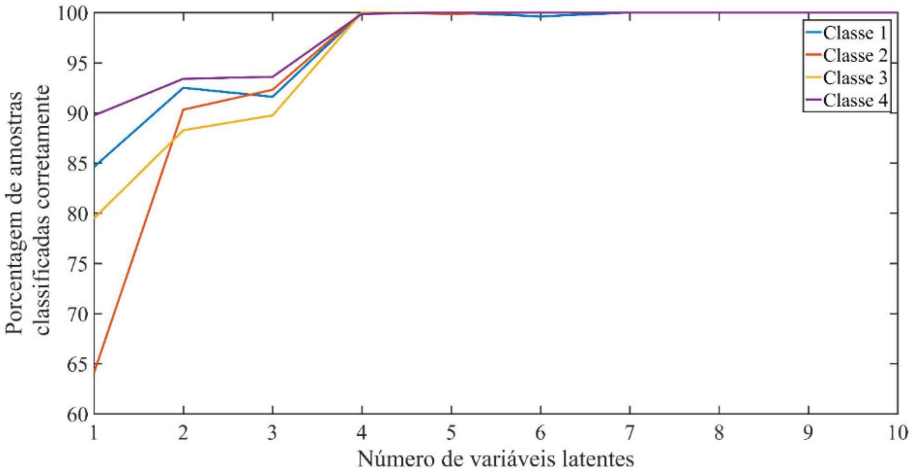

Figura 5. *Porcentagem de amostras classificadas corretamente em função do número de variáveis latentes*

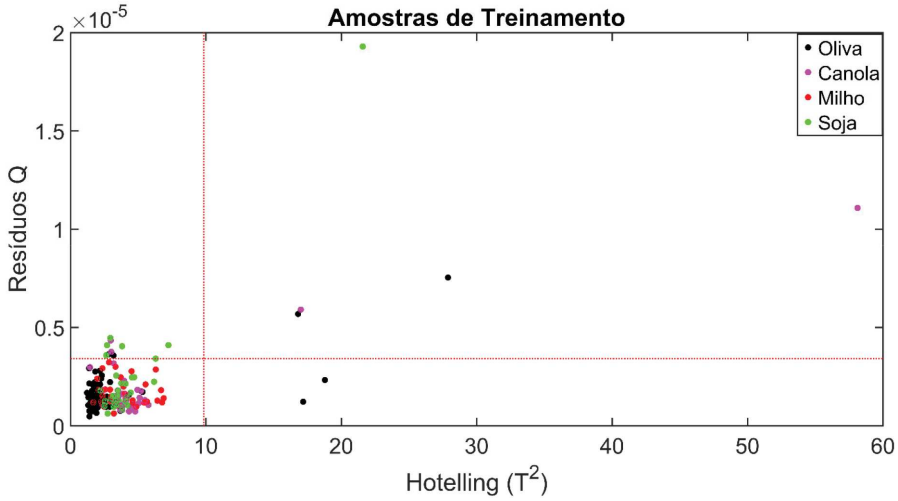

Figura 6. *Resíduos Q versus Hotelling T^2 das amostras de treinamento. A linha tracejada em vermelho representa os limites de Q e T^2 calculados com nível de significância de 5%*

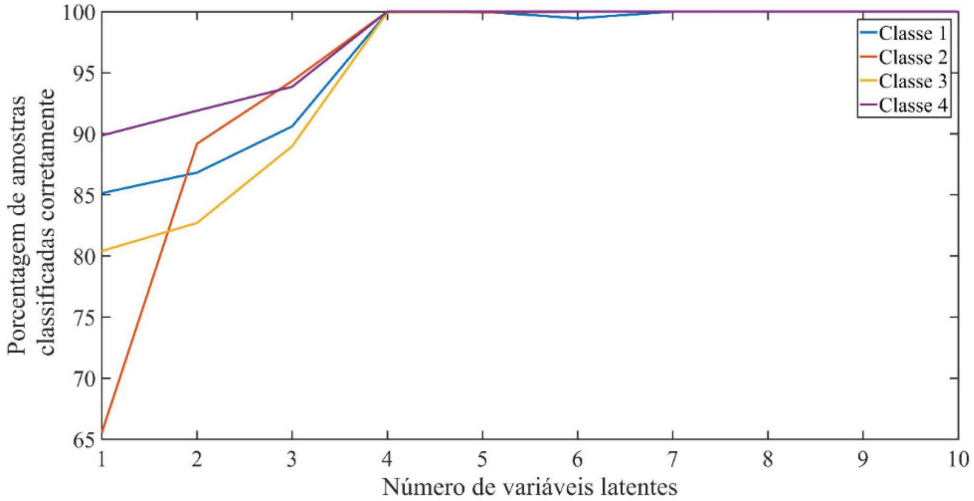

Figura 7. *Porcentagem de amostras classificadas corretamente em função do número de variáveis latentes após a primeira exclusão de amostras anômalas*

(12) Analisando novamente a presença de amostras anômalas para o modelo PLS-DA no conjunto de treinamento usando a função "my_calc_qt_limits_cal" (Figura 8)

```
>> [model] = my_calc_qt_limits_cal (xcal,ycal,xval,4);
```

Como não foram observadas amostras anômalas no conjunto de treinamento será avaliado a presença de amostras anômalas no conjunto de teste usando a função "my_calc_qt_limits_val". O resultado obtido deverá ser um gráfico semelhante à Figura 9.

```
>> [model] = my_calc_qt_limits_val (xcal,ycal,xval,4);
```

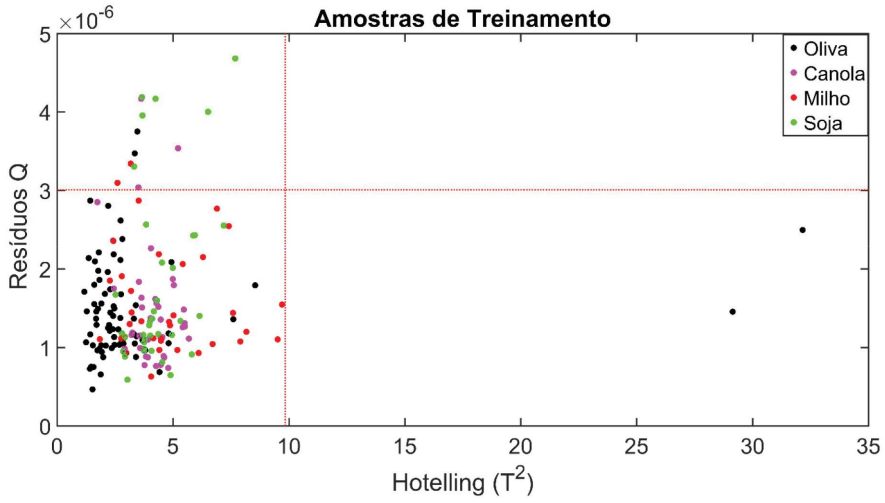

Figura 8. *Resíduos Q versus Hotelling T^2 das amostras de treinamento após a primeira exclusão das amostras anômalas. A linha tracejada em vermelho representa os limites de Q e T^2 calculados com nível de significância de 5%*

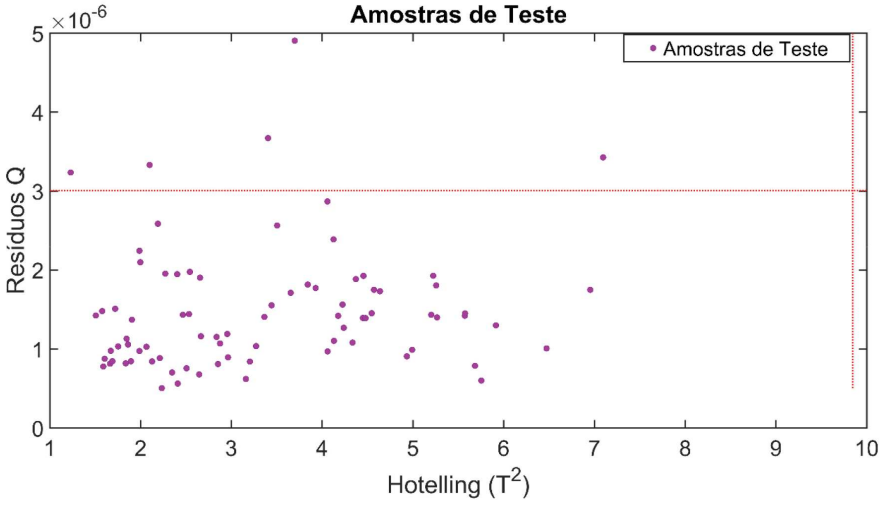

Figura 9. *Resíduos Q versus Hotelling T^2 das amostras de teste. A linha tracejada em vermelho representa os limites de Q e T^2 calculados na etapa de treinamento com nível de significância de 5%*

(13) Como não foram observadas amostras anômalas nos conjuntos de treinamento e teste, será construído o modelo final de classificação PLS-DA com 4 variáveis latentes. Em seguida serão previstas as amostras de treinamento:

```
>>[yprev_cal] = previsto_pls (xcal,ycal,xcal,0,4);
>>% Calculando os limites entre as classes
>>ts = [];
>>for ki = 1:size (ycal,2)
>>plsda_thres = plsdafindthr (yprev_cal(:,ki),ycal(:,ki));
>>ts = [ts,plsda_thres.class_thr];
>>end
>>% Prevendo as amostras de treinamento
>>for u = 1:size (ycal,2)
>>yprev_calts (:,u) = yprev_cal(:,u) >= ts(:,u);
>>end
```

(14) Prevendo as amostras de teste empregando o modelo PLS-DA:

```
>>[yprev_val] = previsto_pls (xcal,ycal,xval,0,4);
>>for u = 1:size(yval,2)
>>yprev_valts(:,u) = yprev_val(:,u)>=ts(:,u);
>>end
```

(15) Para facilitar a interpretação dos resultados será construído o gráfico das classes previstas pelo modelo PLS-DA para as amostras do conjunto de treinamento (Figura 10a):

```
>>Nome{1,1} = 'Oliva' ; Nome{1,2} = 'k';
>>Nome{2,1} = 'Canola' ; Nome{2,2} = 'm';
>>Nome{3,1} = 'Milho' ; Nome{3,2} = 'r';
>>Nome{4,1} = 'Soja' ; Nome{4,2} = 'g';
>>figure
>>for ki = 1:size (yprev_cal,2)
>>subplot (2,2,ki)
>>tp1 = find (ycal(:,ki)); tp2 = setxor(1:length(ycal),tp1);
>>plot (1:length(yprev_cal),yprev_cal(:,ki),'o'),
>>hold on;
>>marc= strcat (Nome{ki,2},'o');
>>plot (tp1,yprev_cal(tp1,ki),marc), hold on
>>hline (ts(ki),'r')
>>title (Nome{ki,1})
>>xlabel ('Amostra')
>>ylabel (sprintf('Classe %g',ki))
>>end
```

(16) Construindo o gráfico das classes previstas pelo modelo PLS-DA

para as amostras do conjunto de validação/teste (Figura 10b):

```
>>for ji = 1:size(yprev_val,2)
>>subplot (2,2,ji)
>>tv1 = find (yval(:,ji)); tv2 = setxor (1:length(yval),tv1);
>>plot (1:length(yprev_val),yprev_val(:,ji),'o'), hold on
>>marc = strcat(Nome{ji,2},'o');
>>plot (tv1,yprev_val(tv1,ji),marc), hold on
>>hline (ts(ji),'r')
>>title (Nome{ji,1})
>>xlabel ('Amostra')
>>ylabel (sprintf('Classe %g',ji))
>>end
```

Para ilustrar uma situação que pode vir a ocorrer em modelos de classificação, suponha que uma amostra *B* seja prevista como pertencente às classes 2 e 3 (óleo de canola e milho, respectivamente). Para esses casos, além de utilizarmos os limiares para predizer a qual classe uma amostra pertence, também é analisada qual a maior probabilidade dessa amostra pertencer às classes previstas. A Tabela 2 apresenta os valores hipotéticos de predição da amostra *B* do conjunto de treinamento para cada uma das classes, sendo que nela também estão incluídos os valores hipotéticos dos limiares calculados para cada modelo.

Nessa situação, a amostra *B* é prevista inicialmente como pertencente simultaneamente às classes óleo de canola e milho. Entretanto, observa-se que o valor previsto para a classe óleo de canola está muito próximo do seu limiar, enquanto para a classe óleo de milho esse valor está mais distante do respectivo limiar. Calculando-se as diferenças entre os valores previstos e os respectivos limiares, observa-se que a diferença é maior para a classe dos óleos de milho. Além disso, o valor previsto nesse caso é mais próximo de 1, indicando que a probabilidade dessa amostra ser um óleo de milho é maior. Logo, essa amostra é prevista como óleo de milho. Por outro lado, existe a possibilidade da presença de misturas de óleos no modelo, desse modo, a amostra *B* poderia ser uma blenda dos óleos de canola e milho. A escolha dentre essas duas possibilidades é realizada de acordo com o problema avaliado. Esse é apenas um exemplo de como os modelos PLS-DA podem ser usados com flexibilidade em função do conhecimento que o analista tem do problema estudado. Dessa maneira, diferentes restrições podem ser impostas ao modelo de classificação.

Construído o modelo de classificação, o próximo passo é avaliar sua eficiência. Nesse tutorial serão utilizadas a tabela de contingência e as figuras de mérito descritas anteriormente.

(17) Calculando a tabela de contingência e as figuras de mérito para as amostras dos conjuntos de treinamento e teste (Tabela 3 e Tabela 4).

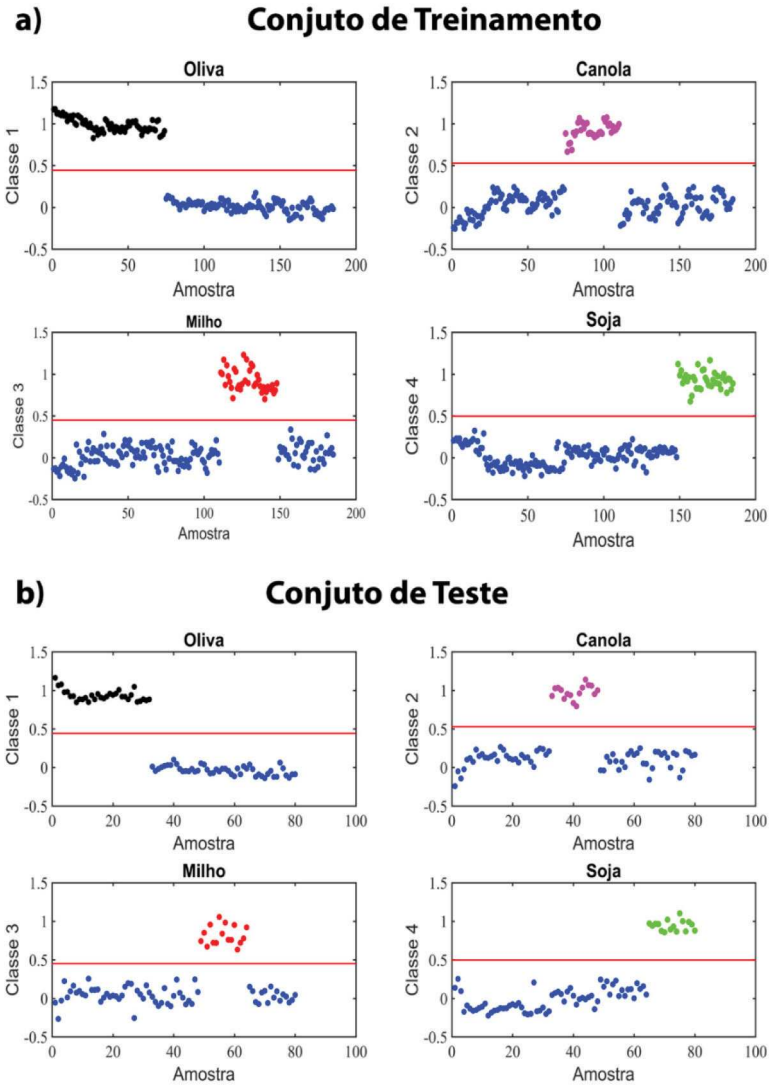

Figura 10. *Resultados de previsão do modelo de classificação PLS-DA para as amostras do conjunto de treinamento (a) e teste (b). A linha vermelha representa o limiar (threshold) entre as classes*

Tabela 2. Valores hipotéticos de previsão da amostra *B* do conjunto de treinamento e valores hipotéticos dos limiares calculados para cada modelo

Amostra *B*	Azeite de Oliva	Óleo de Canola	Óleo de Milho	Óleo de Soja
Valores previstos	0,008	0,607	1,067	-0,682
Limiar	0,502	0,491	0,491	0,440
Diferença do valor previsto até o limiar	Não aplica	0,116	0,576	Não aplica

Tabela 3. Tabela de contingência obtida pelo modelo de classificação PLS-DA para as amostras dos conjuntos de treinamento e teste

Conjunto de Treinamento				
Classe original	Previsto			
	Azeite de Oliva	Óleo de Canola	Óleo de Milho	Óleo de Soja
Azeite de Oliva	74	0	0	0
Óleo de Canola	0	36	0	0
Óleo de Milho	0	0	38	0
Óleo de Soja	0	0	0	37
Conjunto de Teste				
Classe original	Previsto			
	Azeite de Oliva	Óleo de Canola	Óleo de Milho	Óleo de Soja
Azeite de Oliva	32	0	0	0
Óleo de Canola	0	16	0	0
Óleo de Milho	0	0	16	0
Óleo de Soja	0	0	0	16

Tabela 4. Figuras de mérito obtidas pelo modelo de classificação PLS-DA para as amostras dos conjuntos de treinamento e teste

Conjunto de Treinamento				
	Azeite de Oliva	Óleo de Canola	Óleo de Milho	Óleo de Soja
Acurácia %	100	100	100	100
Sensibilidade %	100	100	100	100
Especificidade %	100	100	100	100
Conjunto de Teste				
	Azeite de Oliva	Óleo de Canola	Óleo de Milho	Óleo de Soja
Acurácia %	100	100	100	100
Sensibilidade %	100	100	100	100
Especificidade %	100	100	100	100

```
>>[Tcal, Tcal2] = my_ConfTable_cal (ycal,yprev_cal,ts)
>>[Tval, Tval2] = my_ConfTable_val (yval,yprev_val,ts)
```

Esses resultados são gerados automaticamente em um arquivo do *Microsoft Excel* chamado "Resultados_PLSDA.xlsx" através das funções "my_ConfTable_cal" e "my_ConfTable_val". Caso sejam utilizados outros conjuntos de dados, é necessário abrir as funções "my_ConfTable_cal" e "my_ConfTable_val" para adequar o arquivo do *Microsoft Excel* que será gerado.

Os valores das figuras de mérito obtidas pelo modelo de classificação PLS-DA foram excelentes para os conjuntos de treinamento e teste, com valores de sensibilidade e especificidade iguais a 100% em todas as classes, ou seja, todas as amostras de óleos foram classificadas corretamente de acordo com a sua oleaginosa. A acurácia/eficiência obtida pelo modelo de classificação PLS-DA nos conjuntos de treinamento e teste foi de 100%, ou seja, o modelo foi capaz de prever corretamente 100% das amostras deste conjunto.

No exemplo utilizado nesse tutorial, o modelo PLS-DA foi capaz de classificar corretamente todas as amostras em todas as classes, no entanto, nem sempre isso acontece e nesses casos a avaliação das figuras de mérito tornam-se muito relevantes para inferir sobre a aplicabilidade do modelo desenvolvido.

CONCLUSÃO

Dando continuidade à série de tutoriais de Quimiometria em Matlab, o reconhecimento supervisionado de padrões foi explorado através do emprego da PLS-DA para a classificação de óleos vegetais comestíveis usando espectroscopia na região do infravermelho médio. O mesmo experimento empregado no primeiro tutorial dessa série foi revisitado e explorado para classificar as amostras de óleos vegetais, levando-se em consideração o conhecimento prévio das classes formadas por amostras de cada tipo, conforme preconizado pelo PLS-DA. Foram apresentados todos os comandos e rotinas em Matlab necessários para realizar essa análise multivariada e entender seu funcionamento de forma prática e detalhada. Os dados empregados nesse estudo estão disponíveis por meio de livre acesso na internet[22] ou enviando e-mail para os autores desse trabalho. Uma

vez compreendidos os conceitos teóricos e práticos dos métodos apresentados, encoraja-se o leitor a praticar e aplicar as rotinas disponibilizadas em outras matrizes de dados e aplicações diversas.

MATERIAL SUPLEMENTAR

O algoritmo NIPALS para o PLS1 e as Figuras 1S e 2S estão no material suplementar disponível em http://quimicanova.sbq.org.br, na forma de arquivo PDF, com acesso livre. Na Figura 1S são apresentados os espectros das amostras selecionadas para os conjuntos de treinamento e teste, e na Figura 2S são mostrados os espectros derivados e suavizados das amostras selecionadas para os conjuntos de treinamento e teste.

AGRADECIMENTOS

Os autores agradecem ao Instituto Nacional de Ciência e Tecnologia de Bioanalítica (INCTBio), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasil, 465389/2014-7 e 303994/2017-7), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brasil, Código de financiamento 001), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Brasil, 2014/508673), empresa Speclab Holding S.A e a Embrapa (Projeto número MP5 14.05.01.001.01.00. 00) pelo suporte financeiro.

REFERÊNCIAS

1. Jiménez-Carvelo, A. M.; González-Casado, A.; Bagur-González, M. G.; Cuadros-Rodríguez, L.; *Food Res. Int.* **2019**, *122*, 25.
2. De Souza, A. M.; Poppi, R. J.; *Quim. Nova* **2012**, *35*, 223.
3. De Souza, A. M.; Breitkreitz, M. C.; Filgueiras, P. R.; Rohwedder, J. J. R.; Poppi, R. J.; *Quim. Nova* **2013**, *36*, 1057.
4. Da-Col, J. A.; Dantas, W.; Poppi, R.; *Quim. Nova* **2017**, *41*, 345.
5. Breitkreitz, M. C.; Souza, A. M. de; Poppi, R. J.; *Quim. Nova* **2014**, *37*, 564.
6. Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y.; *Anal. Chim. Acta* **2016**, *914*, 17.
7. Barker, M.; Rayens, W.; *J. Chemom.* **2003**, *17*, 166.
8. Bombarda, I.; Dupuy, N.; Le Van Da, J.-P.; Gaydou, E. M.; *Anal. Chim. Acta* **2008**, *613*, 31.
9. Dominguez-Vidal, A.; Pantoja-De La Rosa, J.; Cuadros-Rodríguez, L.; Ayora-Cañada, M. J.; *Food Chem.* **2016**, *190*, 122.
10. Visani, V.; Netto, J. M. S.; Honorato, R. S.; de Araújo, M. C. U.; Honorato, F. A.; *Microchem. J.* **2017**, *133*, 480.
11. De Santana, F. B.; Gontijo, L. C.; Mitsutake, H.; Mazivila, S. J.; De Souza, L. M.; Borges Neto, W.; *Food Chem.* **2016**, *209*, 228.
12. de Santana, F. B.; Neto, W. B.; Poppi, R. J.; *Food Chem.* **2019**, *293*, 323.
13. Skoog, D. A.; Crouch, S. R.; Holler, F. J.; *Principles of Instrumental Analysis*, 7ª ed., Cengage Learning: Stamford; 2018.
14. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**, *185*, 1.
15. Ballabio, D.; Consonni, V.; *Anal. Methods* **2013**, *5*, 3790.
16. Ferreira, M. M. C.; *Quimiometria: Conceitos, Métodos e Aplicações*, 1ª ed., Editora da Unicamp: Campinas, 2015.
17. Haaland, D. M.; Thomas, E. V.; *Anal. Chem.* **1988**, *60*, 1193.
18. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; *An Introduction to Statistical Learning*, Springer, 2013.
19. Hibbert, D. B.; Armstrong, N.; *Chemom. Intell. Lab. Syst.* **2009**, *97*, 211.
20. Brereton, R. G.; Lloyd, G. R.; *J. Chemom.* **2014**, *28*, 213.
21. Savitzky, A.; Golay, M. J. E.; *Anal. Chem.* **1964**, *36*, 1627.
22. Rinnan, Å.; *Anal. Methods* **2014**, *6*, 7124.
23. Pasquini, C.; *Anal. Chim. Acta* **2018**, *1026*, 8.
24. Laursen, K.; Frederiksen, S. S.; Leuenhagen, C.; Bro, R.; *J. Chromatogr. A* **2010**, *1217*, 6503.
25. Jackson, J. E.; Mudholkar, G. S.; *Technometrics* **1979**, *21*, 341.
26. Bro, R.; Smilde, A. K.; *Anal. Methods* **2014**, *6*, 2812.
27. Isabel López, M.; Pilar Callao, M.; Ruisánchez, I.; *Anal. Chim. Acta* **2015**, *891*, 62.
28. GitHub Tutorial-PLS-DA-Quimiometria: https://github.com/felipebachion/Tutorial-PLS-DA-Quimiometria, acessado em Janeiro 2019.
29. Kennard, R. W.; Stone, L. A.; *Technometrics* **1969**, *11*, 137.