UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

DEPARTAMENTO DE QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

ANÁLISE DISCRIMINANTE LINEAR EM DUAS DIMENSÕES PARA CLASSIFICAÇÃO DE DADOS QUÍMICOS DE SEGUNDA ORDEM

Adenilton Camilo da Silva

João Pessoa-PB,

Agosto de 2017

*Adenilton Camilo da Silva**

ANÁLISE DISCRIMINANTE LINEAR EM DUAS DIMENSÕES PARA CLASSIFICAÇÃO DE DADOS QUÍMICOS DE SEGUNDA ORDEM

Exame de doutorado submetido ao Programa de Pós-Graduação em Química da Universidade Federal da Paraíba como parte dos requisitos para obtenção do título de Doutor em Química, área de concentração Química Analítica.

Orientador: Prof. Dr. Mário Cesar Ugulino de Araújo

**Bolsista CNPq*

João Pessoa-PB,

Agosto de 2017

Catalogação na publicação
Setor de Catalogação e Classificação

S586a Silva, Adenilton Camilo da.
Análise discriminante linear em duas dimensões para classificação de dados químicos de segunda ordem / Adenilton Camilo da Silva. - João Pessoa, 2017.
118 f. : il. -

Orientador(a): Prof. Dr. Mário Cesar Ugulino de Aráujo.
Tese (Doutorado) – UFPB/CCEN/PPGQ

1. Química analítica. 2. Técnicas quimiométricas. 3. Técnicas de redução de dimensionalidade (TUCKER, PARAFAC). 4. Análise discriminante linear (2D – LDA). 5. Técnica de classificação multivariada (U-PLS-DA). I. Título.

UFPB/BC CDU - 543(043)

Análise discriminante linear em duas dimensões para classificação de dados químicos de segunda ordem.

Tese de Doutorado apresentada pelo aluno Adenilton Camilo da Silva e aprovada pela banca examinadora em 21 de agosto de 2017.

Prof. Dr. Mário César Ugulino de Araújo
Orientador/Presidente

Prof. Dr. Kássio Michell Gomes de Lima
Examinador

Prof. Dr. José Germano Veras Neto
Examinador

Prof. Dr. Edvan Cirino da Silva
Examinador

Prof. Dr. Wallace Duarte Fragoso
Examinador

“Pois o sol de um novo dia vai brilhar
E essa luz vai refletir na nossa estrada
Clareando de uma vez a caminhada
que nos levará direto ao apogeu
Tenha fé, nunca perca a fé em Deus.”

(Diogo Nogueira)

A ***Deus*** pelo os dons,

a minha querida mãe ***Maria do Socorro***,

e a minha ***família*** por todo o amor que nos fortalece,

Dedico...

Agradecimentos

A minha esposa e meu filho, ***Ayla Thamires e João Pedro***, que são presença firme de Deus na minha vida, e que estão sempre ao meu lado, torcendo pela nossa família e por nossas vitórias.

- Aos meus irmãos, ***Adailton e Aline***, por suportarem minhas ausências, mas torcerem com tanto entusiasmo por minha felicidadee crescimento pessoal e profissional.
- Aos meus queridos amigos, ***José Wanderley*** e ***Glaucia Maria*** e sua família, por terem contribuído com exemplo de vida de ***Diva Glaucy*** (in memoriam), e por me amarem como filho.
- Aos meus avós, ***Anésio*** (*in memoriam*) e ***Francisca*** (*in memoriam*), que cuidaram com tanto amor e carinho de seus filhos e netos, mostrando-os caminhos da retidão, amor e honestidade.
- Ao professor ***Mario Cesar Ugulino de Araújo***, por todo apoio e orientação. Sou extremamente grato pela confiança, pelas oportunidades, por todas as contribuições valiosas para minha vida.
- Ao amigo ***Valber Elias***, pelas conversas, por toda alegria em dividir sua moradia comigo por tanto tempo.
- Aos amigos ***Sófacles Figueredo e Matias Insausti*** por todas as contribuições e principalmente por acreditarem e me apoiarem no decorrer de toda a pesquisa.
- Ao professor ***Roberto Kawakami*** por contribuir de forma expressiva com a tese ao me receber com tamanha atenção no ITA.
- Aos ***amigos do LAQA***, pelo o apoio nos momentos difíceis, pelas alegrias, conversas e contribuições valiosas.
- Ao ***CNPq*** pelo financiamento da bolsa de pesquisa.

Agradeço de coração a todos que contribuíram e contribuem com a minha caminhada. Este trabalho é nosso! Muito obrigado!

SUMÁRIO

LISTA DE FIGURAS ..... **iv**
LISTA DE TABELAS ..... **vii**
SIGLAS E ABREVIATURAS ..... **ix**
RESUMO ..... **xi**
ABSTRACT ..... **xii**
1. INTRODUÇÃO **1**
1.1 Caracterização geral da problemática e proposta ***1***
1.2 Objetivos ***4***
1.2.1 Objetivos Gerais *4*
1.2.2 Objetivos Específicos *4*
2.FUNDAMENTAÇÃO TEÓRICA **6**
2.1 Classificações dos dados analíticos e técnicas quimiométricas ***6***
2.1.1 Bilinearidade/trilinearidade dos dados *8*
2.2 Técnicas de redução de dimensionalidade para dados de segunda ordem **11**
2.2.1 Desdobramento ("unfolding") *11*
2.2.2 Métodos de Tucker *13*
2.2.3 PARAFAC *15*
2.3- Técnicas de classificação multivariada **17**
2.3.1 Análise Discriminante Linear - LDA *17*
2.3.2 Mínimos Quadrados Parciais para Analise Discriminante - PLS-DA *20*
2.3.3 N-PLS-DA *23*
3 ALGORITMO PROPOSTO **26**
3.1 Analise discriminante linear em duas dimensões (2D-LDA) **26**
3.1.1 Notação *26*
3.1.2 Determinação dos vetores de projeção *27*
3.1.3 Procedimento de classificação *31*
4. EXPERIMENTAL **34**
4.1 Conjunto de dados **34**
4.1.1 Conjunto de dados simulados *34*
4.1.1.1 Dados simulados I *34*

4.1.1.2 Dados simulados II *37*
4.1.2 Dados de Presunto de Parma curado a seco *38*
4.1.3 Dados de óleo vegetal comestível *40*
4.2 Programas **43**
4.2.1 2D-LDA *43*
4.2.2 No feature extraction ("NFE") *43*
4.2.3 U-PLS-DA *44*
4.2.4 PARAFAC-LDA e TUCKER-3-LDA *45*
5. RESULTADOS E DISCUSSÃO **47**
5.1 Conjuntos de dados simulados de EEM **47**
5.1.1 2D-LDA *47*
5.1.2 PARAFAC-LDA *49*
5.1.3 TUCKER-3-LDA *53*
5.1.4 U-PLS-LDA *55*
5.1.5 NFE (No feature extraction) *57*
5.2 Conjuntos de dados de presunto de Parma curado a seco **60**
5.2.1 2D-LDA *60*
5.2.2 PARAFAC-LDA *63*
5.2.3 TUCKER-3 ..... *65*
5.2.4 U-PLS-DA *66*
5.2.5 NFE (No feature extraction) *68*
5.3 Conjuntos de dados de óleo vegetal comestível **70**
5.3.1 2D-LDA *70*
5.3.2 PARAFAC-LDA *72*
5.3.3 TUCKER-3 ..... *73*
5.3.4 U-PLS-DA *75*
5.3.5 NFE (No feature extraction) *76*
6. CONCLUSÃO **81**
6.1 Propostas futuras **82**
REFERÊNCIAS **83**
Apêndice 1 ..... **92**

Apêndice 2 ..... **96**
Apêndice 3 ..... **97**
Anexo 1 ..... **98**
Anexo 2 ..... **99**

LISTA DE FIGURAS

Figura 1 – Representação das diferentes matrizes de dados para uma amostra ou conjunto de amostras e nomenclatura empregada. (Figura adaptada da referência: [4]). 8

Figura 2 - Gráfico de contorno de uma matriz de dados (EEM ou LC-DAD com níveis de intensidade crescentes de azul para vermelho). (Figura da referência: [6]). 9

Figura 3 - Representação do cubo de dados $\underline{\mathbf{X}}(I,J,K)$ gerado a partir do empilhamento matrizes de dados das amostras. 10

Figura 4 - Maneiras de realizar o rearranjo de um cubo de dados usando desdobramento. 12

Figura 5 - Desdobramento de uma matriz de dados simulados de EEM na direção I x JK. 12

Figura 6- Representação gráfica para o modelo Tucker-3 (Figura adaptada da referência: [38]). 14

Figura 7 - Representação gráfica para o modelo PARAFAC. Decomposição em F tríades de vetores de peso. (Figura da referência: [38]). 16

Figura 8 - Representação do principio da LDA. Sw (Espalhamento intraclasse) e Sb (Espalhamento entre classes). 18

Figura 9 - Estrutura de apresentação das matrizes de dados instrumentais e do vetor de índice de classe. Onde: N = número de amostras, J = número de variáveis, e M = número de descritores. 21

Figura 10 - Decomposição em dois componentes para um cubo de dados em N-PLS-DA. (Figura da referência: [59]). 23

Figura 11- Determinação de cada elemento do vetor de características de uma medida de fluorescência em EEM. 28

Figura 22- Média das amostras de treinamento utilizadas na obtenção das matrizes de espalhamento. 29

Figura 33- Obtenção dos vetores de características a partir o vetor de projeção **b**. 31

Figura 14 - Fatores empregados na geração do conjunto de dados EEM simulado I. (A) Perfis de excitação, (b) perfis de emissão, (c) fator A1, (d) fator A2, (e) fator A3, (f) fator A4**.** 35

Figura 15 – Perfis das classes dos dados de EEM simulado I. (a) Classe 1, (b) Classe 2 e (c) Classe 3. 35

Figura 16 - Fatores empregados na geração do conjunto de dados EEM simulado II. (A) Perfis de excitação, (b) perfis de emissão, (c) fator A1, (d) fator A2, (e) fator A3, (f) fator A4**.** 37

Figura 17 – Perfis das classes dos dados de EEM simulado II. (a) Classe 1, (b) Classe 2 e (c) Classe 3. 37

Figura 18 – Espectros de autofluorescência de superfície para amostras de presunto de Parma maturadas a seco referentes a seguintes classes: (**a**) crua, (**b**) salgada, (**c**) maturada, (**d**) envelhecida. 40

Figura 19 – espectrofluorímetro Aminco Bowman Series 2 utilizado na aquisição dos espectros de fluorescência sincrônica das amostras de óleo vegetal comestível..... 42

Figura 20 – Espectros de fluorescência sincrônica amostras de óleo vegetal comestível referentes a seguintes classes: (**a**) soja, (**b**) milho e (**c**) girassol. 42

Figura 21 – Taxa de classificação correta por validação cruzada obtida por validação cruzada versos número de vetores de projeção para os bancos de dados de EEM simulados (**a**) I e (**b**) II. 47

Figura 22 – Vetores de características 2D-LDA para dados EEM simulados I (**a**) e (**b**) II. 48

Figura 23 – Taxa de classificação correta por validação cruzada versos número de vetores de projeção (Dados simulados (**a**) I e (**b**) II), valor de corcondia (Dados simulados (**c**) I e (**d**) II) e de erro de modelo (Dados simulados (**e**) I e (**f**) II) versos número de fatores PARAFAC. 50

Figura 24 – Perfis recuperados com PARAFAC para os dados simulados I. Usando 2 fatores (Emissão (a) e Excitação (b)) e 4 fatores (Emissão (c) e Excitação (d)). 51

Figura 25 – Perfis recuperados com PARAFAC para os dados simulados I. Usando 2 fatores (Emissão (a) e Excitação (b)) e 4 fatores (Emissão (c) e Excitação (d)). 53

Figura 26 –Taxa de classificação correta por validação cruzada versos: número de Fatores Tucker-3 (Dados simulados (**a**) I e (**b**) II) e diferentes combinações de fatores Tucker-3 (Dados simulados (**c**) I e (**d**) II)..... 54

Figura 27 – Taxa de classificação correta por validação cruzada versos número de variáveis latentes para os bancos de dados de EEM simulados (**a**) I e (**b**) II..... 56

Figura 28 – Perfil de desdobramento dos bancos de dados de EEM simulados (**a**) I e (**b**) II. 58

Figura 29 –Taxa de classificação correta por validação cruzada obtida versos número de vetores de projeção para o banco de dados de presunto de Parma curado a seco. 61

Figura 30 – Vetores de características 2D-LDA das amostras de treinamento do banco e dados de presunto de Parma curado a seco. (a) 1° vetor, (b) 2° vetor, (c) 3° vetor, (d) 4° vetor e (e) 5° vetor. 61

Figura 31 – Taxa de classificação correta por validação cruzada versos número de fatores PARAFAC (**a**), valor de corcondia (**b**) e de erro de modelo (**c**). 63

Figura 32 –Taxa de classificação correta por validação cruzada versos: número de Fatores Tucker-3 (**a**), diferentes combinações de fatores Tucker-3 (**b**). 65

Figura 33 –Taxa de classificação correta por validação cruzada versos número de variáveis latentes para os dados de presunto de Parma. 67

Figura 34 – Perfil de desdobramento das amostras das classes de treinamento do banco de dados e presunto de Parma curado a seco. 69

Figura 35 –Taxa de classificação correta por validação cruzada obtida versos número de vetores de projeção para o banco de dados de óleo vegetal comestível..... 70

Figura 36 – Dados de óleo vegetal comestível: Vetores de caraterísticas do 2D-LDA para o grupo de amostras de treinamento obtidos com: (a)primeiro, (b) segundo, (c) terceiro e (d) quarto vetores de projeção. 71

Figura 37–Taxa de classificação correta por validação cruzada versos número de fatores PARAFAC (a), valor de corcondia (b) e de erro de modelo (c). 72

Figura 38 –Taxa de classificação correta por validação cruzada versos: número de Fatores Tucker-3 (**a**), diferentes combinações de fatores Tucker-3 (**b**). 74

Figura 39 –Taxa de classificação correta por validação cruzada versos número de variáveis latentes para os dados de óleo vegetal comestível. 75

Figura 40 – Perfil de desdobramento das amostras das classes de treinamento do banco de dados e óleo vegetal comestível..... 77

LISTA DE TABELAS

Tabela 1–Valores de média e desvio padrão (Std) empregados na geração aleatória dos coeficientes das combinações lineares usadas para a construção dos conjuntos de dados simulados. 36

Tabela 2 – Divisão das amostras do banco de dados de EEM simulado I e II em grupo de treinamento e teste. 38

Tabela 3 – Divisão das amostras do banco de dados de presunto de Parma curado a seco, em dois grupos: treinamento e teste. 40

Tabela 4 – Divisão das amostras do banco de óleos vegetais comestíveis, em dois grupos: treinamento e teste. 43

Tabela 5 -Dados simulados: Resultado de classificação usando NFE para o grupo de Teste.57

Tabela 6 -Dados simulados: Resultado de classificação do grupo de Teste usando diferentes estratégias de modelagem. 59

Tabela 7 -Resultado de classificação usando 2D-LDA para o grupo de Teste dos dados de presunto de Parma curado a seco. 62

Tabela 8- Resultado de classificação usando PARAFAC-LDA com 5 e 6 fatores para o grupo de Teste dos dados de presunto de Parma curado a seco 64

Tabela 9- Resultado de classificação usando TUCKER-3com 5 fatores para o grupo de Teste dos dados de presunto de Parma curado a seco. 66

Tabela 10- Resultado de classificação usando U-PLS-DA com 7 fatores para o grupo de Teste dos dados de presunto de Parma curado a seco. 67

Tabela 11- Resultado de classificação usando NFE para o grupo de Teste dos dados de presunto de Parma curado a seco. 68

Tabela 12 -Dados de presunto de Parma curado a seco: Resultado de classificação do grupo de Teste usando diferentes estratégias de modelagem. 69

Tabela 13- Resultado de classificação usando PARAFAC-LDA com 5 fatores para o grupo de Teste dos dados de óleo vegetal comestível. 73

Tabela 14 - Resultado de classificação usando TUCKER-3-LDA com 4 e 5 fatores para o grupo de Teste dos dados de óleo vegetal comestível. 75

Tabela 15 - Resultado de classificação usando U-PLS-DA com 5 variáveis latentes para o grupo de Teste dos dados de óleo vegetal comestível. 76

Tabela 16 - Resultado de classificação usando NFE para o grupo de Teste dos dados de óleo vegetal comestível. 77

Tabela 17 -Dados de óleo vegetal comestível: Resultado de classificação do grupo de Teste usando diferentes estratégias de modelagem. 78

Tabela 18 -Resultados comparativos: Taxas de classificação corretas obtidas nos conjuntos de teste. O número de vetores de projeção, fatores ou variáveis latentes empregadas em cada modelo é indicado entre parênteses. 79

SIGLAS E ABREVIATURAS

2D-LC (Cromatografia líquida bidimensional, do inglês *Two-dimensional liquid chromatography*).

2D-LDA (Análise discriminante linear em duas dimensões, do inglês *Two dimensional discriminant analysis*).

APQLD (Decomposição quadrilinear alternada, do inglês A*lternating penalty quadrilinear decomposition*).

AWRCQLD (Decomposição quadrilinear alternada com restrição, do inglês *Alternating weighted residue constraint quadrilinear decomposition*).

CG-MS (Cromatografia gasosa acoplada à espectrometria de massa, do inglês *Gas chromatography-mass spectrometry*).

CORCONDIA (Diagnostico de consistência de nuclear, do inglês *Core Consistency Diagnostic*).

EEMs (Espectroscopia de fluorescência em matriz de excitação-emissão, do inglês *Excitation-Emission Matrix spectroscopy*).

EMWTM (Matriz de tempo de onda modulada por excitação, do inglês *Excitation modulated wavelength time matrix*).

HPLC-DAD (Cromatografia líquida de alta eficiência com detecção por arranjo de diodo, do inglês *High-Performance Liquid Chromatography with Diode-Array Detection*).

LC-MS (Cromatografia liquida acoplada à espectrometria de massa, do inglês *Liquid chromatography–mass spectrometry*).

LDA (Análise discriminante linear, do inglês *Linear discriminant analysis*).

MS-MS (Espectrometria de massa acoplada, do inglês *Tandem mass spectrometry*).

NFE – (Sem extração de características, do ingês *No feature extraction*).

NMF(Fatoração de matriz não negativa, do inglês *Non-negative matrix factorization*).

N-PLS (Minimos quadrados parciais multivias, do inglês *N-way Partial least square*).

N-PLS-DA (Mínimos quadrados parciais para analise discriminante multivias, do inglês *N-Way partial least square discrminant analysis*).

N-SIMCA (Modelagem Independente por Analogia de Classe multivias, do inglês *Multi-way Soft independent modelling by class analogy*).

OD (Distância ortogonal, do inglês *Ortogonal Distance*).

PARAFAC(Análise de fatores paralelos, do inglês *Parallel factor analysis*).

PCA (Análise por componentes principais, do inglês *Principal component analysis*).

PCR (Regressão por componentes principais, do inglês *Principal componente regression*).

PLS-DA (Mínimos quadrados parciais para análise discriminante, do inglês *Partial least square discriminant analysis*).

PLSR (Regressão por mínimos quadrados parciais, do inglês *Partial least square regression*).

Sb (Espalhamento entre as classes, do inglês *Scattering between*).

SD (Distancia de escores, do inglês *Score Distance*).

Sw (Espalhamento no interclasse, do ingês *Scattering within*).

TCC (Taxa de classificação correta).

Xresid(Matriz de resíduos de **X**).

yresid (Matriz de resíduos de **y**).

RESUMO

Com os avanços na instrumentação analítica tem sido cada vez mais comum a obtenção de dados de segunda ordem, principalmente pelo uso de técnicas hifenadas. Apesar das vantagens obtidas com o aumento do número informações sobre a amostra, a partir de detectores, a interpretação direta dos dados pode ser um desafio dada a complexidade de algumas matrizes. Diante disso, é importante que novas estratégias quimiométricas sejam propostas a fim de auxiliar na interpretação desse tipo de dado, como é o caso do algoritmo de análise linear discriminante em duas dimensões (2D-LDA). O 2D-LDA foi originalmente proposto no contexto do processamento de imagens de face para a extração de vetores características com alto poder discriminante. Apesar do seu desempenho promissor em tratamento de imagens, o algoritmo 2D-LDA ainda não foi utilizado em aplicações que envolvem dados químicos. Neste trabalho foi investigado o uso de 2D-LDA em problemas de classificação envolvendo dados químicos de segunda ordem. Quatro conjuntos de dados foram utilizados:dois conjuntos de dados simulados de espectrometria de fluorescência de matriz de excitação/emissão; um conjunto de espectros de autofluorescência de presunto de Parma e outro conjunto de espectro de fluorescência sincrônica total de óleo vegetais comestíveis. Os resultados foram comparados com aqueles obtidos utilizando: Classificação sem extração de características (NFE); U-PLS-DA (Análise discriminante por mínimos quadrados parciais em dados desdobrados) e LDA usando escores de TUCKER-3 ou PARAFAC. No primeiro conjunto de dados simulados, todos os modelos alcançaram uma taxa de classificação correta de 100%. Contudo, no segundo conjunto de dados simulado apenas o modelo NFE apresentou erros de classificação (30%). Os conjuntos de dados do presunto de Parma e dos óleos vegetais obtiveram maior taxa de classificação utilizando 2D-LDA (86%) e TUCKER-3-LDA (100%), em comparação com os modelos sem extração de características (76% e 77%), U-PLS -DA (81% e 92%) e PARAFAC-LDA (86% e 92%). Em geral, o 2D-LDA apresentou resultados comparáveis aos demais algoritmos avaliados, podendo ser considerado como uma estratégia promissora na classificação de dados químicos de segunda ordem.

Palavras chave: 2D-LDA, segunda ordem; classificação; PARAFAC; TUCKER, U-PLS-DA

ABSTRACT

With advances in analytical instrumentation has been increasingly common to obtain second order data by using primarily hyphenated techniques. Despite the advantages obtained by increasing the number of detectors used in sample measurement, the direct data interpretation can be a challenge given the complexity of some matrices. Thus it is important that new chemometric strategies are proposed to support in the interpretation of the data type, such as the two-dimensional discriminant linear analysis algorithm (2D-LDA). 2D-LDA was originally proposed in the image processing context for extraction of characteristic vectors with high discriminant power. Despite its promising performance in image processing, the 2D-LDA algorithm has not used in applications involving chemical data. This work investigates the use of 2D-LDA in classification problems involving second order chemical data. Four datasets were used: 2 simulated datasets of excitation / emission matrix fluorescence spectrometry; Auto fluorescence Spectrometry of Parma Ham, Total Synchronous Spectrometry of Edible Vegetable Oil. The results were compared with following algorithms: no feature extraction (NFE); U-PLS-DA (Partial least squares discriminant analysis in unfolded data) and LDA by using TUCKER-3 or PARAFAC scores. In the first simulated data set all models achieved a correct classification rate of 100%. However, in the second simulated data set only NFE model presented classification errors (30%). The Parma ham and vegetable oils data sets obtained the best classification rates by using 2D-LDA and TUCKER-3-LDA (86% and 100%) compared to the models without extraction of characteristics (76% and 77%), U-PLS-DA (81% and 92%) and PARAFAC-LDA (86% and 92%). In general, the 2D-LDA presented comparable results to the other algorithms and could be considered as a promising strategy in the classification of second order chemical data.

Keywords: 2D-LDA, second order, classification, PARAFAC, TUCKER, U-PLS-DA

CAPÍTULO 1

INTRODUÇÃO

1. INTRODUÇÃO

1.1 Caracterização geral da problemática e proposta

Com os avanços na instrumentação analítica tem sido possível obter conjuntos de dados com diferentes quantidades de informações, que podem ser valores escalares, vetores, matrizes ou tensores, dependendo da configuração instrumental empregada na análise [1]. Para uma melhor caracterização da dimensão dos dados, os mesmos podem ser classificados como: dados de ordem zero, primeira ordem, segunda ordem, terceira ordem e assim por diante [2].

Os dados de segunda ordem apresentam destaque em um grande número de aplicações encontradas na literatura, proporcionado principalmente pelos avanços na combinação de técnicas espectroscópicas e cromatográficas [3]. Para que os dados possam ser caracterizados como sendo de segunda ordem é necessário que os mesmos apresentem como configuração uma matriz por amostra. Normalmente, são registradas simultaneamente as informações obtidas por dois detectores diferentes [3-4]. Dentre as vantagens desse arranjo de sensores destaca-se a obtenção de um grande número de informações sobre as amostras. Como conseqüência, é possível obter um aumento na seletividade e quantificação de analitos na presença de interferentes não modelados ("*vantagem de segunda ordem*") [4-5].

Apesar dos ganhos associados aos dados de segunda ordem, a complexidade dos mesmos pode revelar dificuldades em sua interpretação. Nesse contexto, faz-se necessário o uso de ferramentas quimiométricas que possam auxiliar na extração de informações que possam ser utilizadas para identificação, classificação [8-10] ou quantificação de substâncias presentes nas amostras [6-7].

A classificação consiste na determinação e supervisão de padrões presentes em um

conjunto de amostras de treinamento, que possam ser utilizados para identificação de uma nova observação (amostra). Dentre as características que podem ser exploradas em estudos de classificação, estão: autenticidade, denominação de origem, prazo de validade, adulteração, etc. Essas propriedades podem ser facilmente acessadas, permitindo a execução de respostas analíticas rápidas, confiáveis e com menos custos [11-12].

Diferentes aplicações podem ser encontradas na literatura envolvendo a classificação de dados de segunda ordem, como por exemplo:categorização de amostras de vinagre de Jerez [9], caracterização e classificação de amostras de mel [10], diferenciação de bactérias [5], verificação da autenticidade de amostras de azeite de oliva [13]e classificação de vinho de acordo com a variedade da uva [14]. No geral, as estratégias utilizadas nesses estudos estão baseadas em técnicas de redução de dimensionalidade seguida de técnicas de classificação de primeira ou de segunda ordem, de acordo com o algoritmo escolhido.

Dentre as técnicas encontradas na literatura para redução da dimensionalidade de dados de segunda ordem, estão: Desdobramento (do inglês *Unfolding*) [15], Análise de fatores paralelos (PARAFAC, do inglês *Parallel factor analysis*) [16] e Tucker-3 [17].As mesmas podem ser associadas a técnicas de classificação de primeira ordem, como é o caso da: Análise discriminante linear (LDA, do inglês *Linear discriminant analysis*) [18] e Mínimos quadrados parciais para análise discriminante (PLS-DA, do inglês *Partial least square discriminant analysis*) [19]. Ainda, técnicas que utilizam a informação tridimensional dos dados podem ser aplicadas, como: Modelagem Independente por Analogia de Classe multivias (N-SIMCA, do inglês *N-way Soft independent modelling by class analogy*) [8] e Mínimos quadrados parciais para analise discriminante multivias (N-PLS-DA, do inglês *N-way partial least square discrminant analysis*) [14,20].

Apesar da quantidade de trabalhos envolvendo classificação de dados químicos de segunda ordem, poucos são os algoritmos que não são oriundos da adaptação da estratégias de calibração de segunda ordem (PARAFAC, TCUKER, etc...). Nessa direção, o desenvolvimento de algoritmos de classificação para dados de segunda ordem se torna uma área de grande demanda de estudos e aplicações.

Pode-se argumentar que o tratamento dos dados analíticos de segunda ordem poderia se beneficiar do uso de algoritmos de processamento de imagem, que também estão preocupados com estruturas de matriz (uma imagem digital é uma matriz onde linhas e colunas indicam a distribuição de pixels na imagem). Uma abordagem interessante para a classificação de imagens é o algoritmo de análise discriminante linear em duas dimensões (2D-LDA, do inglês *Two dimensional discriminant analysis*), que foi originalmente proposto por Li et al. [21] e tem sido aplicado, até o momento, apenas no contexto de processamento de imagens de face [22-23].

O algoritmo 2D-LDA baseia-se na extração características das matrizes de dados a partir da utilização de vetores de projeção otimizados com o critério de Fisher [24-26]. Semelhanças entre diferentes imagens podem então ser avaliados em termos da distância entre os vetores de característica correspondente.

Uma das principais vantagens de 2D-LDA, que pode contribuir diretamente na classificação dos dados químicos, consiste na redução da dimensão das matrizes de dados ao mesmo tempo que preserva a informação relevante para efeitos de classificação. Diferente de outros algoritmos, no 2D-LDA a redução de dimensionalidade dos dados da amostra, de matriz para vetor, é realizada a fim de sintetizar as características que auxiliam na classificação a partir do uso do critério de Fisher. Dessa maneira, não é necessário aplicar uma

um algoritmo de primeira ordem para classificação das amostras. Espera-se ainda que essas características do 2D-LDA tornem não prioritárias a seleção de variáveis ou intervalos.

Apesar do promissor desempenho demonstrado em trabalhos de processamento de imagem, o algoritmo 2D-LDA ainda não foi empregado em aplicações que envolvam dados químicos. Diante disso, a fim de preencher esta lacuna, o presente trabalho visa adequar e avaliar o uso do algoritmo 2D-LDA para análise de dados químicos, a fim de propor o mesmo como uma nova ferramenta quimiométria de classificação.

1.2 Objetivos

1.2.1 Objetivos Gerais

Adaptar o algoritmo 2D-LDA, proposto para estudos em de reconhecimento de imagens, para que possa ser aplicado como uma nova proposta em problemas envolvendo classificação baseada em dados químicos de segunda ordem.

1.2.2 Objetivos Específicos

- Implementar, em ambiente Matlab, o algoritmo 2D-LDA para classificação de dados químicos de segunda ordem;
- Simular conjuntos de dados de segunda ordem bilineares/não trilineares para estudos e avaliação do algoritmo proposto;
- Aplicar o algoritmo proposto em estudo envolvendo monitoramento da maturação de presunto de Parma usando fluorescência (dados trilineares);
- Aplicar o algoritmo proposto a um novo c*onjunto d*e dados: Dados de espectrometria de fluorescência sincrônica (não bilineares) para análise "*screening*" de matéria prima de óleos comestíveis;
- Comparar os resultados obtidos com os apresentados por outras técnicas disponíveis na literatura (PARAFAC-LDA, TUCKER-3-LDA e U-PLS-DA).

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Classificações dos dados analíticos e técnicas quimiométricas

Dependendo da configuração instrumental utilizada para medida de uma amostra, podem ser obtidos conjuntos de dados com diferentes dimensões. Essa condição serve de base para classificação dos dados e das ferramentas quimiométricas.

Se uma medida de um sistema químico gera um único valor, de forma que a repetição dessas medidas seja equivalente a um vetor, os dados são considerados de ordem zero ou de uma via ("*one-way*"). Os mesmos podem ser analisados a partir de parâmetros como, como: média, desvio padrão e variância [1]. Entre as técnicas que geram esse tipo de dados, estão: titulação, potenciometria, fotometria de chama, etc. Caso sejam necessárias medidas com um número maior de detectores para caracterizar as amostras, de forma que ocorra um aumento na dimensão dos dados, os mesmos poderão ser nomeados da seguinte maneira: dados de primeira ordem, segunda ordem, terceira ordem, e assim por diante.

Dados de primeira ordem ou de duas vias ("*two-way*") são obtidos por técnicas que geram um vetor por amostra, consequentemente, uma matriz de dados para um conjunto de amostras [4]. Aespectroscopia no infravermelho próximo (NIR, do inglês *Near-infrared spectroscopy*), Espectroscopia UV-Vis e a voltametria estão entre as técnicas que geram esse tipo de dado.Neste caso, ferramentas quimiométricas de primeira ordem podem ser utilizadas, como: Análise por componentes principais (PCA, do inglês *Principal component analysis*) [27], Regressão por mínimos quadrados parciais (PLSR, do inglês *Partial least square regression*) [28] e Regressão por componentes principais (PCR, do inglês *Principal componente regression*) [29].Dados de segunda ordem ou de três vias ("*three-way*") são obtidos por técnicas que geram uma matriz de dados por amostra, portanto, um cubo de dados para um conjunto de amostras [2]. Dentre essas técnicas estão: Espectroscopia de

fluorescência em matriz de excitação-emissão (EEM), ou a partir de técnicas hifenizadas como: Cromatografia líquida de alta eficiência com detecção por arranjo de diodo (HPLC-DAD), Cromatografia gasosa acoplada à espectrometria de massa (CG-MS) e Cromatografia liquida acoplada à espectrometria de massa (LC-MS). Para análise quimiométrica podem ser utilizadas técnicas de segunda ordem, como: PARAFAC [16], TUCKER-3 [17] e o N-PLS-DA [20,30]. Alguns desses algoritmos apresentam restrições em sua aplicação de acordo com a característica dos dados, apresentadas naseção 2.1.1.

Dados com número de ordem superior a dois, que apresentam mais que três vias ("*higher-way*") [2], podem ser obtidos por técnicas que apresente arranjos de detectores que forneçam três modos instrumentais para aquisição de dados por amostra. Entre os exemplos de técnicas que geram este tipo de dados estão: Cromatografia líquida bidimensional (2D-LC, do inglês *Two-dimensional liquid chromatography*) [31], comprimento de onda de excitação controlado no tempo (EMWTM, do inglês *Excitation modulated wavelength time matrix*) [32] e matriz de excitação emissão registrada com variação de pH (*excitation-emission-pH*) [33]. Todas essas técnicas geram matrizes do tipo quatro vias ("*four-way*"), ou seja, um cubo de dados por amostra. Ainda não existem muitas ferramentas quimiométricas disponíveis na literatura para tratar matrizes com essas dimensões, sendo mais pontualmente utilizados o PARAFAC [32], decomposição quadrilinear alternada (APQLD, do inglês A*lternating penalty quadrilinear decomposition*) e decomposição quadrilinear alternada com restrição (AWRCQLD, do inglês *Alternating weighted residue constraint quadrilinear decomposition*) [33].

Na **Figura 1** é apresentado um resumo da classificação dos dados químicos de acordo com a dimensão da matriz por amostra ou por conjunto de amostras.

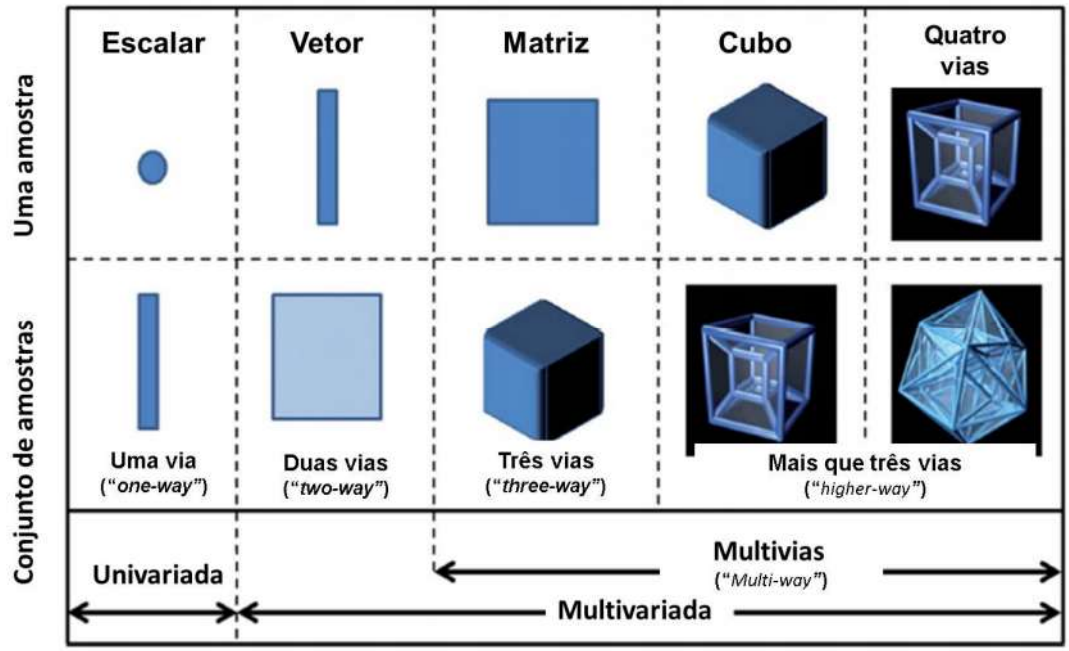


Figura 1 – Representação das diferentes matrizes de dados para uma amostra ou conjunto de amostras e nomenclatura empregada. (Fonte: Figura adaptada da referência: [4]).

2.1.1 Bilinearidade/trilinearidade dos dados

Bilinearidade/trilinearidade são características dos dados de segunda ordem e estão diretamente associadas à instrumentação analítica e a peculiaridades do sistema químico,e devem ser consideradas para escolha do algoritmo de modelagem [34].

Para um melhor entendimento do conceito de bilinearidade é apresentado na **Figura 2** uma matriz **X**, de dimensões J x K, obtida com HPLC-DAD para uma mistura de dois analitos. As colunas representam os espectros (b) e as linhas equivalem aos cromatogramas (c) registrados em cada comprimento de onda [35-36].

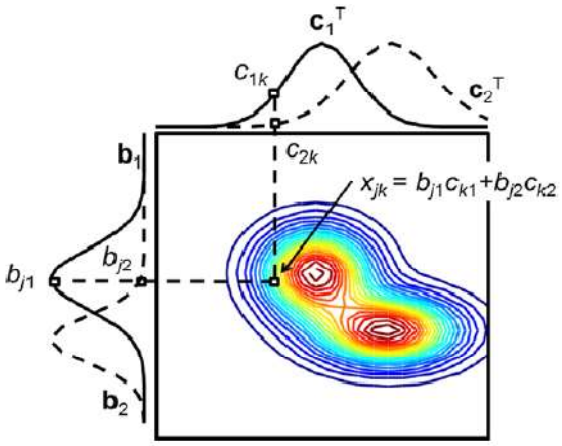


Figura 2 - Gráfica de contorno de uma matriz de dados obtidos com HPLC-DAD (com os níveis de intensidade crescentes de azul para vermelho). (Figura da referência: [6]).

A matriz **X**(**Figura 2**) será do tipo bilinear se a mesma puder ser representada por um produto de matrizes [34], como apresentado na **Equação 1**.

$$\mathbf{X} = BC^T \tag{1}$$

Sendo que as matrizes **B** e **C** sintetizam a soma das contribuições dos N analitos, de acordo com a **Equação 2**.

$$BC^T = b_1c_1^T + \ldots + b_Nc_N^T \tag{2}$$

Onde T na equação é equivalente à operação de transposição de c.

Se o número de N termos da **Equação 2** utilizada para modelar a matriz **X** for equivalente ao número de analitos, a matriz de dados será do tipo bilinear [1]. Caso essa condição não ocorra, a matriz de dados será do tipo não bilinear.

Nas matrizes não bilineares resposta instrumental R(b,c) para cada analito não pode ser representada por um único produto de duas funções individuais.Esse fenômeno ocorre principalmente quando os modos instrumentais *J* e *K* são mutuamente dependentes.

Um exemplo de dados não bilineares é o espectro de fragmentação (MS-MS). Neste

caso, é gerada uma matriz **X** (*JK*), onde cada *J* íons precursores são fragmentados para obter o *K* espectro de massas dos íons dos produtos gerados [37]. Logo, os perfis da razão m/z encontrados em *K* variam em função de duas condições: (1) estrutura das espécies precursoras presentes em *J*; (2) características de processo de quebra e ionização inerente à instrumentação/técnica. Neste caso, a matriz **X** (*JK*) não poderá ser representada pelo produto da função resposta nos modos *JK*, pois os termos em *K* são dependentes do comportamento das medidas no modo *J*.

O conceito de trilinearidade pode ser entendido como uma extensão da bilinearidade, só que aplicado a um tensor de dados $\underline{\mathbf{X}}$ (IJK). Neste caso, além dos modos *JK*, um novo modo (dimensão) chamado de *I* é acrescentado, sendo este relacionado às contribuições dos analitos presentes nas amostras individuais (**Figura 3**).

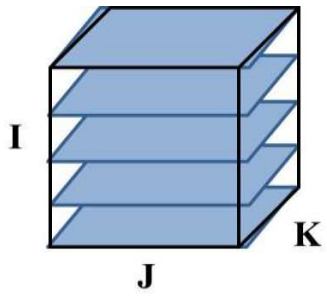


Figura 3 - Representação do cubo de dados $\underline{\mathbf{X}}$ (*I,J, K*) gerado a partir do empilhamento matrizes de dados das amostras.

Para que o tensor de dados $\underline{\mathbf{X}}$ seja trilinear, as matrizes das amostras devem ser do tipo bilinear e a quantidade de contribuições dos analitos no modo *I* devem ser iguais à quantidade desses nos modos *JK*. Dessa forma, um elemento (constituinte do tensor) $\mathbf{x}_{ijk}$ poderá ser determinado por meio da **Equação 3**.

$$x_{ijk} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} \quad (3)$$

Onde: a_{in} é proporcional à concentração do constituinte n na amostra i, b_{jn} à contribuição do constituinte n no modo instrumental j, e c_{kn} é equivalente à contribuição do elemento n no modo k.

A quebra da trilinearidade do tensor de dados $\underline{\mathbf{X}}$ pode ocorrer se alguma das amostras empilhadas não apresente bilinearidade ou se as condições de medição dos modos *J* e *K* forem diferentes entre as amostras. Neste caso, o numero de componentes para modelar o tensor de dados $\underline{\mathbf{X}}$ pode ser diferente entre os modos *I*, *J* e *K*, de forma que não possa constituir um produto de matrizes.

2.2 Técnicas de redução de dimensionalidade para dados de segunda ordem

2.2.1 Desdobramento ("unfolding")

O desdobramento consiste no "rearranjo" dos dados tridimensionais de forma a obter uma matriz bidimensional que permita a utilização de técnicas multivariadas de primeira ordem [15]. Para tanto, é utilizado um procedimento parecido com um "fatiamento" do tensor de dados. Essa estratégia pode ser aplicada a dados trilineares e não trilineares.

Três maneiras de concatenação podem ser aplicadas a um tensor $\underline{\mathbf{X}}$ de dimensões *I* x *J* x *K*, a fim de se realizar o desdobramento: *I* x *JK* (**Figura 4a**), *J* x *KI* (**Figura 4b**) *ou K* x *JI* (**Figura 4c**) [38].

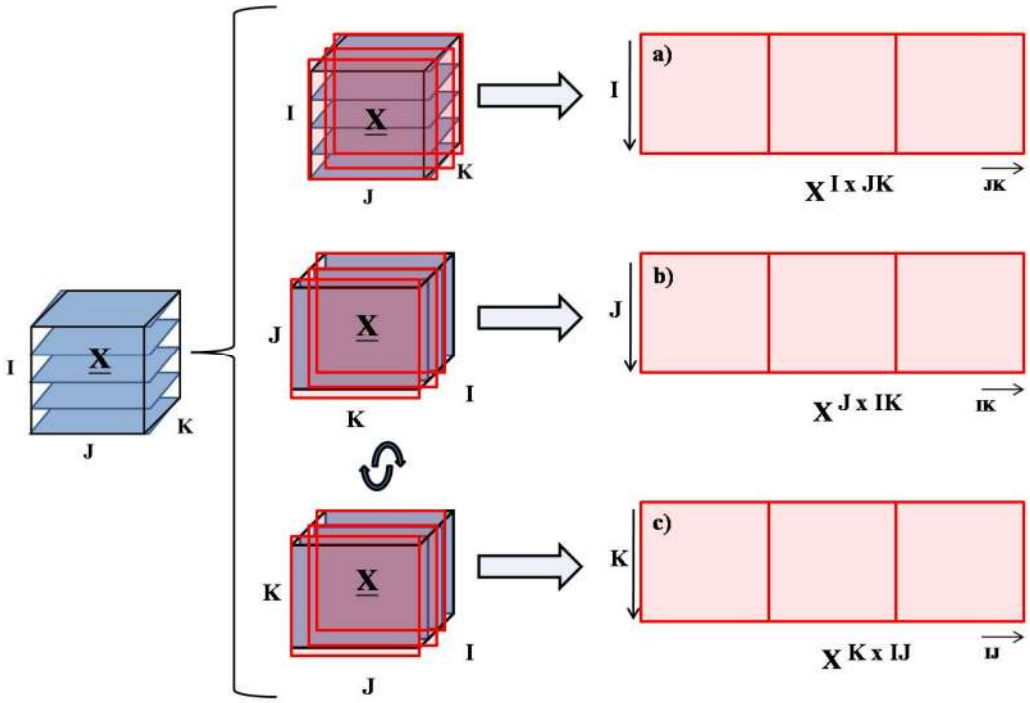


Figura 4 - Maneiras de realizar o rearranjo de um cubo de dados usando desdobramento.

A maneira mais comum de realizar o desdobramento dos dados é concatenando as variáveis de forma a gerar uma matriz do tipo: *I* x *JK* (**Figura 4a**). Esse tipo de concatenação permite a obtenção de um vetor de dados por amostra, com um número final de variáveis igual ao produto do número de variáveis presentes nos modos *J* e *K*. Para uma melhor visualização, na **Figura 5** é apresentado o desdobramento de uma matriz de dados simulados de EEM no sentido *I* x *JK*.

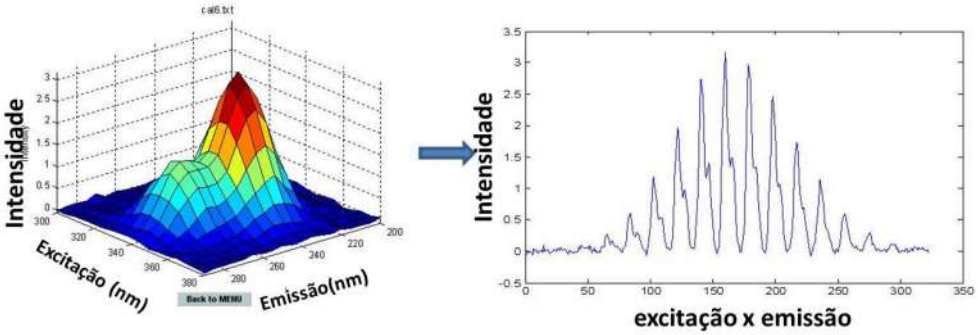


Figura 5 - Desdobramento de uma matriz de dados simulados de EEM na direção I x JK.

2.2.2 Métodos de Tucker

Os métodos de Tucker foram propostos por volta do ano de 1966 e recebem o nome do psicometrista Ledyard R. Tucker [17]. Dependendo da maneira que for empregado para decompor os dados de segunda ordem, três métodos podem ser obtidos: Tucker-1, Tucker-2 e Tucker-3 [38].

O Tucker-1 consiste na aplicação individual de uma PCA nas três formas de desdobramento (*I* x *JK, J* x *KI ou K* x *JI*), apresentado na **Figura 4** da seção *2.2.1*. Em seguida, os pesos de PCA são desconsiderados e os valores de escores são analisados [17]. Uma das principais vantagens do Tucker-1 está na simplicidade de implementação. No entanto, esse modelo apresenta desvantagens quando comparado a modelos que levam em consideração a tridimensionalidade dos dados, já que a analise usando Tucker-1 é realizada individualmente para cada um dos modos de medida [26]. O Tucker-2 é um caso particular, e será abordado após a descrição do Tucker-3.

O Tucker-3 é aplicado em dados de segunda ordem (Trilineares e não-trilineares) e permite a decomposição das informações em três matrizes, por exemplo, no caso de HPLC-DAD: **A** (composição), **B** (espectros) e **C** (tempo de eluição). Essas matrizes apresentam interações que são dadas a partir de uma matriz conectora **G** ("*core array*"). Os elementos do tensor **G** indicam a importância de cada interação entre as respostas dos fatores (quando bem ajustado, equivale aos analitos). Neste caso, valores da matriz conectora G que estão próximos de zero indicam um baixo nível de interação entre os fatores. Graficamente o método Tucker-3 pode ser representado de acordo com a **Figura 6** [38,39].

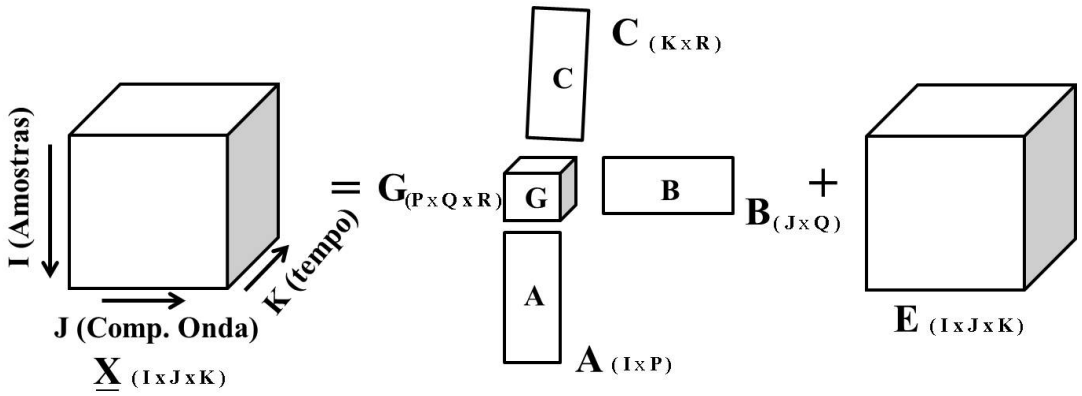


Figura 6- Representação gráfica para o modelo Tucker-3 (Figura adaptada da referência: [38]).

Na **Equação 4** é apresentada decomposição de um tensor de dados $\underline{\mathbf{X}}$ usando o Tucker-3 [26].

$$\underline{\mathbf{X}} = \mathbf{AG}(\mathbf{C} \otimes \mathbf{B})^t + \underline{\mathbf{E}} \tag{4}$$

Onde: as matrizes **A** (*I* x *P*), **B** (*J* x *Q*) e **C** (*K* x *R*) contêm os perfis ("loadings") dos fatores relativos às dimensões *I*, *J* e *K* respectivamente; A matriz **G** (*P* x *Q* x *R*) é a matriz conectora; Os índices *P*, *Q* e *R* são o número de fatores obtidos em cada um dos modos, os quais podem apresentar valores diferentes entre si; $\underline{\mathbf{E}}$ (*I* x *J* x *K*) com erros de aproximação do modelo; O símbolo $\otimes$ equivale ao produto de Kronecker (**Anexo 1**) [40-41].

Na **Equação 5** temos a equação do Tucker-3 para um único elemento do cubo de dados [39].

$$x_{ijk} = \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} \right) + e_{ijk} \tag{5}$$

Onde: x_{ijk} equivale a um elemento do cubo de dados $\underline{\mathbf{X}}$; a_{ip}, b_{jq}, c_{kr} e g_{prq}, são os valores correspondentes à x_{ij} nas matrizes **A**, **B** e **C** obtidos para os modos *I*, *J* e *K*; e_{ijk} está relacionado ao erro de aproximação para o valor de x_{ijk}.

Os fatores obtidos com o Tucker-3 por muitas vezes são restringidos a ser ortogonais entre si, permitindo a obtenção de resultados mais simples e com menor demanda de tempo computacional.No entanto, esse tipo de modelo não apresenta "unicidade", ou seja, diferentes soluções similares podem ser obtidas.

O Tucker-2 possui descrição semelhante a do Tucker-3, porém o modelo não explora totalmente a estrutura de três vias dos dados. Neste caso, um dos modos da matriz não é descompactado, ou seja, o mesmo não é considerado no cálculo do modelo [42], resultando na **Equação 6**.

$$x_{ijk} = \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip} b_{jq} g_{pq} \right) + e_{ij} \qquad (6)$$

Onde: x_{ijk} equivale a um elemento do tensor de dados; a_{ip},b_{jq} e g_{pq}, são os valores correspondentes à x_{ij} nos pesos *A* e *B* obtidos para os modos *I*,*J* e g_{pr} equivale ao elemento da matriz de núcleo referente à x_{ij}.

2.2.3 PARAFAC

O PARAFAC foi inicialmente proposto por Harshman [43] e por Carroll e Chang [44], e consiste na decomposição de tensores visando à identificação e quantificação de perfis de fatores (em condições bem ajustadas, os fatores equivalem aos analitos presentes nas amostras).

De forma semelhante ao Tucker-3, o PARAFAC gera três matrizes de pesos ("loadings"): **A**,**B** e **C**, como apresentado na **Figura 7**.

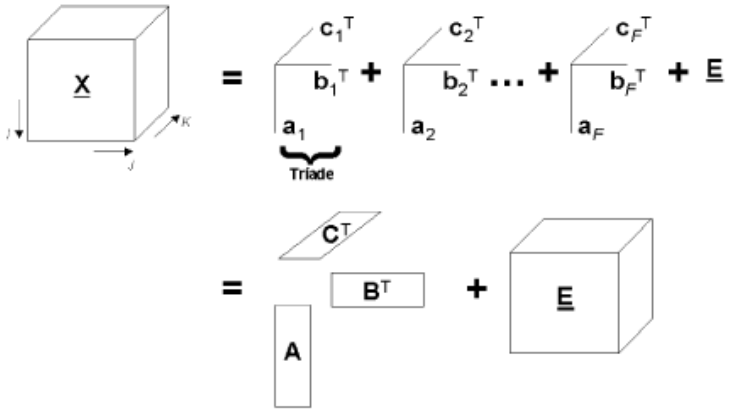


Figura 7-Representação gráfica para o modelo PARAFAC. Decomposição em F tríades de vetores de peso. (Figura da referência: [38]).

O modelo trilinear calculado busca minimizar a soma dos quadrados dos resíduos [16], como apresentado equação na **Equação 7** para um elemento do tensor $\underline{\mathbf{X}}$.

$$x_{ijk} = \sum_{f=1}^{F} a_{ip} b_{jq} c_{kr} + e_{ijk} \qquad (7)$$

Onde: x_{ijk} equivale a um elemento do tensor de dados; a_{ip},b_{jq},c_{kr} são os valores correspondentes à x_{ijk} nas matrizes de peso **A**, **B** e **C** respectivamente, obtidos para os modos *I*,*J* e *K*.

Algumas características diferenciam o PARAFAC e o fazem ser visto como um caso restrito do Tucker-3, dentre eles estão: igualdade do número de fatores para as três matrizes de pesos ("loadings"); a matriz de núcleo ("*core array*") comporta-se como uma matriz superidentidade; a solução do modelo gerado apresenta unicidade ("*uniqueness*"), ou seja, um único resultado é obtido para as mesmas condições de entrada no algoritmo. Essas características tornam o uso do PARAFAC popular na quimiometria, especialmente na determinação de perfis de compostos puros [45].

O número de fatores (F) obtidos com um modelo PARAFAC, quando bem ajustado, pode ser relacionado ao número de constituintes químicos presentes na amostra. Existem diferentes estratégias para escolha do número apropriado de fatores, dentre eles: conhecimento da composição química da amostra, validação cruzada e/ou reamostragem ("*split-half*") [38], variância explicada pelo modelo ou por método automático usando diagnóstico de concordância de núcleo (CORCONDIA, do inglês *Core Consistency Diagnostic*).

O CORCONDIA está associado à interpretação do PARAFAC como uma restrição do Tucker-3. Desta forma, caso os dados apresente tendência a trilinearidade, espera-se que os elementos da superdiagonal da matriz conectora (G) apresente valores próximos de 1 e os demais elementos valores próximos de zero [38, 46]. Um valor de CORCONDIA próximo de 100% indica a adequação dos dados ao PARAFAC, um valor em torno de 50% indica deficiência de trilinearidade e valores próximos de zero ou negativos indicam inconsistência trilinear [38]. O cálculo do valor de CORCONDIA pode ser encontrado no **anexo 1**.

2.3- Técnicas de classificação multivariada

2.3.1 Análise Discriminante Linear - LDA

A LDA foi inicialmente utilizada por Barnard [47] e Fisher [18], sendo bastante aplicada para exploração, classificação e controle de qualidade em estudos envolvendo diferentes conjuntos de dados. A classificação de uma amostra pertencente a uma matriz **X** (*I*, *J*), onde *I* equivale ao modo das amostras e *J* ao de variáveis, pode ser dada por intermédio da combinação linear das j características (variáveis), obtida através de funções discriminantes geradas com base no critério de projeção linear de Fisher [24-25]. Esse critério visa maximizar a razão “espalhamento entre as classes (Sb) pelo espalhamento no interclasse

(Sw)", como apresentado na **Figura 8**.

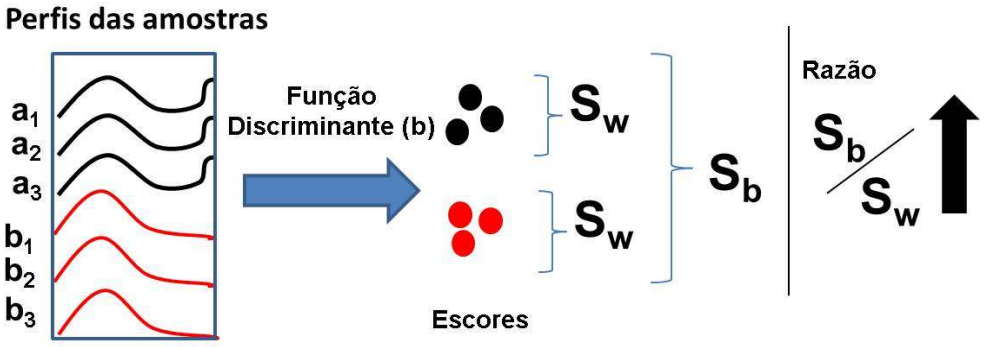


Figura 8 - Representação do principio da LDA. Sw (Espalhamento intra classe) e Sb (Espalhamento entre classes). (Fonte: própria)

Na **Equação 8** é possível encontrar a operação matemática correspondente à determinação da função discriminante (b) dada pelo critério de Fisher [22-24].

$$\mathbf{b} = \arg\max_{\mathbf{b}} \frac{\mathbf{b}^T \mathbf{S}_B \mathbf{b}}{\mathbf{b}^T \mathbf{S}_W \mathbf{b}} \tag{8}$$

Onde o $\mathbf{S_B}$ e o $\mathbf{S_W}$ são, respectivamente, as matrizes de espalhamento entre classes e intra classe, e podem ser determinados pelas **Equações 9** e **10**:

$$\mathbf{S_B} = \sum_{p=1}^{L} N_p (\overline{\mathbf{X}}_p - \overline{\mathbf{X}})^T (\overline{\mathbf{X}}_p - \overline{\mathbf{X}}) \tag{9}$$

$$\mathbf{S_W} = \sum_{p=1}^{L} \sum_{k \in I_p} (\mathbf{X}_k - \overline{\mathbf{X}}_p)^T (\mathbf{X}_k - \overline{\mathbf{X}}_p) \tag{10}$$

Sendo que $\overline{\mathbf{X}}_p$ equivale a média das amostras pertencentes a uma classe p de um total de L classes. $\overline{\mathbf{X}}$ está relacionado a media de todas as amostras. As médias podem ser obtidas como apresentado na **Equação 11** e **12**.

$$\overline{\mathbf{X}}_p = \frac{1}{N_p} \sum_{k \in I_p} \mathbf{X}_k \tag{11}$$

$$\overline{\mathbf{X}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{X}_k \tag{12}$$

O calculo de autovalores generalizado pode ser utilizado na obtenção das funções discriminantes, de acordo com a **Equação 13**:

$$\mathbf{S_W^{-1}S_B b} = \lambda \mathbf{b} \tag{13}$$

Onde λ são os autovalores para a razão $\mathbf{S_W^{-1}S_B}$.

O autovetor com maior autovalor equivale à primeira função discriminante linear (**b**1), o segundo maior autovalor estará relacionado à segunda função discriminante linear (**b**2), e assim consecutivamente até se determinar a quantidade de funções que possam ser adequadas para o problema.

A classificação de um objeto desconhecido pode ser dada a partir da distância euclidiana entre as projeções desse objeto e a media das L classes. A amostra será atribuída à ao centro da classe com menor distância. Como apresentado na **Equação 14** [48].

$$\min_j \left\| b^T (x_{desconhecido} - \bar{x}_p) \right\| \text{ com p= 1,2, …,L} \tag{14}$$

A LDA é encontrada na literatura aplicada a dados químicos de primeira ordem obtidos com diferentes técnicas, como: Espectroscopia UV-Vis [49], NIR [50], eletroanalítica [51], HPLC com espectrometria de massas [52], ressonância magnética nuclear [53], entre outros. Vale ressaltar que na maioria dessas aplicações os dados químicos apresentam um número de variáveis maior que o de amostras. Essa característica gera uma impossibilidade na determinação da pseudo-inversa apresentada na **Equação 13**. Logo, estratégias para redução de dimensionalidade de varáveis devem ser adotadas [54], como: análise por componentes principais (PCA) [27], algoritmo das projeções sucessivas (SPA) [55], algoritmo genético

(GA) [56], entre outros.

Quando se trata de dados químicos de segunda ordem, os trabalhos encontrados na literatura com aplicação da LDA necessitam de uma etapa de decomposição prévia, neste caso, cada amostra será representada por uma matriz de dados, necessitando de outros métodos para redução do numero de variáveis. Diante desse inconveniente, neste trabalho são aplicados PARAFAC e Tucker-3 como técnicas para decomposição do tensor como etapa prévia à aplicação da LDA.

2.3.2 Mínimos Quadrados Parciais para Analise Discriminante - PLS-DA

A técnica de classificação por mínimos quadrados parciais para análise discriminante (PLS-DA) [19] é utilizada para discriminação entre diferentes classes de amostras atravésda obtenção de um modelo de regressão entre duas matrizes: matriz de dados instrumentais (**X**) e matriz de classes (**Y**) [57]. Para tanto, o mesmo utiliza-se dos princípios do método de regressão por mínimos quadrados parciais (PLS) [28], no qual é realizada a predição da variável dependente **Y**a partir da maximizaçãoda covariância da mesma com as variáveis independentes pertencentes a **X** em um novo subespaço de projeção das matrizes. Caso existam apenas duas classes, o PLS1 é aplicado e a matriz Y será considerada um vetor composto por [0 1], onde: 0 (atribuído para amostras da classe 1) e 1 (atribuído para amostras da classe 2) (**Figura 9**). Caso existam mais classes, valores binários devem ser atribuídos às amostras, e o PLS2 deve ser considerado [57].

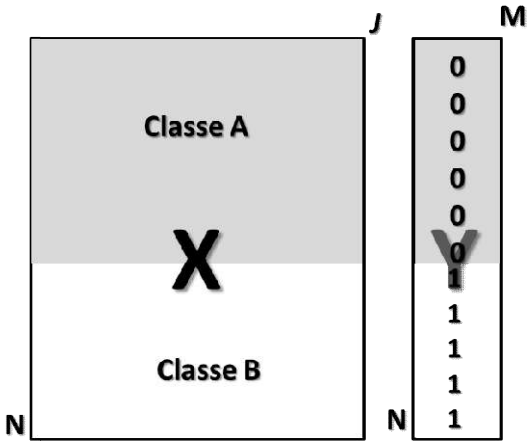


Figura 9 - Estrutura de apresentação das matrizes de dados instrumentais e do vetor de índice de classe. Onde: N = número de amostras, J = número de variáveis, e M = número de descritores. (Fonte: Própria)

As operações matemáticas fundamentais para modelagem PLS-DA para classificação são apresentadas nas equações **15** e **16** [19].

$$\mathbf{X} = tp + \mathbf{E} \tag{15}$$

$$\mathbf{y} = tq + \mathbf{f} \tag{16}$$

Note que **X** e **y** são decompostos em estruturas semelhantes, onde o termo t (escore) é semelhante para as duas equações. As matrizes **E** e**f** equivalem aos resíduos e os parâmetros *p* e *q* são pesos ("loadings") das matrizes **X** e **y**. Vale salientar que é importante realizar todos os pré-processamentos necessários (seleção de intervalo de variáveis informativas, correção de linha de base, remoção de ruídos, normalização, centragem na média) antes executar a rotina do PLS-DA.

A seguir, são apresentados os passos do algoritmo PLS1, de acordo com Brereton et al. [57]:

Passo 1 - Calculo do **w**(vetor de pesos PLS) (**Equação 17**), que será utilizado para estimar os valores de escores.

$$\mathbf{w} = \mathbf{X}'\mathbf{y} \tag{17}$$

Passo 2 - Calculo dos valores de escores (**t**) comuns às matrizes **X** e **Y.** (**Equação 18**).

$$\mathbf{t} = \frac{\mathbf{Xw}}{\sqrt{\sum w^2}} \tag{18}$$

Passo 3 -Determinação dos pesos ("loadings") (**Equação 19** e **20**), para a matriz **X (p)** e **Y (q).**

$$\mathbf{p} = \frac{t'X}{\sum t^2} \tag{19}$$

$$\mathbf{q} = \frac{y't}{\sum t^2} \tag{20}$$

Passo 4- Calculo dos valores de resíduos par a matriz **X** (**E**) (**Equação 21)**e **y** (**f**) (**Equação 22).**

$$\mathbf{X}^{\mathbf{resid}} = \mathbf{X} - \mathbf{tp} \tag{21}$$

$$\mathbf{y}^{\mathbf{resid}} = \mathbf{y} - \mathbf{tq} \tag{22}$$

Os valores de escores, pesos e resíduos são referentes ao primeiro componente de PLS. Caso sejam necessários mais componentes PLS para explicar os dados, as matrizes de resíduos(**X**resid e **y**resid) devem ser utilizadas no lugar das matrizes **X** e **y** a partir do passo 1, de maneira a se obter os parâmetros do segundo componente de PLS. Esse procedimento deve ser repetido até que um critério de parada seja obedecido.

Passo 5 Estimativa do vetor **b** de coeficientes de regressão (**Equação 23**).

$$\mathbf{b} = \mathbf{W}(\mathbf{PW})^{-1}\mathbf{q} \tag{23}$$

Passo 6 Predição do valor de amostra desconhecida (**Equação 24**).

$$\hat{y} = \mathbf{xb} \tag{24}$$

Passo 7 Atribuição de classe à amostra desconhecida.

Umas das formas de decidir a qual classe a amostra desconhecida será atribuída é avaliando os valores dos índices de classes para as amostras de treinamento. Considerando duas classes, sendo que a primeira delas recebeu índice 1 e a segunda recebeu índice -1, a amostra desconhecida pertencerá à primeira classe se o valor predito for maior que 0, caso contrário, a mesma pertencerá à segunda classe. Outras formas de determinar o limiar de classificação podem ser usadas, como a determinação dos pontos centrais a partir da medida da média das médias das classes ou a partir da avaliação de curvas ROC [58].

2.3.3 N-PLS-DA

O N-PLS-DA tem uma configuração muito semelhante à apresentada pelo PLS-DA, no entanto, o mesmo se baseia no algoritmo de regressão N-PLS (*Partial least square n-way*) [39]. Neste caso, é maximizada a covariância entre os dados instrumentais de segunda ordem $\underline{\mathbf{X}}$ (*I*,*J*,*K*) e o vetor **y** que contém os índices de classes. Para tanto,uma decomposição de $\underline{\mathbf{X}}$ (*I*,*J*,*K*) e **y** é realizada. Dessa forma são obtidas as matrizes de escores (**T**) e pesos (**Wj** e **Wk**) para $\underline{\mathbf{X}}$, e os escores (**U**) para **y** como apresentado na **Figura 10**.

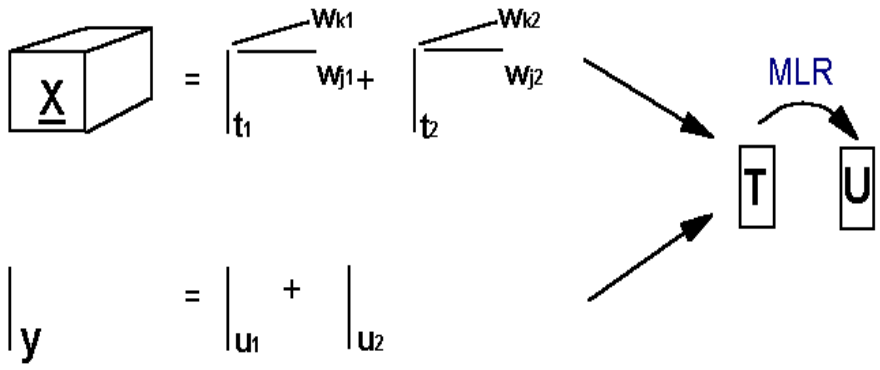


Figura 10 - Decomposição em dois componentes para um cubo de dados em N-PLS-DA. (Figura da referência: [59]).

O algoritmo N-PLS-DA segue uma rotina muito semelhante aos passos do PLS-1, de forma a obter os escores de $\underline{\mathbf{X}}$ e **y** simultaneamente, a fim de se obter um vetor de coeficientes (V) (**Equação 25**) que contribuam na predição de amostras futuras ($\hat{y}$) (**Equação 26**) [34].

$$v = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{y} \tag{25}$$

$$\hat{y} = u^t v \tag{26}$$

A atribuição de classe a uma nova amostra será realizada avaliando os valores de **Y** das amostras de treinamento. Considerando duas classes, sendo que a primeira delas recebeu índice 1 e a segunda recebeu índice -1, a amostra desconhecida pertencerá à primeira classe se o valor de $\hat{y}$ for acima de 0, caso contrário, a mesma pertencerá à segunda classe.

CAPÍTULO 3

ALGORITMO PROPOSTO

3. ALGORITMO PROPOSTO

3.1Analise discriminante linear em duas dimensões (2D-LDA)

Neste capítulo é apresentada uma descrição do algoritmo 2D-LDA. O mesmo foi originalmente proposto por Li et al. [21] no contexto de processamento de imagem de face, e foi adaptado nessa tese para ser utilizado com dados químicos de segunda ordem. Inicialmente o banco de dados é divido em dois: amostras de treinamento e amostras de teste. Em seguida são obtidos os vetores de projeção otimizados, os quais são extraídos de acordo com o critério de projeção linear de Fisher [24-26] a partir as amostras de treinamento. Vetores de características das amostras são gerados através da transformação das matrizes de dados utilizando os vetores de projeção. A classificação das amostras é dada em termos da distância entre os vetores de características correspondentes a amostra desconhecida e os vetores das amostras de cada classe de treinamento.

3.1.1 Notação

Matrizes, vetores e escalares são representados por letras maiúsculas em negrito, letras minúsculas em negrito e letras em itálico (maiúscula ou minúsculas), respectivamente. O T e -1 sobrescritos denotam transposta e inversa de uma matriz. A dimensão de matrizes e vetores são indicadas entre parentes. Um elemento (i, j) de uma matriz $\mathbf{X}$ é denotado como $X_{i,j}$.

As L classes envolvidas no problema são indicadas por C_1, C_2, ..., C_L. É assumido que o grupo de amostras de treinamento é formado por um conjunto de N matrizes $\mathbf{X}_k$ (k = 1, 2, ..., N), e corresponde às amostras de classe conhecida. Os índices de classe de N_p amostra de treinamento, sendo a classe pth será denotada por I_p, com p = 1, 2, ..., L. A notação $\sum_{k \in I_p} \mathbf{X}_k$ será utiliza para indicar a soma das matrizes correspondente às amostras na pth classe.

A notação traço(**M**) indica o traço (soma dos elementos da diagonal) de uma matriz quadrada **M**. A partir dessa notação, a soma dos quadrados dos elementos de uma matriz **X** (*m*×*n*) pode ser expressa por:

$$\text{tra}\,ço(\mathbf{X}^T\mathbf{X}) = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{i,j})^2 \tag{33}$$

3.1.2 Determinação dos vetores de projeção

Dada uma matriz **X** (*m*×*n*) de dados referentes a uma amostra, onde (*m*) e (*n*) correspondem aos registros realizados nos dois modos instrumentais de uma técnica de segunda ordem, é possível obter vetor de características **y** (*m*× 1),multiplicando **X** por um vetor de projeção **b** (*n*× 1), de acordo com a **Equação 34**.

$$\mathbf{y} = \mathbf{Xb} \tag{34}$$

O *i*th componente do vetor **y** é dado por um produto escalar entre a *i*th linha da matriz **X** e o vetor de projeção **b**, como apresentado na **Equação 35**.

$$y_i = \sum_{j=1}^{n} X_{i,j}b_j \tag{35}$$

No caso de dados de EEM, por exemplo, o número de linhas (*m*) corresponde aos comprimentos de onda de excitação e o número de colunas (*n*) é condizente aos comprimentos de onda de emissão. O *i*th componente do vetor de características **y** equivale à combinação linear dos comprimentos de onda de excitação, com coeficientes referentes aos valores do vetor **b**,para cada um dos comprimentos de onde de emissão (**Figura 11**).

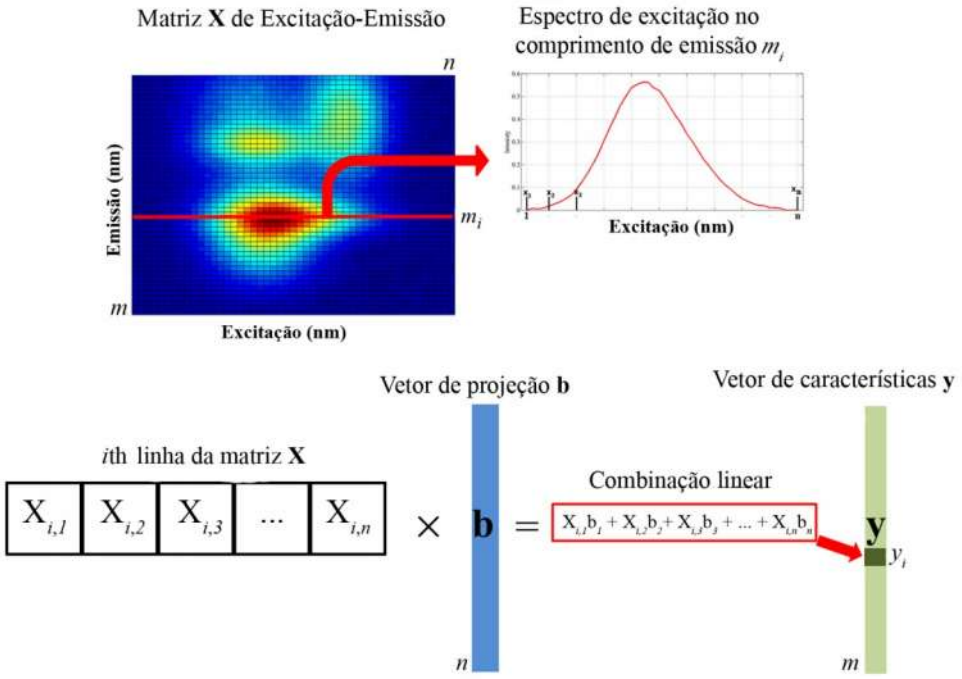


Figura 11- Determinação de cada elemento do vetor de características de uma medida de fluorescência em EEM.

O objetivo principal do algoritmo 2D-LDA é encontrar um vetor de projeção **b** que realce as características discriminantes de cada uma das classes, auxiliando na classificação das amostras a partir do vetor de características **y**.

Um vetor de projeção ótimo ($\mathbf{b}_{opt}$) pode ser obtido a partir da maximização do critério de projeção linear de Fisher, como apresentado na **Equação 36**.

$$\mathbf{b}_{opt} = \arg\max_{\mathbf{b}} \frac{\mathbf{b}^T \mathbf{S}_B \mathbf{b}}{\mathbf{b}^T \mathbf{S}_W \mathbf{b}} \tag{36}$$

Onde $\mathbf{S_B}$ ($n \times n$) e $\mathbf{S_W}$ ($n \times n$) denotam o espalhamento entre classes e intra classe, e podem ser calculados a partir dos dados de treinamento, de acordo com as **Equações 37** e **38**.

$$\mathbf{S_B} = \sum_{p=1}^{L} N_p (\overline{\mathbf{X}}_p - \overline{\mathbf{X}})^T (\overline{\mathbf{X}}_p - \overline{\mathbf{X}}) \tag{37}$$

$$\mathbf{S_W} = \sum_{p=1}^{L} \sum_{k \in I_p} (\mathbf{X}_k - \overline{\mathbf{X}}_p)^T (\mathbf{X}_k - \overline{\mathbf{X}}_p) \tag{38}$$

Onde $\overline{\mathbf{X}}_p$ equivale à media das matrizes **X** para o conjunto de amostras de treinamento pertencentes à classe C_p; e $\overline{\mathbf{X}}$ denota a media de todo o conjunto de amostras de treinamento, e podem ser determinados com a **Equação 39** e **40**.

$$\overline{\mathbf{X}}_p = \frac{1}{N_p} \sum_{k \in I_p} \mathbf{X}_k \tag{39}$$

$$\overline{\mathbf{X}} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{X}_k \tag{40}$$

Na **figura 12** é representado graficamente o cálculo das matrizes média utilizadas para determinação dos espalhamentos intra classe e entre classes.

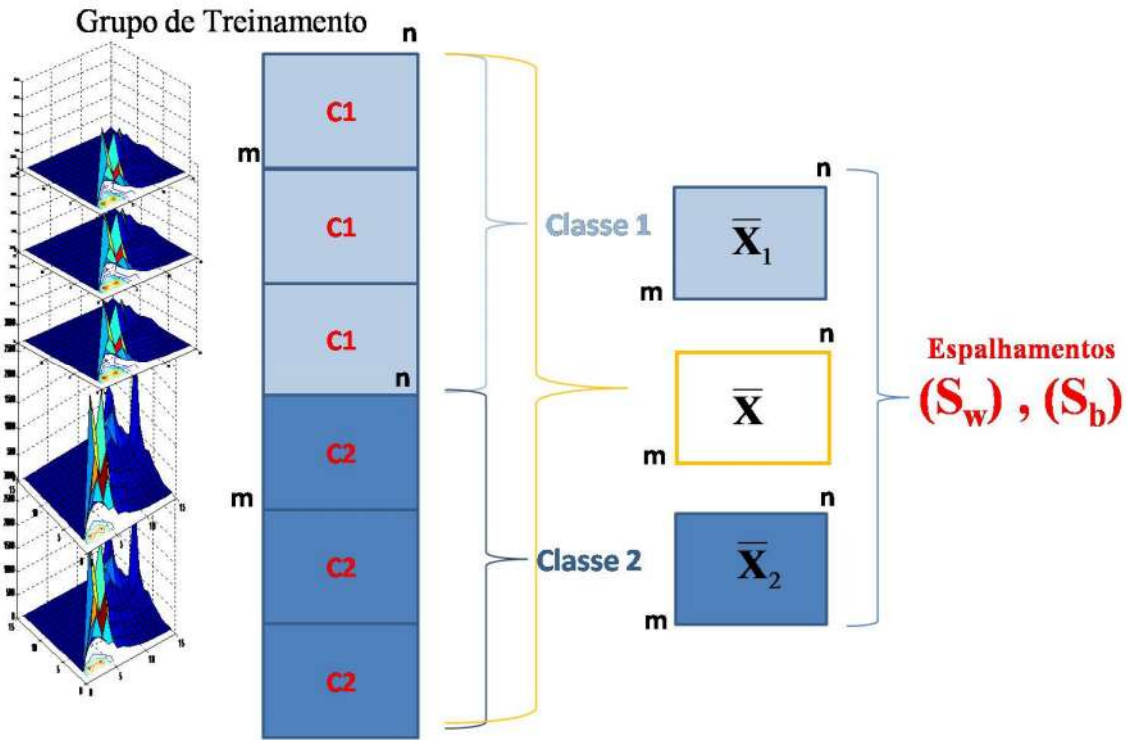


Figura 12- Média das amostras de treinamento utilizadas na obtenção das matrizes de espalhamento

Observação: Condiciona-se que a matriz por amostra, apresentada na figura 12, possua o número de variáveis em n menor do que o apresentado em m. Neste caso a matriz de espalhamento intra classe **Sw** terá dimensão (n x n). Essa estratégia é utilizada a fim de evitar que ocorra alguma inconsistência no cálculo da $\mathbf{S}_{\mathbf{W}}^{-1}$ (pseudo inversa de **Sw**), quando da determinação do vetor de projeções $\mathbf{b}_{opt.}$

Se $\mathbf{S_W}$ for não singular, $\mathbf{b}_{opt}$ pode ser obtido como um autovetor resultante de um problema de autovalores generalizados. Sendo assim, $\mathbf{b}_{opt}$ deve está de acordo com a **Equação 41**:

$$\mathbf{S}_{\mathbf{W}}^{-1}\mathbf{S}_{\mathbf{B}}\mathbf{b}_{\mathbf{opt}} = \lambda\mathbf{b}_{\mathbf{opt}} \tag{41}$$

Onde λ é o maior autovalor referente a matriz $\mathbf{S}_{\mathbf{W}}^{-1}\mathbf{S}_{\mathbf{B}}$ ($n \times n$). Esse procedimento pode ser ampliado quando for necessário mais de um vetor de projeção. Neste caso, um número de M vetores de projeção em ordem decrescente de relevância para classificação. O valor máximo de M será: $M \leq n$, onde n é equivalente ao ranque da matriz $\mathbf{S}_{\mathbf{W}}^{-1}\mathbf{S}_{\mathbf{B}}$. O qth vetor de projeção $\mathbf{b}_q$ é obtido com a resolução da **Equação 42**.

$$\mathbf{S}_{\mathbf{W}}^{-1}\mathbf{S}_{\mathbf{B}}\mathbf{b}_q = \lambda_q\mathbf{b}_q \tag{42}$$

Onde λ_q equivale ao qth maior autovalor de $\mathbf{S}_{\mathbf{W}}^{-1}\mathbf{S}_{\mathbf{B}}$, com $q = 1, 2, ..., M$.

Uma vez que sejam obtidos r vetores de projeção, sendo $r \leq M$, é possível organizar os $\mathbf{b}_1$, $\mathbf{b}_2$,..., $\mathbf{b}_r$ vetores em uma única matriz de projeção **B** ($n \times r$), que permite encontrar uma matriz de características **Y** ($m \times r$) calculada a partir de uma matriz **X** como apresentado na **Equação 43**.

$$\mathbf{Y} = \mathbf{XB} \tag{43}$$

As colunas de **Y** correspondem aos r vetores de características ordenados em ordem decrescente de relevância para o problema de classificação.

Na **figura 13** é apresentado graficamente como é feita a obtenção de um vetor de características para as amostras de treinamento e teste a partir do vetor de projeção **b**.

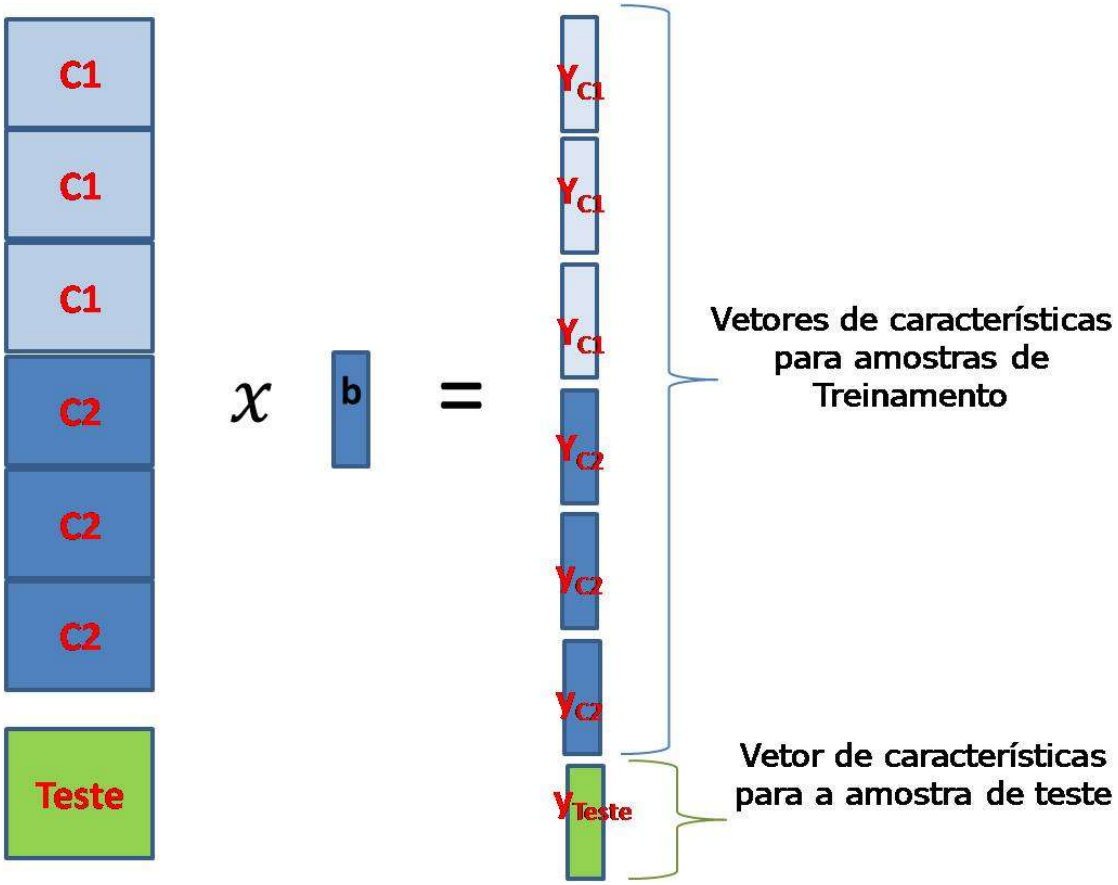


Figura 13- Obtenção dos vetores de características a partir o vetor de projeção **b**.

3.1.3 Procedimento de classificação

A classificação de uma amostra de teste $\mathbf{X}_{test}$ pode ser realizada de acordo com a similaridade da sua matriz de características $\mathbf{Y}_{test} = \mathbf{X}_{test}\mathbf{B}$ com respeito as matrizes de características $\mathbf{Y}_k = \mathbf{X}_k\mathbf{B}$, $k = 1, 2, ..., N$, das amostras pertencentes ao grupo de treinamento. Para determinação do valor da similaridade foi utilizado neste trabalho o calculo de distancia Euclidiana $d(\mathbf{Y}_{test}, \mathbf{Y}_k)$, que pode ser calculado pela **Equação 44**.

$$d(\mathbf{Y}_{test}, \mathbf{Y}_k) = \sqrt{\text{tra}\,\text{ço}[(\mathbf{Y}_{test} - \mathbf{Y}_k)^T (\mathbf{Y}_{test} - \mathbf{Y}_k)]} \quad (44)$$

Para k = 1, 2,..., N, com N igual ao número total de amostras de treinamento.

Uma vez obtidos os valores de distancia d ($\mathbf{Y}_{test}$, $\mathbf{Y}_k$) é possível determinar o valor de distancia média entre a amostra de teste e as N_p amostras de treinamento pertencentes à classe C_p de acordo com a **Equação 45**.

$$\bar{d}(\mathbf{Y}_{test}, C_p) = \frac{1}{N_p} \sum_{k \in I_p} d(\mathbf{Y}_{\text{test}}, \mathbf{Y}_k) \quad (45)$$

Por fim, a amostra de teste será atribuída à classe p* que corresponda ao menor valor de distancia médio, a partir da condição dada na **Equação 46**.

$$\bar{d}(\mathbf{Y}_{test}, C_{p^*}) = \min_{p=1,2,\ldots,L} \bar{d}(\mathbf{Y}_{test}, C_p) \quad (46)$$

CAPÍTULO 4

EXPERIMENTAL

4. EXPERIMENTAL

4.1Conjuntos de dados

4.1.1 Conjunto de dados simulados

O uso de dados simulados pode auxiliar na interpretação das etapas dos algoritmos em condições ajustadas matematicamente pelo programador. Neste trabalho, dados de espectroscopia de fluorescência em matriz de excitação/emissão (EEM) foram simulados a partir do produto de gaussianas, as quais estão teoricamente associadas aos perfis de excitação e emissão de quatro fatores (analitos), que combinados, levaram à formação das diferentes classes. Dois bancos de dados foram gerados: Banco de dados simulado I e Banco de dados Simulado II, de forma que a diferença entre esses está no grau de sobreposição dos perfis dos analitos.

4.1.1.1 Dados simulados I

a) Os perfis foram gerados a partir de comprimentos de onda fictícios que variaram nas seguintes faixas: de 300-398 nm para a excitação (**Figura14a**) e 410-508 nm para emissão (**Figura14b**). Ambos os modos com um intervalo de 2 nm. Como resultado, foi obtida uma matriz de dados de EEM com m = 50 linhas e n = 50 colunas para cada um dos fatores (analitos).

b) As matrizes de amostras foram geradas como a combinação linear de até quatro fatores simulados (A1, A2, A3, A4, que estão representados nas **Figuras 14c**, **14d**, **14e** e**14f**, respectivamente);

c) Três classes foram definidas através da variação dos fatores utilizados na geração das amostras: *Classe 1* (fatores A1, A2, A3), *Classe 2* (fatores de A2, A3, A4) e *Classe 3* (fatores

A1, A2, A3, A4). Exemplos de amostras em cada uma destas classes são apresentados na **Figura 15**.

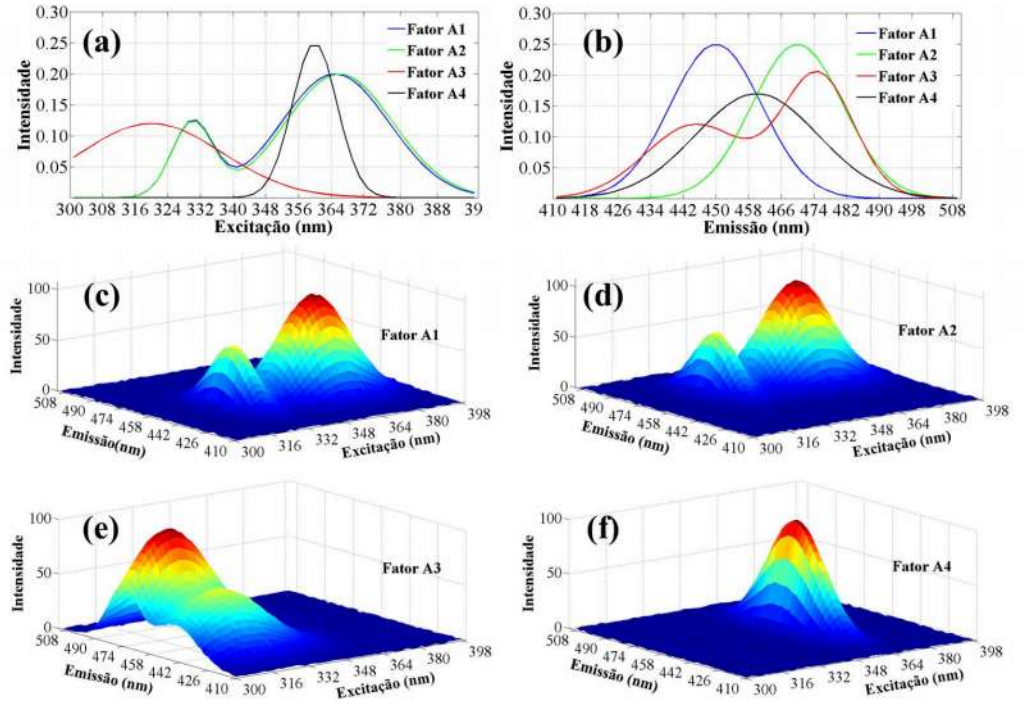


Figura 14 - Fatores empregados na geração do conjunto de dados EEM simulado I. (A) Perfis de excitação, (b) perfis de emissão, (c) fator A1, (d) fator A2, (e) fator A3, (f) fator A4.

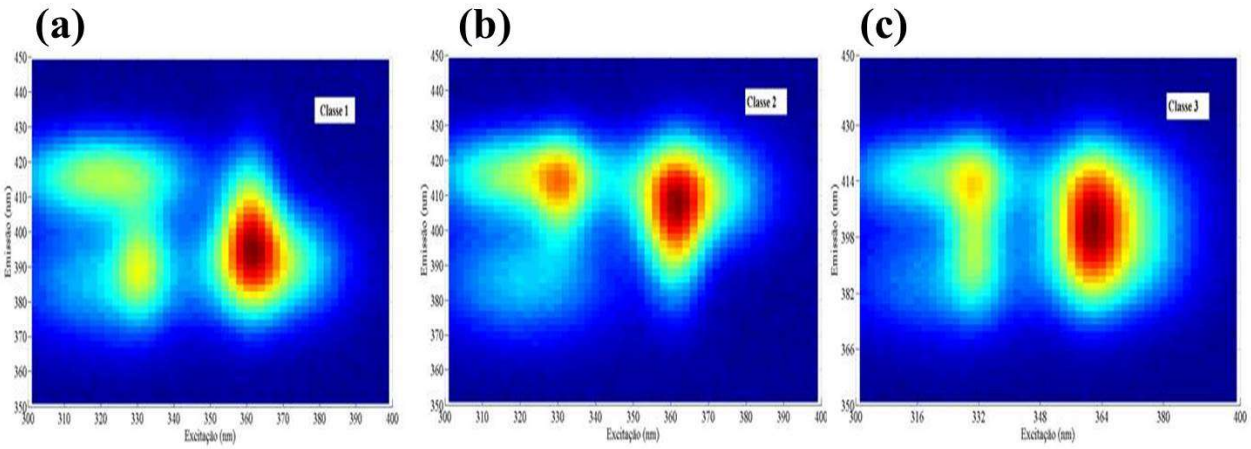


Figura 15–Perfis das classes dos dados de EEM simulado I. (a) Classe 1, (b) Classe 2 e (c) Classe 3.

d) A variabilidade nas classes foi acrescentada através de modificações nos coeficientes da combinação linear utilizados para a geração das amostras (**Tabela 1**);

e) Um ruído Gaussiano foi adicionado às matrizes de EEM resultantes, como uma intensidade de 0,5% do valor máximo de pico para cada uma das amostras.

Tabela 1–Valores de média e desvio padrão (Std) empregados na geração aleatória dos coeficientes das combinações lineares usadas para a construção dos conjuntos de dados simulados.

Grupo de Treinamento						
Fatores	**Classe1**		**Classe2**		**Classe3**	
	Média	**Std**	**Média**	**Std**	**Média**	**Std**
A1	0,8675	0,0954	-	-	0,8674	0,0765
A2	0,8194	0,0948	0,8352	0,0975	0,8563	0,0893
A3	0,8546	0,0836	0,8664	0,0949	0,8427	0,0889
A4	-	-	0,8833	0,0892	0,8501	0,0658
Grupo de Teste						
Fatores	**Classe1**		**Classe2**		**Classe3**	
	Média	**Std**	**Média**	**Std**	**Média**	**Std**
A1	0,8893	0,0661	-	-	0,8744	0,0771
A2	0,9342	0,05396	0,8870	0,0636	0,8782	0,0481
A3	0,9147	0,0650	0,8914	0,0536	0,9001	0,0555
A4	-	-	0,9134	0,0446	0,9066	0,0576

*Std= Desvio Padrão.

4.1.1.2 Dados simulados II

Os passos adotados para o banco de dados simulados I foram aplicados na geração do banco de dados simulados II, modificando o grau de sobreposição dos analitos. Para tanto, novos fatores foram obtidos, dando origem aos perfis: Excitação (**Figura 16a**), Emissão (**Figura 16b**), fatores (**Figura 16c**, **16d**, **16e** e **16f**) e classes (Figura**17a**, **17b** e **17c**).

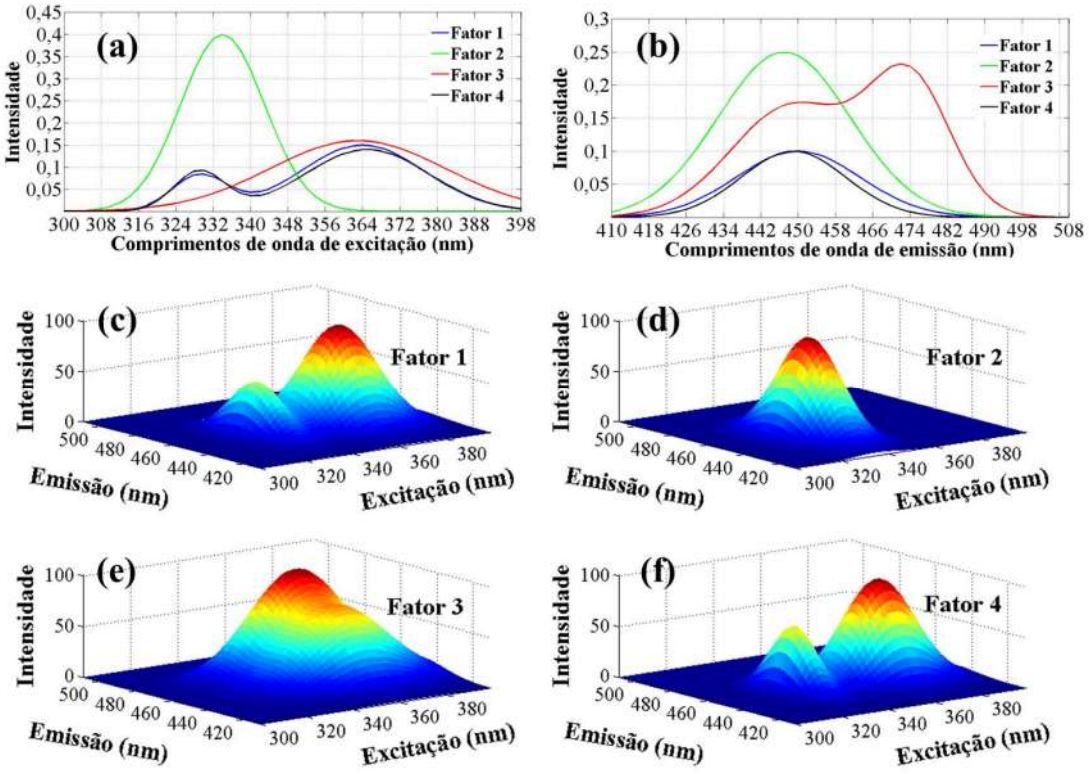


Figura 16 - Fatores empregados na geração do conjunto de dados EEM simulado II. (A) Perfis de excitação, (b) perfis de emissão, (c) fator A1, (d) fator A2, (e) fator A3, (f) fator A4.

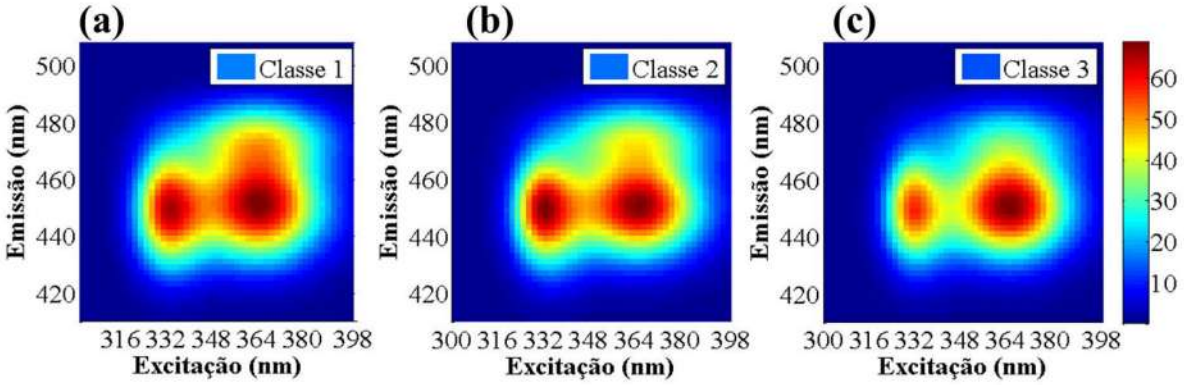


Figura 17 –Perfis das classes dos dados de EEM simulado II. (a) Classe 1, (b) Classe 2 e (c) Classe 3.

Os conjuntos de dados simulados foram divididos aleatoriamente em: conjunto de treinamento e conjunto de teste, como indicado na **Tabela 2**.

Tabela 2 – Divisão das amostras do banco de dados de EEM simulado I e II em grupo de treinamento e teste

Banco de dados	Classe	Grupo de Treinamento	Grupo de Teste	Total
Dados de EEM simulado I e II	Classe 1	20	10	30
	Classe 2	20	10	30
	Classe 3	20	10	30
	Total	60	30	90

4.1.2 Dados de Presunto de Parma curado a seco

O presunto de Parma é um produto alimentar oriundo da perna do porco que é produzido tradicionalmente na cidade de Parma (Itália), e apresenta denominação de origem protegida. A cura consiste na salga de uma carne com o objetivo inicial de conservação e (ou) intensificação de textura e sabor, dando características únicas ao produto [61]. O presunto de

Parma, especificamente, desenvolve um sabor distintivo e aroma após 12 meses de maturação [62-63].

O banco de dados adotado nesse estudo refere-se ao uso da espectroscopia de autofluorescência de superfície como uma alternativa aos métodos tradicionais para a avaliação de parâmetros de qualidade relacionados ao envelhecimento do presunto de Parma. O conjunto de dados foi disponibilizado por Moller et al. [64]em: www.models.life.ku.dk/datasets. Este conjunto foi também empregado por Durante et al. [8] para classificação de amostras de acordo com o estado de envelhecimento utilizando a ferramenta quimiométrica N-SIMCA.

O conjunto de dados compreende um total de 67 amostras, com os espectros de EEM registrados em instrumento BioView (Delta Light e Optics, Lyngby, Dinamarca) equipado com uma sonda de fibra óptica. No presente trabalho, o número de variáveis e a divisão das amostras em classes seguiram estrutura semelhante à apresentada no trabalho de Durante et al. [8]. Assim, a matriz de dados para cada amostra consiste em m = 13 comprimentos de onda de emissão na gama de 350 - 590 nm, e n = 11 comprimentos de onda de excitação na gama de 270 - 470 nm. Quatro classes foram definidas com base no período de envelhecimento da amostra de presunto de Parma: carne crua, salgada (3 meses), maturada (11-12 meses) e envelhecida (15-18 meses).

Na **Figura 18** são apresentados espectros EEM representativas de cada classe. A divisão das amostras em conjuntos de treinamento e teste é apresentada na **Tabela 3**.

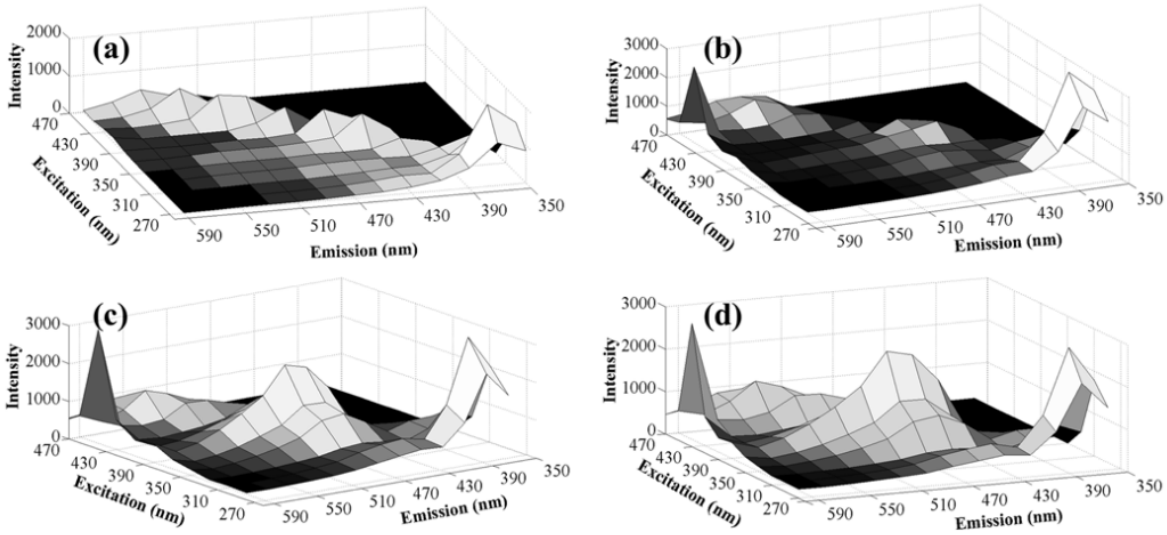


Figura 18–Espectros de autofluorescência de superfície para amostras de presunto de Parma maturadas a seco referentes a seguintes classes: (**a**) crua, (**b**) salgada, (**c**) maturada, (**d**) envelhecida.

Tabela 3 – Divisão das amostras do banco de dados de presunto de Parma curado a seco, em dois grupos: treinamento e teste.

Banco de dados	Classe	Grupo de Treinamento	Grupo de Teste	Total
Presunto de Parma curada a seco	Crua	4	2	6
	Salgada	9	5	14
	Maturada	17	7	24
	Envelhecida	16	7	23
	Total	46	21	67

4.1.3 Dados de óleo vegetal comestível

Óleos vegetais comestíveis, extraídos de sementes das plantas,apresentam propriedades nutricionais que variam de acordo com a matéria-prima, processamento, armazenamento e outros fatores [65-66]. Diante da similaridade entre o óleo oriundo de diferentes oleaginosas, a identificação de matéria-prima pode ser uma tarefa desafiadora. Neste sentido, são propostas

na literatura diferentes metodologias analíticas que utilizam quimiometria e técnicas instrumentais como: fluorimetria [67], nariz eletrônico [68], infravermelho médio (mid-IR) [69] e voltametria de onda quadrada [70].

A fluorescência sincrônica total é uma técnica altamente sensível, que pode ser utilizada como uma alternativa à fluorescência molecular padrão para evitar a superposição de bandas de excitação e emissão, eliminando efeitos de espalhamento [71]. Para tanto, os monocromadores de emissão e excitação são acionados simultaneamente, com uma diferença de comprimento de onda constante ($\Delta\lambda = \lambda emissão - \lambda excitação$) [72].

Dentre as aplicações da fluorescência sincrônica relatadas na literatura estão à avaliação de adulterações no azeite virgem [71], discriminação entre azeite virgem comestível e lampante [72], a classificação de óleos comestíveis em soluções de n-hexano [73] e classificação de amostras de biodiesel em relação a matéria-prima [74].

O conjunto de dados compreende 49 amostras de óleo comestível de três tipos de matéria-prima: soja (13 amostras), milho (20 amostras) e girassol (16 amostras). Os dados espectrais foram adquiridos com um espectrofluorímetro Aminco Bowman Series 2 controlado por computador e equipado com uma fonte de luz de descarga de xenon (150 W) (**Figura 19**). As medições foram realizadas com uma taxa de varredura de 5 $nm.s^{-1}$, com precisão e repetitividade de ± 0,5 nm e ± 0,25 nm, respectivamente. Para cada amostra utilizou-se um volume de 600 uL e obtiveram-se oito espectros síncronos movendo os monocromadores de emissão e de excitação com diferenças constantes de comprimento de onda ($\Delta\lambda$) de 10, 15, 20, 25, 30, 35, 40 e 45 nm. O intervalo de excitação foi o mesmo para todos os espectros (280-430 nm), enquanto a faixa de emissão variou de 290-440 nm a 325-475 nm, de acordo com a diferença de comprimento de onda ($\Delta\lambda$) empregada. Como resultado, os dados espectrais para cada amostra apresentaram o formato em uma matriz com

m = 150 linhas correspondentes aos comprimentos de onda de excitação (λexcitação) e n = 8 colunas correspondentes às diferenças de comprimento de onda (Δλ). Na **Figura 20** são apresentados os espectros para cada classe das amostras.

Figura 19–espectrofluorímetro Aminco Bowman Series 2 utilizado na aquisição dos espectros de fluorescência sincrônica das amostras de óleo vegetal comestível.

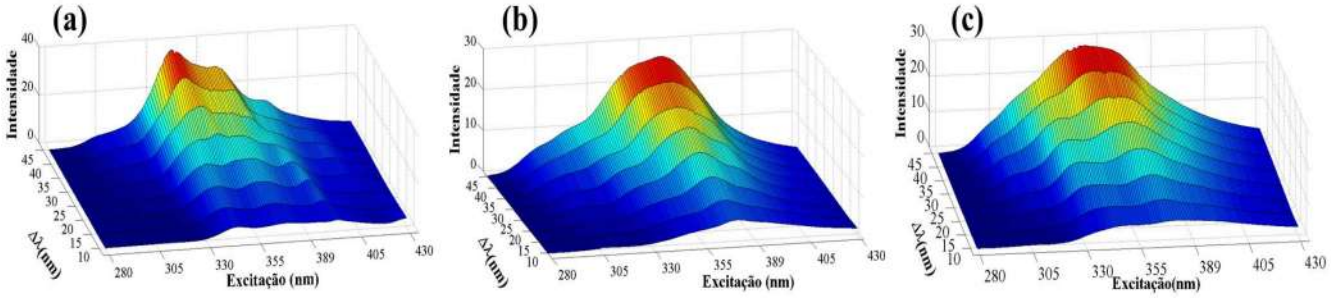


Figura 20–Espectros de fluorescência sincrônica amostras de óleo vegetal comestível referentes a seguintes classes: (**a**) soja, (**b**) milho e (**c**) girassol.

Na **Tabela 4,** é possível visualizar a divisão do conjunto de amostras em: grupo de treinamento e grupo de teste.

Tabela 4 – Divisão das amostras do banco de óleos vegetais comestíveis, em dois grupos: treinamento e teste.

Banco de dados	Classe	Grupo de Treinamento	Grupo de Teste	Total
Óleos comestíves	Soja	10	3	13
	Milho	14	6	20
	Girassol	12	4	16
	Total	36	13	49

4.2 Programas

4.2.1 2D-LDA

O algoritmo 2D-LDA foi implementado no Matlab®2010b (Mathworks) a partir de scripts e rotinas desenvolvidas no laboratório. Para escolha do número ótimo de vetores de projeção (r) que devem ser usados na modelagem (**Seção 3.2**), valores de taxa de classificação correta (TCC) para validação cruzada "leave-one-out" foram avaliados. Por razões de parcimônia, caso um mesmo valor de TCC seja obtido para diferentes valores de r, o menor valor de r será selecionado. Após a determinação dos vetores de projeção, os vetores de características das amostras de treinamento foram avaliados. Em seguida, realizou-se a classificação das amostras pertencentes ao grupo de Teste (**Seção 3.3**).

4.2.2 No feature extraction ("NFE")

Com objetivo de avaliar a influência do procedimento de extração de vetores de características com 2D-LDA, o procedimento de classificação descrito na **Seção 3.3**foi aplicado usando apenas os dados originais desdobrados no modo ***I*** x ***JK*** (**Seção 2.2.1**), no lugar de usar as matrizes de características que são primordiais no algoritmo 2D-LDA. Para

esse propósito a modificação foi realizada no calculo das **Equações 43**, **44** e **45**, onde o Xtest desdobrado e Xk desdobrado foram usados, em vez de Ytest e Yk, respectivamente.Para essa estratégia foi dado o nome de: NFE (*no feature extraction*), ou seja, será utilizada classificação por distancia euclidiana sem extração de vetores de características.

Adicionalmente, outros algoritmos que são propostos na literatura, foram utilizados na comparação de resultados: U-PLS-DA, PARAFAC-LDA e TUCKER-3-LDA.

4.2.3 U-PLS-DA

A operação de desdobramento empregada em U-PLS-DA foi aplicada de acordo com o apresentado na **Seção 2.2.1**. Em seguida, é realiza a classificação a partir do uso do algoritmo PLS-DA (**Seção 2.3.2**) para dados de primeira ordem.

A escolha do número de variáveis latentes do U-PLS-DA foi realizada de forma a maximizar o valor TCC obtido por validação cruzada. O critério de parcimônia para número de variáveis também foi adotado. Neste caso, quando o valor TCC máximo foi obtido com diferentes números de variáveis latentes (VL), o menor valorde VL foi selecionado.

Os cálculos de PLS-DA foram realizados em Matlab®2010b (Mathworks) a partir do uso do "Classification Toolbox 3.1", disponível para download online no endereço eletrônico:http://michem.disat.unimib.it/chm/download/softwares.htm. Mais detalhes sobre o toolbox podem ser encontrados na referencia [58].

4.2.4 PARAFAC-LDA e TUCKER-3-LDA

Para classificação utilizando os algoritmos PARAFAC-LDA e TUCKER-3-LDA, valores de escores obtidos a partir da decomposição dos dados de segunda ordem foram utilizados como dados de entrada do algoritmo de classificação LDA.

Na execução dos algoritmos PARAFAC e TUCKER-3 foram aplicadas as restrições de não negatividade e ortogonalidade. Dessa maneira, nenhum dos perfis dos fatores encontrados no PARAFAC apresentará valores negativos e os fatores obtidos com TUCKER-3 não terão informações redundantes(não correlacionados entre si) [38].

No estudo de decomposição utilizando PARAFAC-LDA, três critérios foram avaliados: maximização do valor TCC para validação cruzada, erro de modelagem e valor de CORCONDIA. No caso do TUCKER-3, foram avaliados: Maximização do valor TCC para validação cruzada com número de fatores iguais nos 3 modos dos dados. Também foram avaliados usando busca exaustiva de combinação de fatores. Para comparação com o 2D-LDA e U-PLS-DA, foram utilizados os resultados obtidos com o critério de maximização da TCC para validação cruzada.

Os cálculos de TUCKER-3-LDA e PARAFAC-LDA foram realizados em Matlab®2010b (Mathworks) usando a caixa de ferramentas "N-way toolbox v. 3.30", disponível online para download no seguinte endereço eletrônico:http://www.models.life.ku.dk/algorithms. O algoritmo LDA foi desenvolvido no laboratório.

CAPÍTULO 5

RESULTADOS E DISCUSSÃO

5. Resultados e discussão

5.1 Conjuntos de dados simulados de EEM

5.1.1 2D-LDA

Para usar o algoritmo 2D-LDA, os conjuntos de treinamento foram inicialmente empregados para calcular as matrizes de espalhamento SB e SW, como apresentado nas **Equações 37** e **38**. Em seguida, foram obtidos os vetores de projeção b1, b2, ..., b50 para cada um dos conjuntos de dados simulados. A determinação do número r de vetores de projeção a serem utilizados na construção do modelo de classificação foi dada a partir dos valores de taxa de classificação correta obtida para validação cruzada nos conjuntos de treinamentos, como apresentado na **Figura 21a** (Dados simulados I) e **Figura 21b** (Dados simulados II).

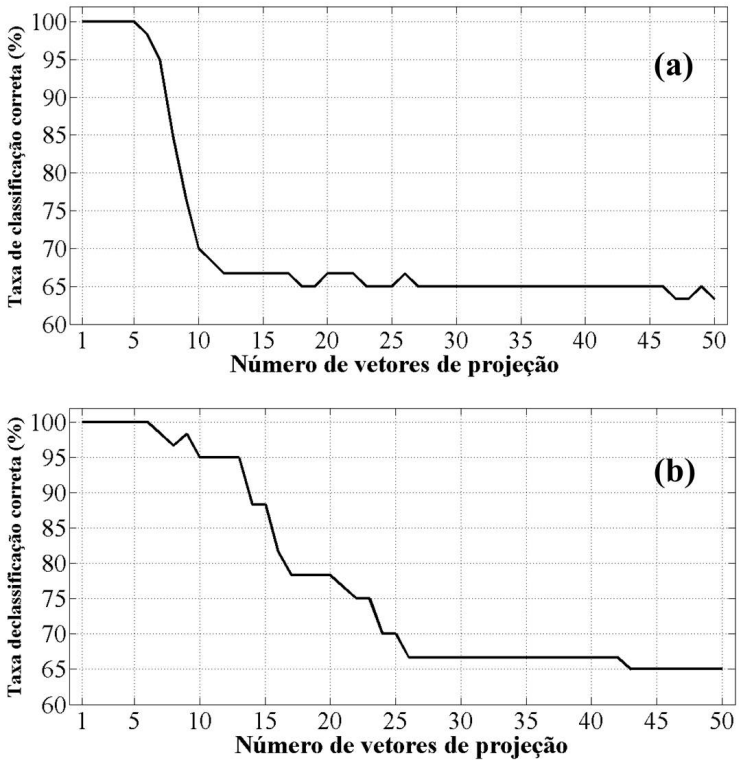


Figura 21 – Taxa de classificação correta por validação cruzada versos número de vetores de projeção para os bancos de dados de EEM simulados (**a**) I e (**b**) II.

Como apresentado na **Figura 21**, considerando o critério de parcimônia, um r equivalente a um vetor de projeção é suficiente para obtenção de taxa de classificação correta de 100% para ambos os bancos de dados simulados.

Na **Figura 22** são apresentados os vetores de características das amostras de treinamento para os dados simulados I (**Figura 22a**) e II (**Figura 22b**), obtidos a partir do produto entre as matrizes individuais das amostras e o vetor de projeção selecionado.

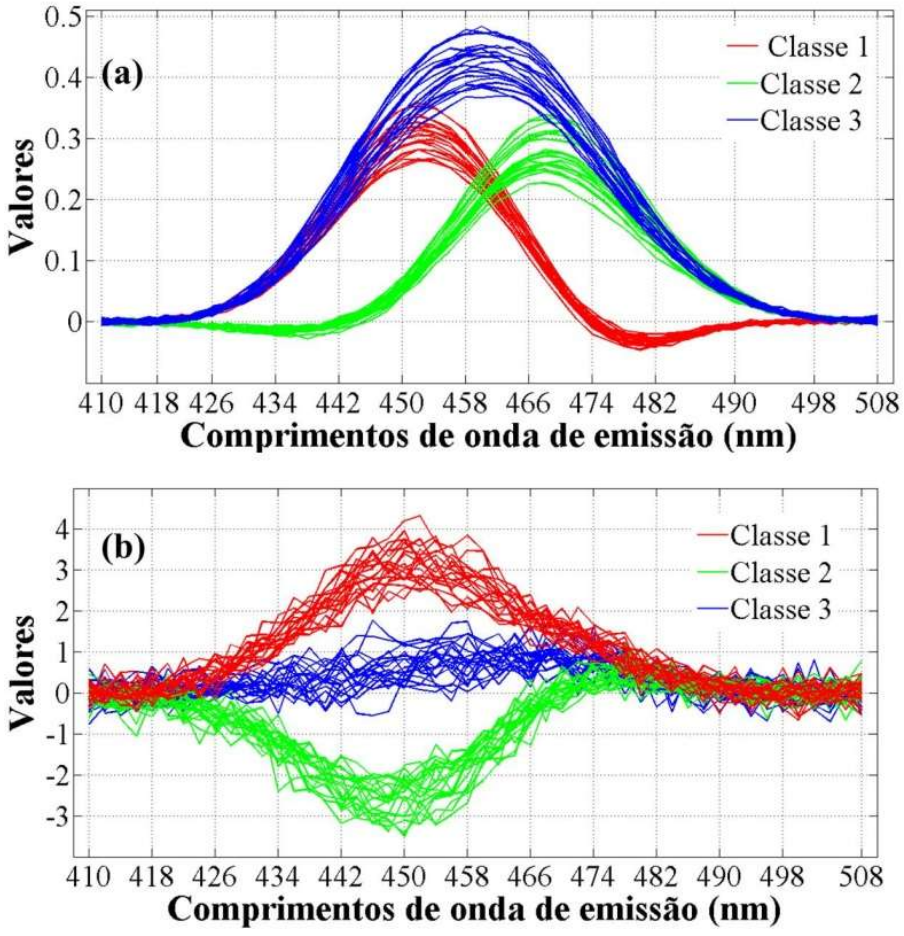


Figura 22 – Vetores de características 2D-LDA para dados EEM simulados I (**a**) e (**b**) II.

Como pode ser observado na **Figura 22**, existe uma clara diferença entre os vetores de características para as amostras de treinamento das classes 1,2 e 3 para os dois conjuntos de dados simulados. No caso dos dados simulados I (**Figura 22a)**, é possível verificar que os perfis encontrados são semelhantes, respectivamente, aos fatores simulados A1, A2 e A4 (**Figura 14**). No caso do banco de dados simulados II não foi possível fazer nenhuma associação com os perfis de emissão simulados apresentados na **Figura 14**. No entanto, para esse banco de dados, verifica-se que os perfis dos vetores de características para as amostras de treinamento apresentam a região discriminante em torno de 426 a 474 nm, que é também a região onde ocorre a emissão dos fatores A1 e A4 (**Figura 16)**, que fornecem as características das classes. Esse resultado nos auxilia a inferir que, para dados com maiores sobreposições, ainda é possível obter informações sobre qual região espectral contém informações que contribuem para discriminação das amostras.

Após a obtenção dos vetores de características para as amostras de treinamento dos dados simulados, foi realizada a classificação das amostras de teste. Uma taxa de classificação correta de 100% foi alcançada para os dois modelos 2D-LDA.

5.1.2 PARAFAC-LDA

A escolha do número de fatores para construção dos modelos PARAFAC-LDA dos dados simulados foi feita a partir da taxa de classificação correta da validação cruzada para dados simulados I(**Figura 23a**)e II (**Figura 23b**). Ainda, os valores de corcondia para I (**Figura 23c**) e II (**Figura 23d**) e soma do quadrado dos erros dos modelos I (**Figura 23e**) e II (**Figura 23f**) com diferentes números de fatores foram avaliados.

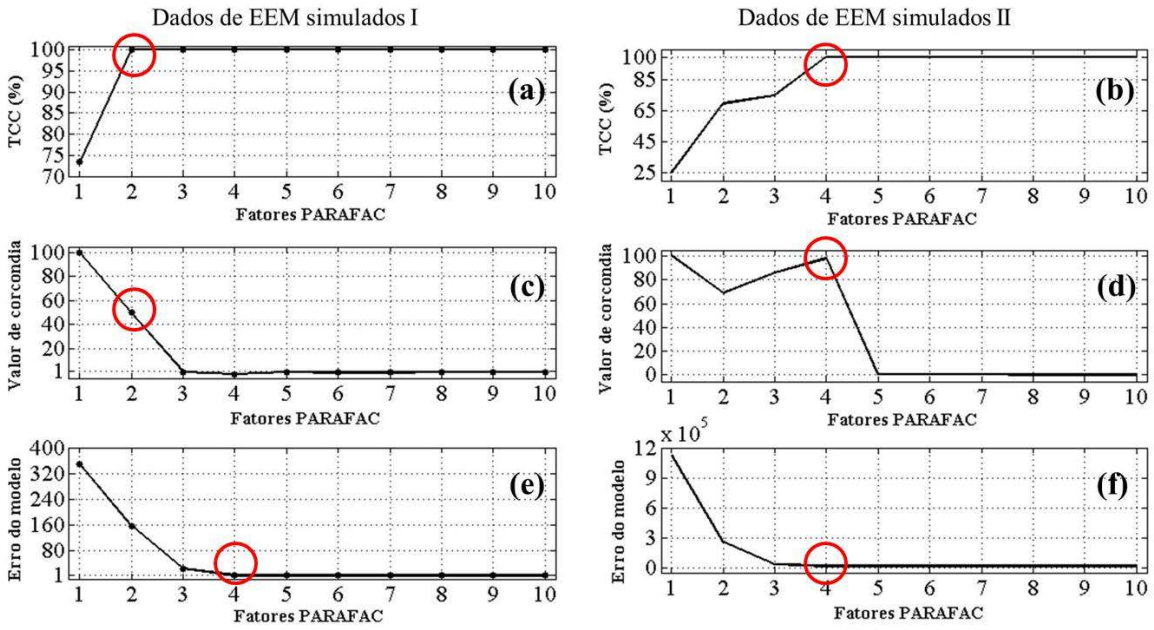


Figura 23 – Taxa de classificação correta por validação cruzada versos número de fatores PARAFAC(Dados simulados I (**a**) e II (**b**)), valor de corcondia (Dados simulados I (**c**) e II (**d**)) e de erro de modelo (Dados simulados I (**e**) e II (**f**)) versos número de fatores PARAFAC.

O modelo de classificação, para o conjunto de dados simulados I, apresenta uma maior parcimônia com um total de 2 fatores, como pode ser visto a partir dos valores de taxa de classificação correta da validação cruzada apresentado na **Figura 23a**. Da mesma maneira, os valores de concordância de núcleo (**Figura 23b**), com valor de 49,8 de corcondia, indicam que até dois fatores existe uma tendência a trilinearidade dos dados. Esse resultado se alinha com o proposto na simulação, pois apenas dois analitos (fatores simulados) estão presentes em todas as amostras.Logo, acima dessa quantidade de fatores ocorre uma quebra da trilinearidade por não haver uniformidade da resposta entre as amostras. No entanto, o perfil recuperado (**Figura 24a,b**) usando o modelo PARAFAC com apenas dois fatores, apresentado a seguir, não é condizente com o perfil proposto na simulação (**Figura 14b**). Espera-seque um melhor resultado em termos de perfis recuperados seja dada pelo modelo

com um menor erro de modelagem. De acordo com a **Figura 23e**, o menor valor de erro de modelagem é encontrado com um número de fatores igual a 4.Avaliando as **Figura 24c** ***e*** **24d**,é possível verificar uma melhor correlação dos perfis recuperados com os propostos na simulação (**Figura 12b**).

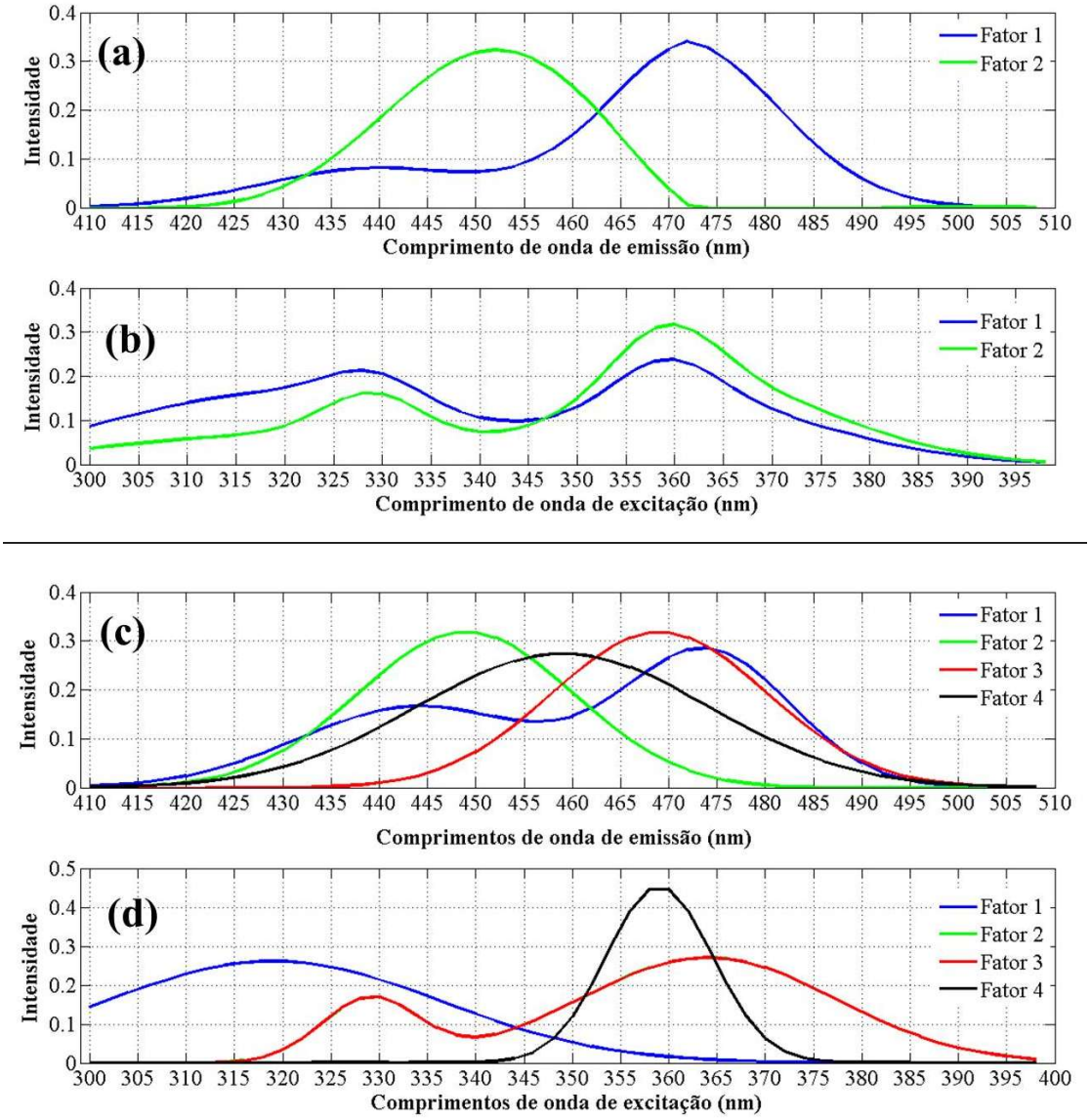


Figura 24 –Perfis recuperados com PARAFAC para os dados simulados I. Usando 2 fatores (Emissão (a) e Excitação (b)) e 4 fatores (Emissão (c) e Excitação (d)).

Apesar da melhor correlação de perfis de emissão e excitação ser obtida quando são utilizados 4 fatores no modelo PARAFAC, apenas 2 fatores foram selecionados. A escolha foi realizada levando em consideração que o modelo de classificação mais parcimonioso que retém informação discriminante, com 100% de taxa de classificação correta da validação cruzada.

Para construção do modelo PARAFAC do banco de dados simulados II, é possível verificar que o uso de 4 fatores permite uma maior taxa de classificação correta da validação cruzada (**Figura 23b**), com boa trilinearidade (corcondia de 98,15 (**Figura 23d**)) e um baixo valor de erro de modelagem (**Figura 23f**). Apesar dos ótimos resultados obtidos, os resultados podem levar a conclusões incorretas, tendo em vista que o número de analitos varia entre as amostras. A "falsa" trilinearidade dos dados pode ser justificada pelo fato dos fatores 1e 4 (**Figura 14**) apresentarem um alto nível de sobreposição, permitindo que os seus perfis sejam previstos em todas as amostras, mesmo que um deles não esteja presente. Avaliando a **Figura 25**, é possível verificar um alto grau de correlação dos perfis recuperados com os propostos na simulação (**Figura 14**).

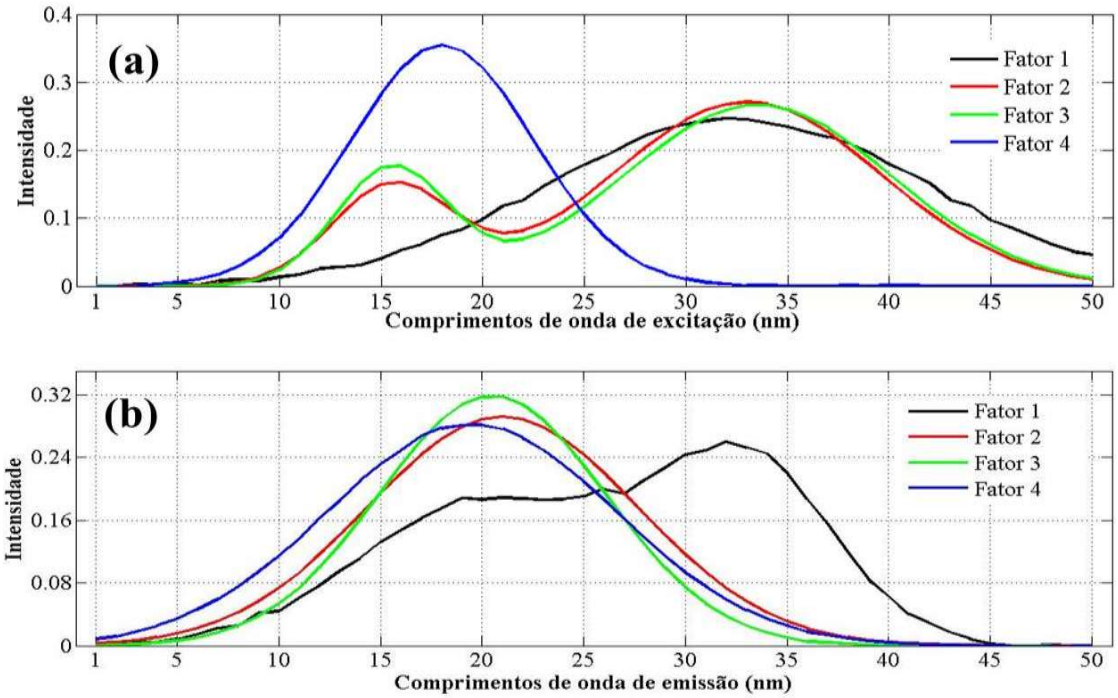


Figura25–Perfis recuperados com PARAFAC para os dados simulados I. Usando 2 fatores (Emissão (a) e Excitação (b)) e 4 fatores (Emissão (c) e Excitação (d)).

Após determinar o número de fatores que serão utilizados nos modelos de classificação PARAFAC-LDA, a determinação das classes das amostras de teste foi realizada. Taxas de classificação correta de 100% foram alcançadas para as amostras de teste dos bancos de dados simulados I e II, de acordo com o apresentado nas matrizes de confusão a seguir.

Os modelos PARAFAC-LDA, de forma geral, conseguem prever com 100% de acerto as amostras de teste.

5.1.3 TUCKER-3-LDA

A escolha do número de fatores para construção dos modelos TUCKER-3-LDA para os dados simulados foi dada a partir da avaliação da taxa de classificação correta para dados quando: número de fatores iguais para todos os modos da matriz (simulados I (**Figura 26a**) e

II (**Figura 26b**)) e a partir de diferentes combinações realizadas com busca exaustiva (simulados I (**Figura 26c**) e II (**Figura 26d**)).

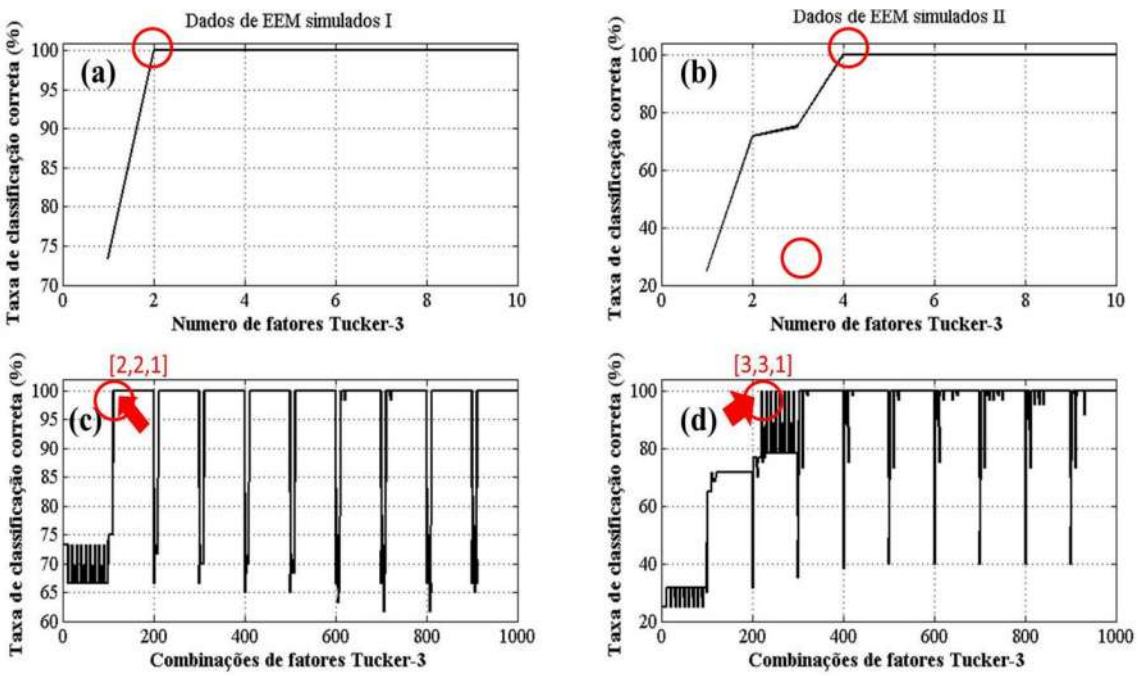


Figura 26 –Taxa de classificação correta por validação cruzada versos: número de Fatores Tucker-3 (Dados simulados I (**a**) e II (**b**) e diferentes combinações de fatores Tucker-3 (Dados simulados I (**c**) e II (**d**))

O modelo de classificação TUCKER-3-LDA, para o banco de dados simulado I, apresenta uma maior parcimônia com um total de 2 fatores, com100% de taxa de classificação correta da validação cruzada (**Figura 26a**). Ainda, é possível verificar que os estudos com busca exaustiva (**Figura 26c**) e com PARAFAC (**Figura 23a**) corroboram com esse resultado, pois ambos necessitaram dois fatores na matriz de escores para obter a mesma taxa de classificação. No entanto, é válido salientar que a busca exaustiva usa número de fatores diferentes em cada uma das matrizes geradas na decomposição (2 fatores para os escores, 2 fatores para o modo de excitação e apenas 1 para emissão). Logo, a correlação entre os escores e os perfis obtidos torna-se inviável, reduzindo assim a possibilidade de interpretação

química a partir dos seus resultados. Outro fator que inibe a interpretação química nos modelos TUCKER-3-LDA, é o fato de não ter sido aplicado a restrição de não negatividade, permitindo assim perfis recuperados com valores negativos, os quais não estão presentes nos dados originais.

Para o banco de dados simulados II, é possível verificar que o uso de 4 fatores permite 100% de acerto na validação cruzada (**Figura 26b**). O resultado encontrado é semelhante ao obtido usando PARAFAC (**Figura 23b**). Quando realizada a busca exaustiva de fatores (**Figura 26d**), verificou-se que um total de 3 fatores para a matriz de escores foram suficientes para gerar um modelo com 100% de acerto. Apesar da menor quantidade de fatores, a correlação entre as matrizes recuperadas é perdida. Logo, a interpretação química dos resultados é comprometida, invalidando os perfis obtidos com Tucker-3. Neste caso,penas a informação discriminante das classes presente na matriz de escores será utilizada para como entrada no LDA. Como resultado da classificação do grupo de treinamento, 100% das amostras foram corretamente classificadas.

5.1.4 U-PLS-LDA

A escolha do número de variáveis latentes para construção dos modelos U-PLS-DA foi realizada a partir da análise da taxa de classificação correta da validação cruzada versos o número de variáveis latentes para os dados simulados I (**Figura 27a**) e II (**Figura 27b**), como apresentado a seguir.

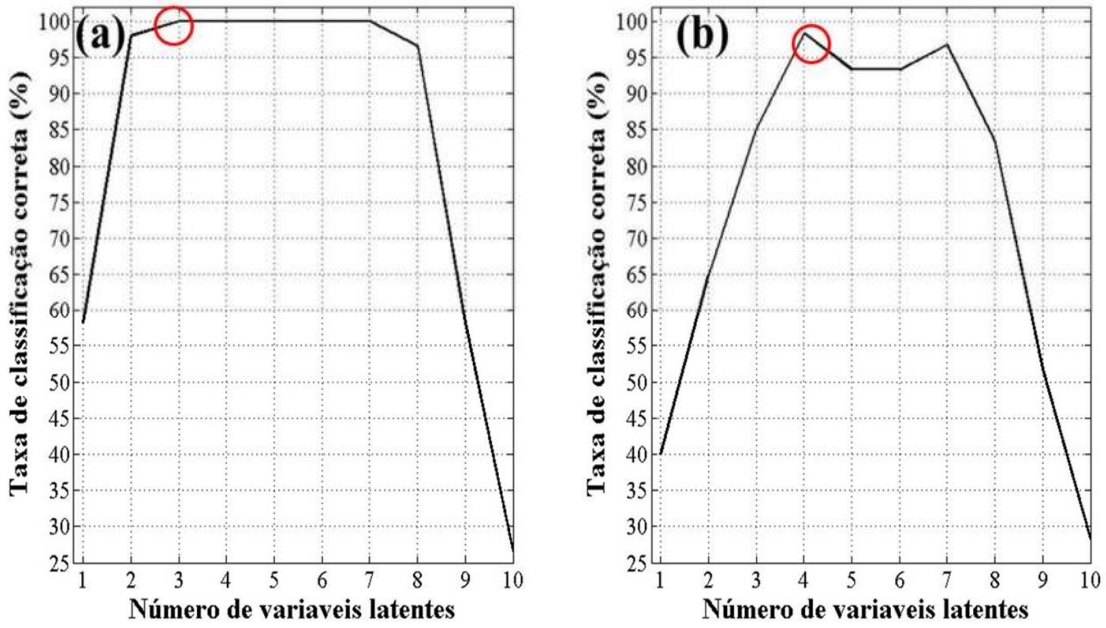


Figura 27 – Taxa de classificação correta por validação cruzada versos número de variáveis latentes para os bancos de dados de EEM simulados I (**a**)e II (**b**).

Avaliando os valores de taxa de classificação correta da **Figura 27**, é possível verificar que os modelos de classificação mais parcimoniosos em termos de número de variáveis latentes de PLS para os dados simulados I e II equivalem, respectivamente, aos resultados com 3 (100% de classificação correta) e 4 (98,3% de classificação correta) variáveis selecionadas. Para o modelo gerado para o banco de dados simulado II, os erros oriundos do uso das 4 variáveis latentes equivale a 1,7% do total de amostra, e está associados a não atribuição de classes (quando a amostra não é classificada ou é atribuída a mais de uma classe). Esse resultado evidencia, de certo modo, o maior grau de sobreposição (semelhança) entre as amostras simuladas no segundo banco de dados.

Após determinar o número de fatores que serão utilizados nos modelos de classificação U-PLS-DA, a determinação das classes das amostras de teste foi realizada. Taxas de classificação correta de 100% foram alcançadas para as amostras de teste dos bancos de dados simulados I e II.

5.1.5 NFE (No feature extraction)

Na **Tabela 5** são apresentadas as classes preditas das amostras de teste pertencentes aos dados simulados I e II. Como pode ser observado, todas as amostras foram corretamente classificadas para os dados simulados I, diferente do resultado obtido para as amostras dos dados simulados II, onde 90% das amostras da classe 2 foram prevista erroneamente como pertencentes a classe 1. Esse resultado pode ser atribuído ao nível de sobreposição simulado em cada banco de dados. Para uma melhor compreensão desse efeito, são disponibilizados na **Figura 28** os perfis médios das matrizes desdobradas para as amostras pertencentes às classes de treinamento para os bancos de dados simulados I (**Figura 28a**) e II (**Figura 28b**).

Tabela 5 - Resultado de classificação usando NFE para o grupo de Teste (conjuntos de dados Simulados). Valores em percentual (%).

NFE		**Classe predita**		
Conjuntos de dados	**Classe**	**Classe 1**	**Classe 2**	**Classe 3**
Simulado I	Classe 1	100%	-	-
	Classe 2	-	100%	-
	Classe 3	-	-	100%
Simulado II	Classe 1	100%	-	-
	Classe 2	90%	10%	-
	Classe 3	-	-	100%

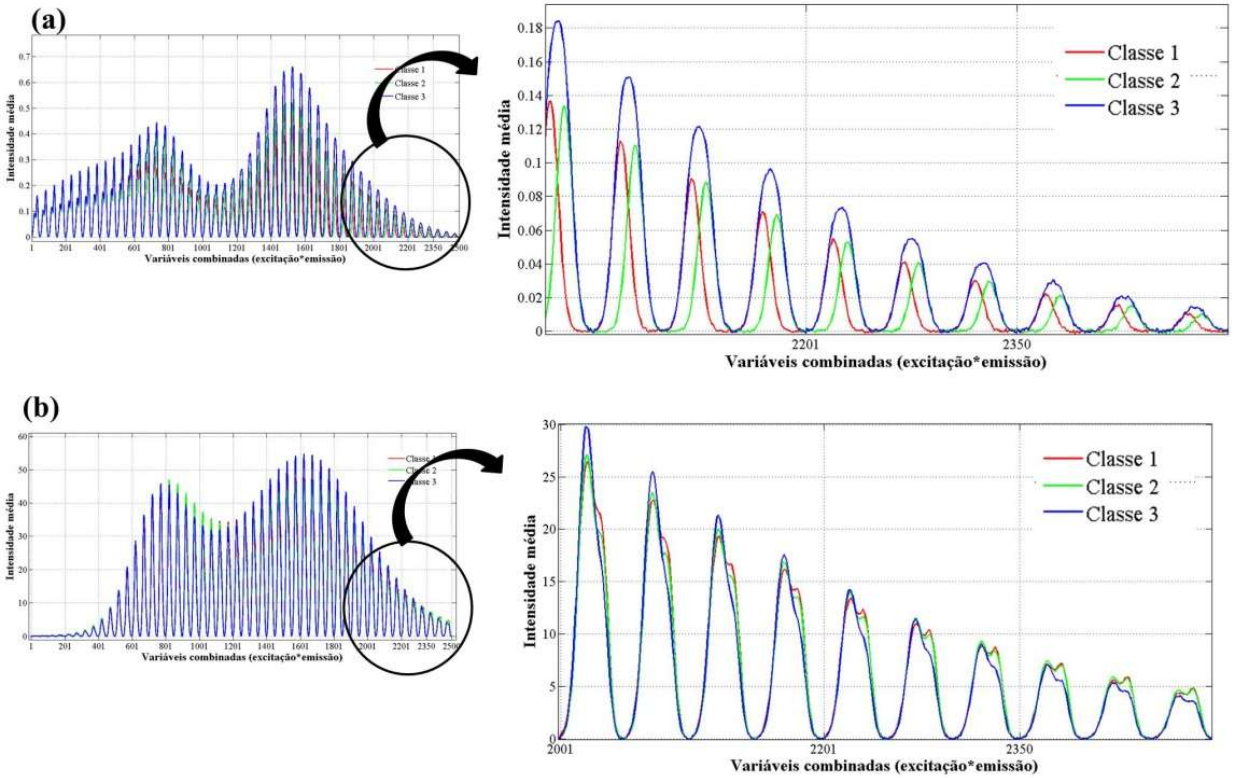


Figura 28 –Perfil de desdobramento dos bancos de dados de EEM simulados I (**a**) e II (**b**).

Observando a **Figura 26**, é possível verificar que para o banco de dados simulados I, os perfis médios das classes apresentam um formato semelhante, porém, se diferenciam principalmente por deslocamento e intensidades. No entanto, no caso dos perfis das amostras do banco de dados simulado II existe uma clara sobreposição das classes em termos de posição das bandas. Além disso, os perfis das classes 1 e 2 são mais semelhantes entre si do que quando comparados com a classe três. Esse resultado corrobora para justificativa dos erros de classificação das amostras de Teste para o banco de dados simulado II.

Na **Tabela 6**, apresentada a seguir, são disponibilizados os valores de taxa de classificação correta (%) das amostras de teste dos conjuntos de dados simulados I e II, quando aplicados os modelos de classificação obtidos com: 2D-LDA, PARAFAC-LDA,

TUCKER-3-LDA, U-PLS-DA e NFE.

Tabela 6- Resultado de classificação do grupo de Teste usando diferentes estratégias de modelagem. (conjuntos de dados Simulados). Valores em percentual (%).

Dados	2D-LDA	PARAFAC-LDA	TUCKER3-LDA	U-PLS-DA	NFE
Simulado I	100 (1)	100 (2)	100 (2)	100 (3)	100
Simulado II	100 (1)	100 (4)	100 (4)	100 (4)	70

*Entre parênteses: numero de vetores de projeção, fatores e variáveis latentes.

Avaliando a **Tabela 11** é possível identificar que praticamente todos os modelos de classificação utilizados para predição das classes das amostras de teste, alcançaram taxa de classificação correta de 100%. A única exceção foi encontrada com o modelo NFE quando aplicado ao conjunto de dados Simulado II, onde apenas 70% de todas as amostras de teste foram corretamente atribuídas à classe verdadeira. Essa menor eficiência apresentada pelo modelo NFE pode ser atribuída principalmente a dois a fatos: o método trabalha apenas com distância sem o uso de extração de características discriminantes (vetores de características, fatores ou variáveis latentes) e o banco de dados Simulado II apresenta maior complexidade em termos de sobreposição dos fatores.

Os resultados obtidos para os dados simulados indicam um excelente potencial do 2D-LDA para classificação de dados de três vias, assim como dos demais algoritmos de classificação avaliados. No entanto, por apresentar um baixo nível de complexidade, os dados simulados foram utilizados apenas para ilustrar as etapas envolvidas nos processos de classificação. As vantagens do 2D-LDA serão melhor evidenciadas em dois estudos de caso envolvendo dados reais de maior complexidade, que serão abordados a seguir.

5.2 Conjuntos de dados de fluorescência para presunto de Parma curado a seco

5.2.1 2D-LDA

Na **Figura 29** são apresentados os valores de taxa de classificação correta obtida para validação cruzada no conjunto de amostras de treinamento.

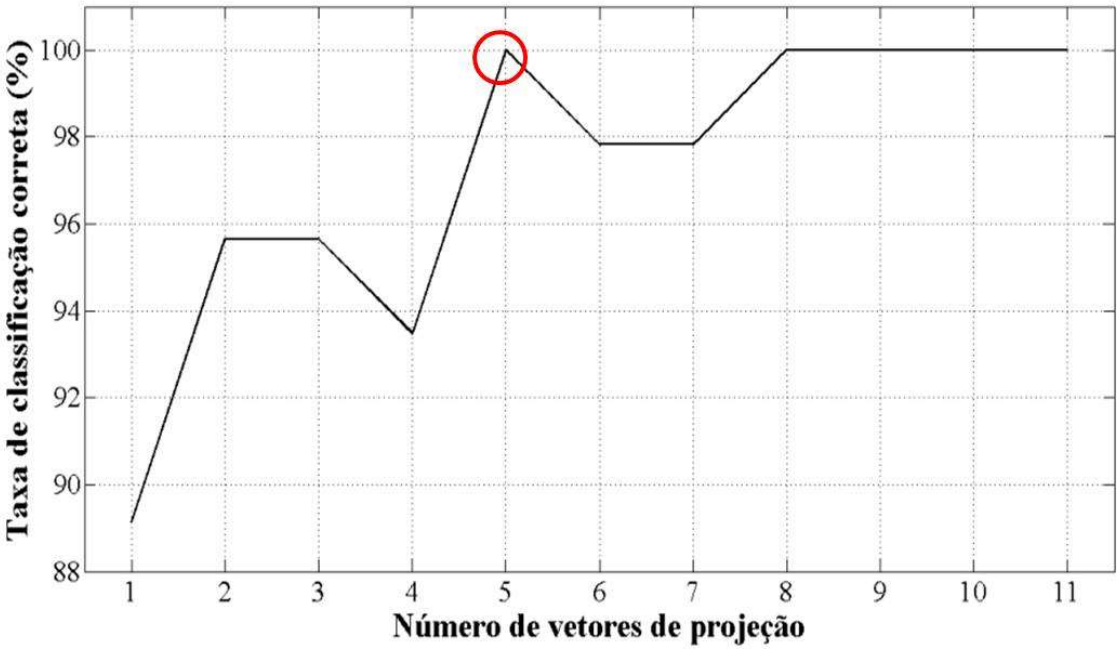


Figura 29 – Taxa de classificação correta por validação cruzada obtida versos número de vetores de projeção para o banco de dados de presunto de Parma curado a seco.

A partir do resultado apresentado na **Figura 29**, considerando o critério de parcimônia para quantidade de variáveis, um r equivalente a 5 vetores de projeção será suficiente para classificação correta de todas as amostras de treinamento no processo de validação cruzada completa. Na **Figura 30**, são apresentados os 5 vetores de características das amostras de treinamento.

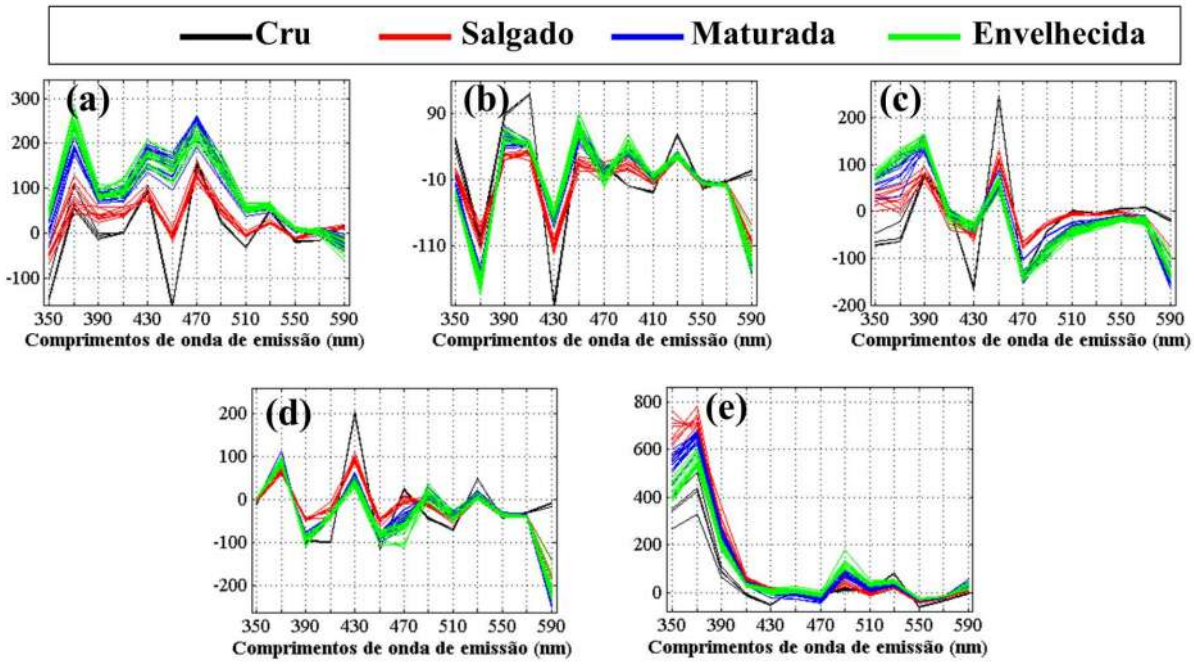


Figura 30 – Vetores de características 2D-LDA das amostras de treinamento do banco e dados de presunto de Parma curado a seco. 1° vetor (**a**), 2° vetor (**b**), 3° vetor (**c**), 4° vetor (**d**)e 5° vetor (**e**).

Como pode ser visto na **Figura 30**, praticamente todos os vetores de características apresentam uma tendência de comportamento com relação às classes de treinamento. As amostras das classes cruas e salgadas, perfis de cor preta e vermelha, se diferenciam entre si e das amostras das classes maturada e envelhecida em comprimentos de onda de emissão em torno de 430 a 470 nm. Esse comportamento pode ser explicado principalmente por influência de fenômenos como: aumento da oxidação de lipídeos, que resultam em outros produtos que passam a modificar o perfil do espectro, e interação de cloretos com compostos orgânicos, podendo ocorrer uma atenuação da emissão "*quenching*" de alguns compostos orgânicos [64]. Para a classe crua esses efeitos não são pronunciados como ocorre na classe de presunto de Parma salgado. Com o decorrer do processo de cura esses efeitos passam a saturar, reduzindo assim as diferenças entre as classes de amostras maturadas e envelhecidas, as quais podem ser

distinguidas, com menor evidencia, a partir do terceiro (**Figura 30c**) e quinto (**Figura 30e**) vetores de características.

As amostras de teste foram classificadas utilizando o procedimento 2D-LDA com os cinco vetores de projeção obtidos no conjunto de treinamento. Os resultados são apresentados na forma de matriz de confusão na **Tabela 7**.

Tabela 7 - Resultado de classificação usando 2D-LDA para o grupo de Teste dos dados de presunto de Parma curado a seco. Valores em percentual (%).

2D-LDA		**Classe predita**			
Conjunto de dados	**Classe real**	Cru	Salgado	Maturado	Envelhecido
Presunto de Parma	Cru	100	-	-	-
	Salgado	-	100	-	-
	Maturado	-	-	86	14
	Envelhecido	-	-	29	71

Todas as amostras cruas e salgadas foram corretamente atribuídas às suas classes verdadeiras, o que é consistente com a clara separação de classes observada nos vetores de característica. Alguns erros de classificação envolvendo amostras maturadas e envelhecidas foram obtidos, o que também é consistente com os resultados do conjunto de treinamento.

5.2.2 PARAFAC-LDA

A escolha do número de fatores para construção dos modelos PARAFAC-LDA foi dada a partir da taxa de classificação correta da validação cruzada (**Figura 31a**), valor de corcondia (**Figura 31b**)e soma do quadrado dos erros do modelo (**Figura 31c**) com diferentes números de fatores avaliados.

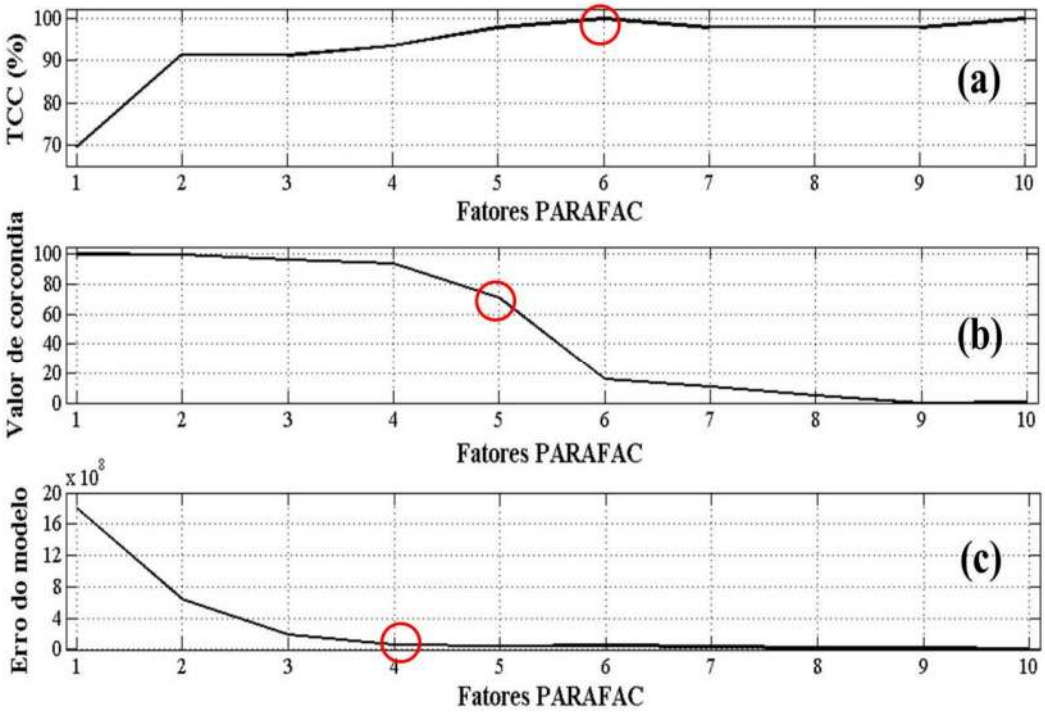


Figura 31 – Taxa de classificação correta por validação cruzada versos número de fatores PARAFAC (**a**), valor de corcondia (**b**) e de erro de modelo (**c**).

De acordo com a **Figura 31a**, o modelo apresenta 100% de taxa de classificação correta da validação cruzada com menor numero de variáveis quando 6 fatores PARAFAC são utilizados. No entanto, apenas 5 fatores são necessários para garantir um menor erro de modelagem (**Figura 31c**) com uma trilinearidade razoável (corcondia de 70,81) (**Figura 31b**). Logo, para que possa ser comparado de forma justa com o 2D-LDA, dois resultados serão avaliados: Acrescentando um fator a mais ao modelo a fim de ampliar a informação

discriminante presente nos escores de PARAFAC, mesmo que esse fator não contribua para trilinearidade dos dados; e com apenas 5 fatores que garantam a modelagem adequada para o PARAFAC, considerando que os dados de fluorescência em matriz de excitação emissão são trilineares. Os resultados de predição para as amostras de teste são apresentados na **Tabela 8**, na forma de matriz de confusão.

Tabela 8- Resultado de classificação usando PARAFAC-LDA com 5 e 6 fatores para o grupo de Teste dos dados de presunto de Parma curado a seco. Valores em percentual **(%).**

PARAFAC-LDA		**Classe predita**			
Conjunto de dados	**Classe real**	Cru	Salgado	Maturado	Envelhecido
Presunto de Parma (5)	Cru	100	-	-	-
	Salgado	-	100	-	-
	Maturado	-	-	71,4	28,6
	Envelhecido	-	-	28,6	71,4
Presunto de Parma (6)	Cru	100	-	-	-
	Salgado	20	80	-	-
	Maturado	-	-	71.4	28.6
	Envelhecido	-	-	28,6	71,4

Avaliando os resultados presentes **Tabela 13** é possível verificar que o modelo com 5 fatores, que levou em consideração a trilinearidade dos dados, foram obtidos os melhores resultados quando comparado com o modelo com 6 fatores (que teve maior taxa de classificação correta da validação cruzada). Esse resultado indica que ao levar em conta apenas a taxa de classificação correta da validação cruzada pode ocorrer sobre ajuste ("*over*

fitting") no modelo, levando a erros da predição, que podem ser evitados ao se considerar a trilinearidade inerente aos dados.

5.2.3 TUCKER-3

A escolha do número de fatores para construção dos modelos TUCKER-3-LDA foi dada mediante a avaliação da taxa de classificação correta para dados quando: número de fatores iguais para todos os modos da matriz (**Figura 32a**) e a partir de diferentes combinações realizadas com busca exaustiva (**Figura 32b**).

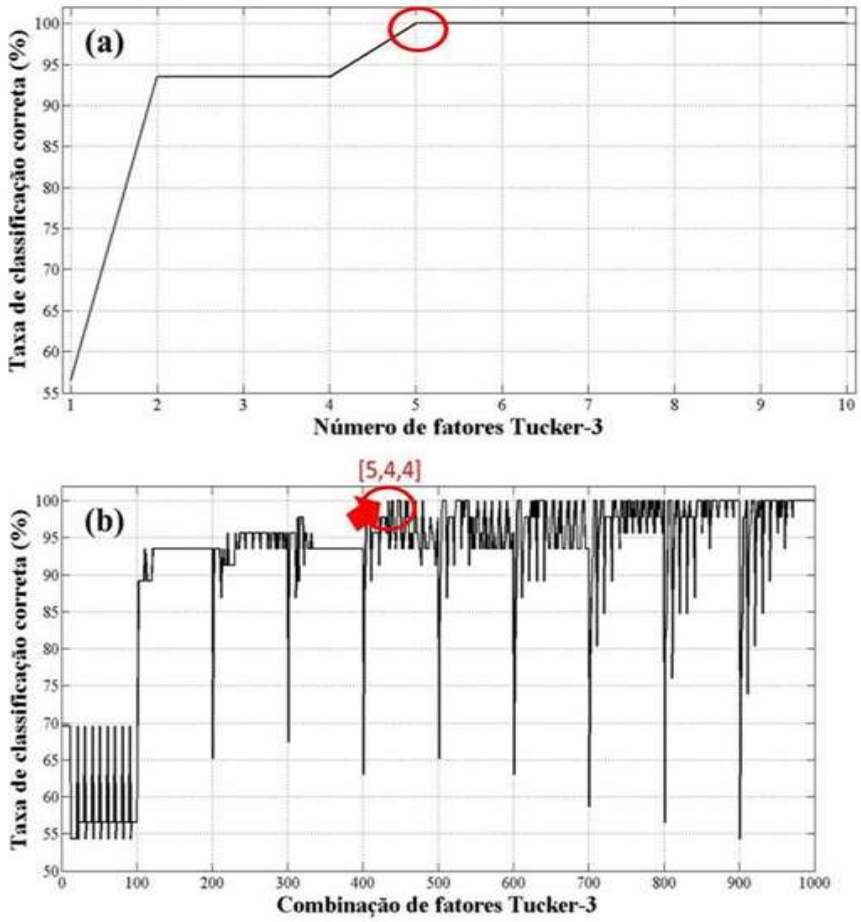


Figura 32 – Taxa de classificação correta por validação cruzada versos: número de Fatores Tucker-3 (**a**), diferentes combinações de fatores Tucker-3 (**b**).

O modelo de classificação TUCKER-3-LDA, para o banco de dados de presunto de Parma, apresenta uma maior parcimônia com um total de 5 fatores, com 100% de taxa de classificação correta da validação cruzada usando número de fatores fixos (**Figura 32a**) e combinações de fatores com busca exaustiva (**Figura 32b**). Em geral, os resultados dos modelos Tucker-3, quando aplicado com restrição de ortogonalidade a dados trilineares, são bem semelhantes aos do PARAFAC. A matriz de confusão relacionada à predição das amostras de teste usando Tucker-3 é apresentada na **Tabela 9**.

Tabela 9- Resultado de classificação usando TUCKER-3com 5 fatores para o grupo de Teste dos dados de presunto de Parma curado a seco. Valores em percentual **(%).**

TUCKER-3-LDA		**Classe predita**			
Conjunto de dados	**Classe real**	Cru	Salgado	Maturado	Envelhecido
Presunto de Parma	Cru	100	-	-	-
	Salgado	-	100	-	-
	Maturado	-	-	86	14
	Envelhecido	-	-	29	71

5.2.4 U-PLS-DA

Na **Figura 33** é apresentado o gráfico do número de variáveis latentes versos a taxa de classificação correta de validação cruzada. Um total de 7 variáveis latentes foram selecionadas por ser menor número de variáveis com 100% de acerto de classificação das amostras de treinamento.

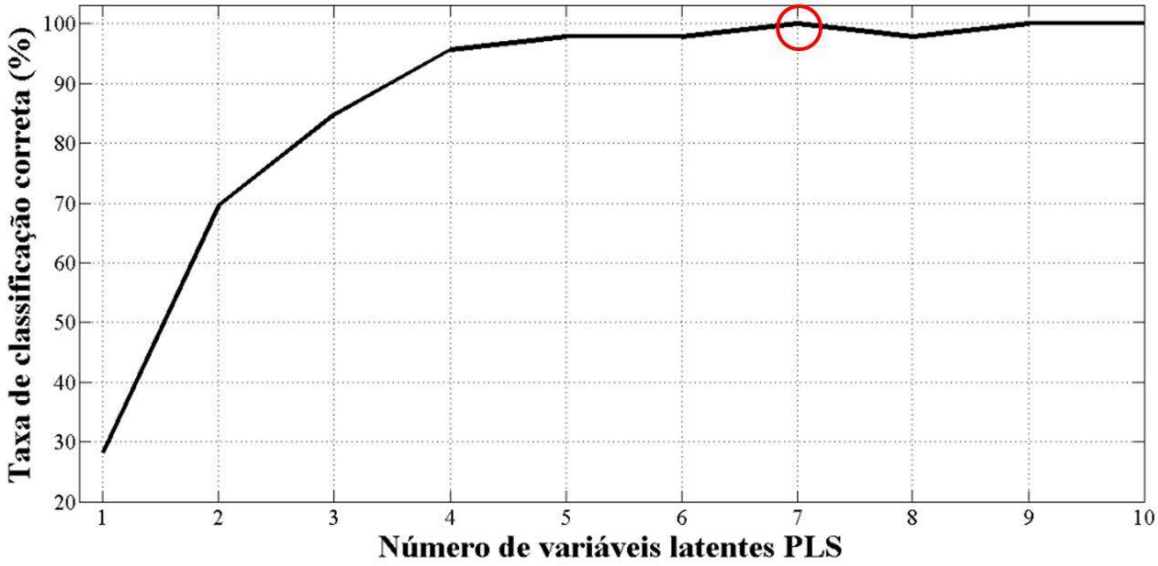


Figura 33 –Taxa de classificação correta por validação cruzada versos número de variáveis latentes para os dados de presunto de Parma.

Após determinar o número de fatores que serão utilizados nos modelos de classificação U-PLS-DA, a determinação das classes das amostras de teste foi realizada. Taxas de classificação correta de 100% foram alcançadas para as amostras de teste, de acordo com as matrizes de confusão presente na **Tabela 10**.

Tabela 10- Resultado de classificação usando U-PLS-DA com 7 fatores para o grupo de Teste dos dados de presunto de Parma curado a seco. Valores em percentual **(%).**

U-PLS-DA		**Classe predita**				
Conjunto de dados	**Classe real**	Cru	Salgado	Maturado	Envelhecido	Classe não atribuída
Presunto de Parma	Cru	100	-	-	-	
	Salgado	-	100	-	-	
	Maturado	-	-	86	14	
	Envelhecido	-	-	14	57	29

Avaliando a **Tabela 10** verifica-se que o modelo U-PLS-DA possui um baixo grau de discriminação entre as classes de amostras Maturadas e Envelhecidas, quando comparado com as demais classes. Tal conclusão possui maior evidência para as amostras envelhecidas que por sua vez apresentaram erros do tipo classe não atribuídas.

5.2.5 NFE (*No feature extraction*)

Para verificar se realmente os vetores discriminantes contribuem para otimizar a discriminação das amostras, foi realizada a predição das amostras de teste utilizando apenas distancia sobre os dados desdobrados. Na **Tabela 11** são apresentados os resultados de predição a partir da matriz de confusão para as amostras de teste.

Tabela 11- Resultado de classificação usando NFE para o grupo de Teste dos dados de presunto de Parma curado a seco. Valores em percentual **(%).**

NFE		**Classe predita**			
Conjunto de dados	**Classe real**	Cru	Salgado	Maturado	Envelhecido
Presunto de Parma	Cru	100	-	-	-
	Salgado	-	100	-	-
	Maturado	-	-	71	29
	Envelhecido	-	-	43	57

Como pode ser observado na **Tabela 11**, o modelo NFE apresentou baixa capacidade para discriminar as classes de amostras maturas e envelhecidas entre si. Esse resultado pode ser atribuído a não extração de características das classes que possam auxiliar na capacidade preditiva do modelo. Ainda, é possível observar uma forte sobreposição dos espectros de fluorescência desdobrados (**Figura 34**).

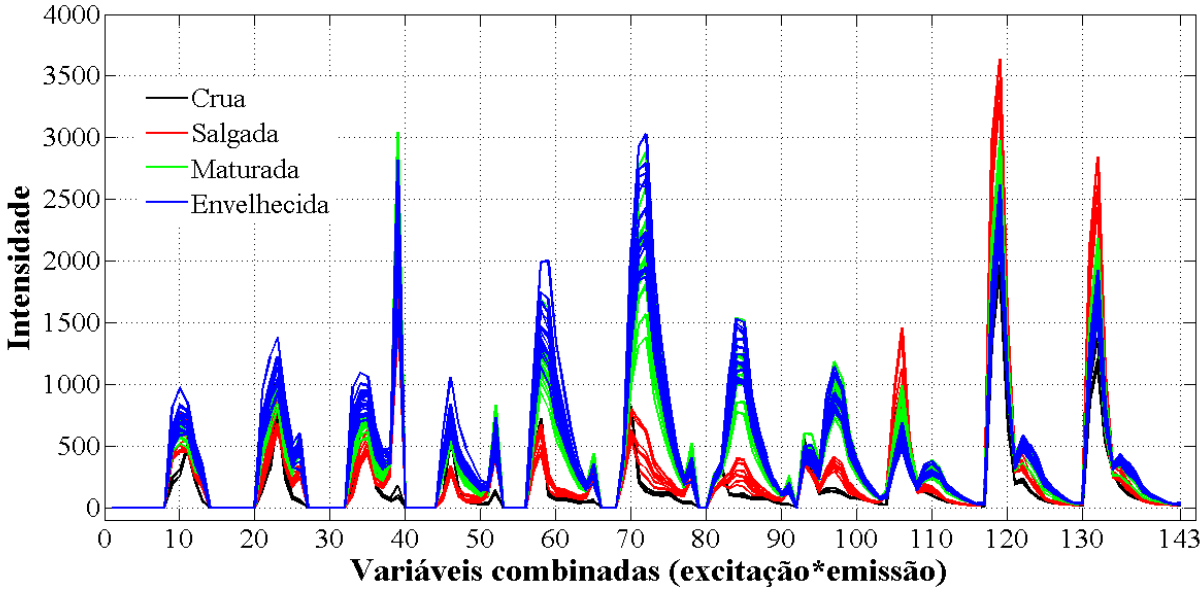


Figura 34 –Perfil de desdobramento das amostras das classes de treinamento do banco de dados e presunto de Parma curado a seco.

Na **Tabela 12**, apresentada a seguir, são disponibilizados os valores de taxa de classificação correta (%) das para o grupo de teste usando: 2D-LDA, PARAFAC-LDA, TUCKER-3-LDA, U-PLS-DA e NFE.

Tabela 12 - Resultado de classificação do grupo de Teste usando diferentes estratégias de modelagem.(Dados de presunto de Parma curado a seco) Valores em percentual (%).

Dados	2D-LDA	PARAFAC-LDA	TUCKER-3-LDA	U-PLS-DA	NFE
Presunto de Parma Curado a seco	86 (5)	86 (5) 76 (6)	86 (5)	81 (7)	76

*Entre parênteses: numero de vetores de projeção, fatores e variáveis latentes.

A partir da **Tabela 12** é possível identificar que o 2D-LDA apresentou resultados comparáveis aos encontrados com outras estratégias de modelagem presentes na literatura. Em geral, o valor de taxa de classificação correta das amostras de teste ficou entre 76% e

86%, sendo que os erros estão concentrados nas amostras de teste rotuladas como pertencentes à classe de amostras maturadas e envelhecidas, evidenciando assim a pouca diferenciação desses dados pela saturação do processo de maturação em termos de fenômenos que possam ser detectados pela técnica de autofluorescência de superfície.

5.3 Conjuntos de dados de óleo vegetal comestível

5.3.1 2D-LDA

Na **Figura 35** são apresentados os valores de taxa de classificação correta obtida para validação cruzada no conjunto de amostras de treinamento.

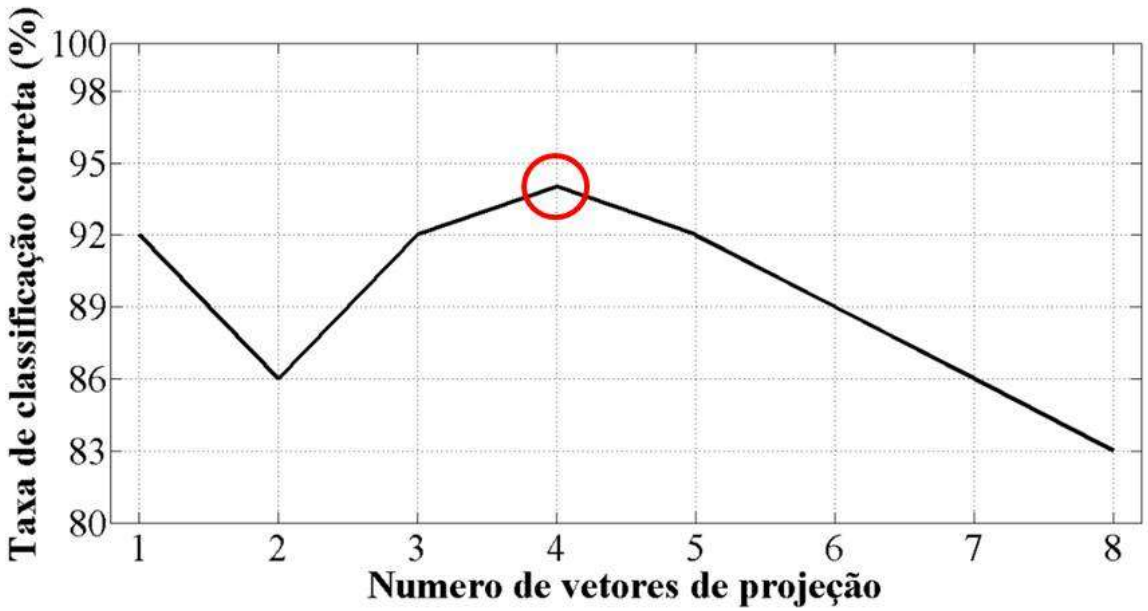


Figura 35 –Taxa de classificação correta por validação cruzada obtida versos número de vetores de projeção para o banco de dados de óleo vegetal comestível.

A partir do resultado apresentado na **Figura 35**, o modelo 2D-LDA que apresenta maior taxa de classificação correta para validação cruzada é o que apresenta um número de vetores de projeção igual a 4. Na **Figura 36** são apresentados os 4 vetores de características obtidos a partir do produto entre os vetores de projeção e as matrizes de treinamento.

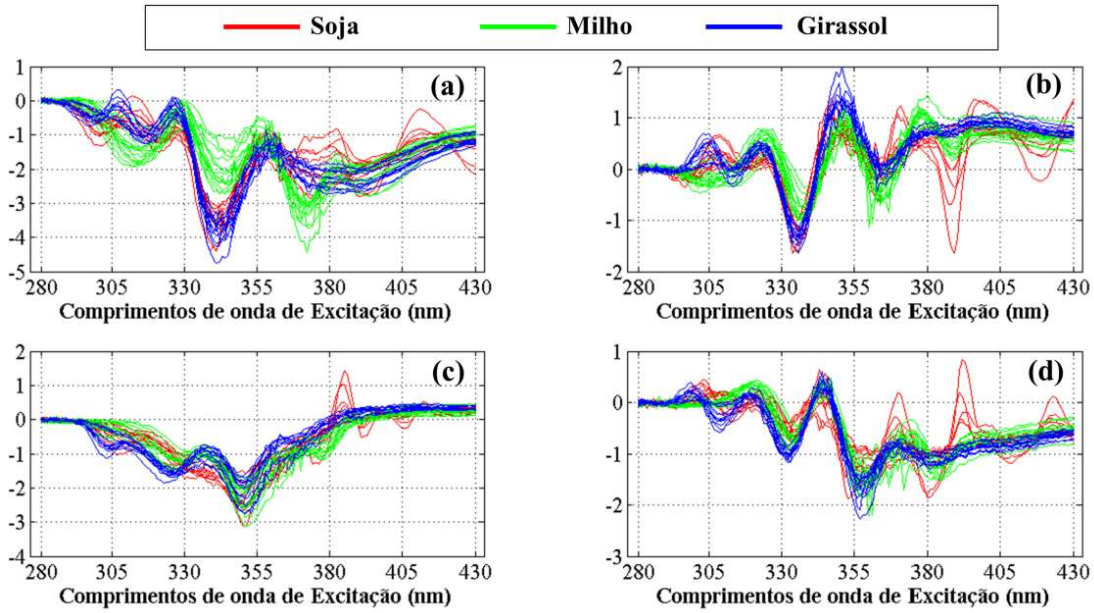


Figura 36- Vetores de características do 2D-LDA para as amostras de treinamento obtidos com:primeiro (a), segundo (b), terceiro (c) e quarto (d) vetores de projeção.

Avaliando os vetores de características apresentados na **Figura 36**, verifica-se uma melhor discriminação entre os perfis das amostras de soja e as demais classesna faixa entre330 e 350 nm (terceiro vetor de projeção, **Figura 36c**). Já as amostras de milho têm um perfil característico que se distingue na faixa de comprimentos de onda que se entende de 305 a 380 nm (primeiro vetor de projeção, **Figura 36a**). No caso das amostras de girassol, a melhor discriminação das outras classes é observada no terceiro e quarto vetores de projeção (**Figura 36c**, **36d**), entre 305 e 330 nm. Estes intervalos de comprimento de onda são semelhantes aos empregados em um estudo anterior [71] sobre o uso de fluorescência sincrônica para a quantificação de adulterações de azeite por soja (315-365 nm), milho (315-392 nm), girassol (315-365 nm)e outros óleos. As principais contribuições para os espectros estão associados a presença de antioxidantes fenólicos (em torno de 300 – 330 nm) e algumas vitaminas que apresentam excitação na faixa entre 350 – 600 nm [74-75].

Após a obtenção dos vetores de características para as amostras de treinamento, foi realizada a classificação das amostras de teste. Como resultado, todas as amostras foram corretamente classificadas.

5.3.2 PARAFAC-LDA

A escolha do número de fatores para construção dos modelos PARAFAC-LDA foi dada mediante a avaliação da taxa de classificação correta da validação cruzada (**Figura 37a**), valor de corcondia (**Figura 37b**) e soma do quadrado dos erros do modelo (**Figura 37c**) com diferentes números de fatores avaliados.

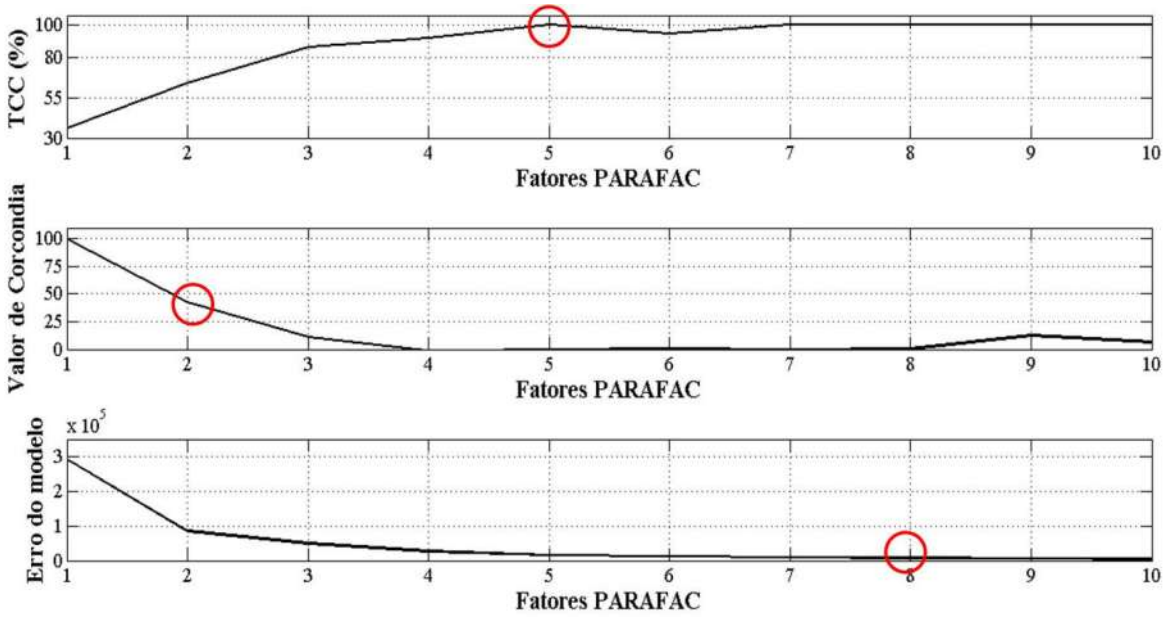


Figura 37 –Taxa de classificação correta por validação cruzada versos número de fatores PARAFAC (a), valor de corcondia (b) e de erro de modelo (c).

Na **Figura 37** verifica-se que o modelo PARAFAC-LDA com melhor taxa de classificação correta na validação cruzada necessita de pelo menos5 fatores. A análise dos demais resultados possibilita concluir que os valores de corcondia decaem rapidamente com o número de fatores, de forma que a partir de dois fatores o mesmo já apresenta um valor

abaixo de 50%. Ainda, os erros do modelo apresentam um valor mínimo próximo do número total de deltas utilizados na medida, indicando que o modelo necessita de muitos fatores para explicar os dados. Essa característica está associada ao uso de PARAFAC em dados do tipo não trilineares.

Diferente do que ocorreu nos dados de presunto de Parma, as ferramentas usadas para determinação do número de fatores não convergem para um mesmo resultado ou para valores próximos. Diante disso, apenas o critério de taxa de classificação correta da validação será adotado e um total de 5 fatores serão utilizados no modelo PARAFAC-LDA para prever as amostras de teste, como apresentado na matriz de confusão contida na **Tabela 13**.

Tabela 13- Resultado de classificação usando PARAFAC-LDA com 5 fatores para o grupo de Teste. (Dados de óleo vegetal comestível) Valores em percentual (%).

PARAFAC-LDA		**Classe predita**		
Banco de dados	**Classe real**	Soja	Milho	Girassol
Óleo vegetal comestível	Soja	100	-	-
	Milho	-	100	-
	Girassol	-	25	75

5.3.3 TUCKER-3

A escolha do número de fatores para construção dos modelos TUCKER-3-LDA foi realizada por intermédio de duas estratégias de avaliação da taxa de classificação correta: número de fatores iguais para todos os modos da matriz (**Figura 38a**) e a partir de diferentes combinações realizadas com busca exaustiva (**Figura 38b**).

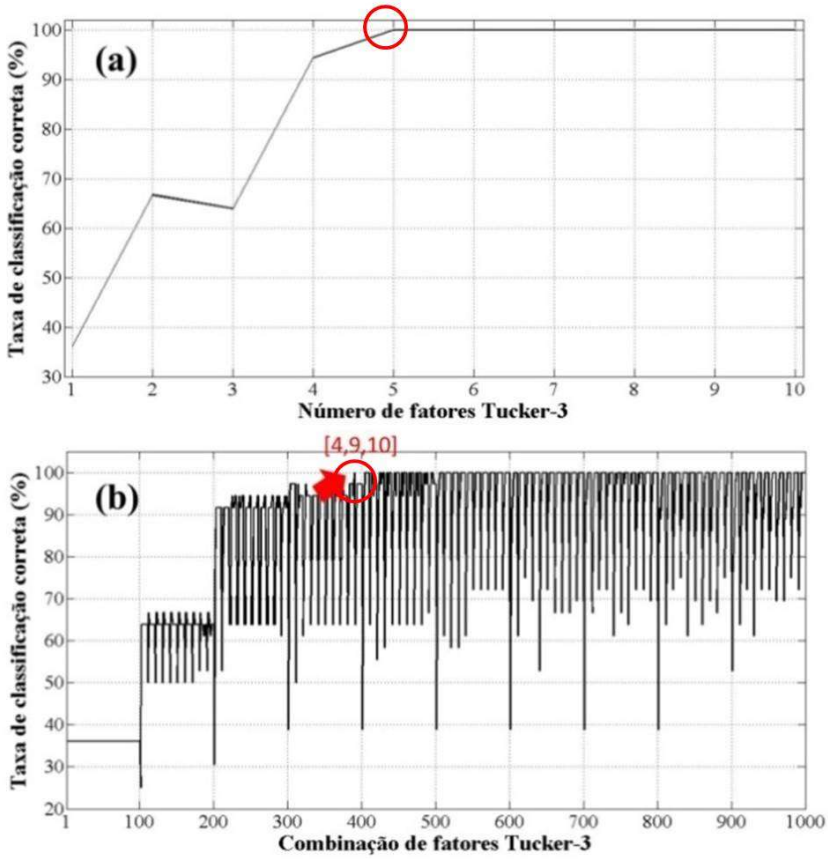


Figura 38 –Taxa de classificação correta por validação cruzada versos: número de Fatores Tucker-3 (**a**), diferentes combinações de fatores Tucker-3 (**b**).

O modelo de classificação TUCKER-3-LDA, para o banco de dados de óleo vegetal comestível, apresenta 100% de taxa de classificação da validação cruzada, quando utilizados 5 fatores (**Figura 38a**). Ao se fazer uma busca aleatória, apenas 4 fatores de escores (**Figura 38b**) foram utilizados para classificação correta de todas as amostras. Diante disso, foram gerados dois modelos utilizando 4 e 5 fatores. A matriz de confusão relacionada à predição das amostras de teste é apresentada na **Tabela 14**.

Tabela 14 - Resultado de classificação usando TUCKER-3-LDA com 4 e 5 fatores para o grupo de Teste.(Dados de óleo vegetal comestível) Valores em percentual (%).

TUCKER-3-LDA		**Classe predita**		
Banco de dados	**Classe real**	Soja	Milho	Girassol
Óleo vegetal comestível (4 fatores Tucker)	Soja	100	-	-
	Milho	-	100	-
	Girassol	-		100
Óleo vegetal comestível (5 fatores Tucker)	Soja	100	-	-
	Milho	-	100	-
	Girassol	-	25	75

5.3.4 U-PLS-DA

Na **Figura 39** é apresentado o gráfico do número de variáveis latentes versus a taxa de classificação correta de validação cruzada. Um total de 5 variáveis latentes foram selecionadas por ser menor número de variáveis com 100% de acerto de classificação das amostras de treinamento.

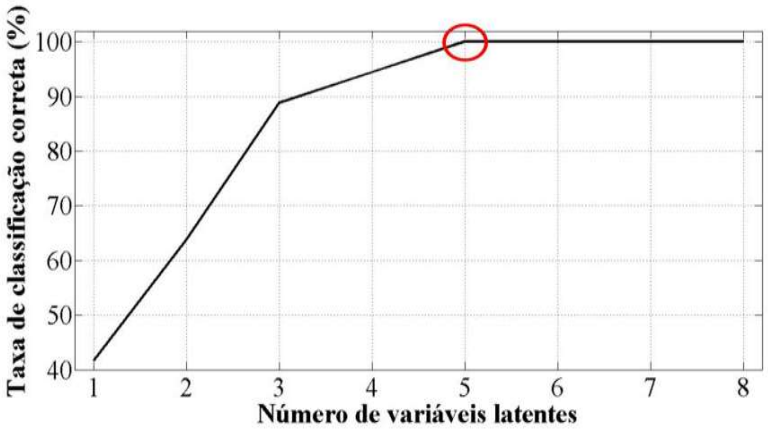


Figura 39 – Taxa de classificação correta por validação cruzada versos número de variáveis latentes para os dados de óleo vegetal comestível

Após determinar o número de fatores que serão utilizados nos modelos de classificação U-PLS-DA, a determinação das classes das amostras de teste foi realizada. Taxas de classificação correta de 100% foram alcançadas para as amostras de teste, de acordo com as matrizes de confusão presente na **Tabela 15**.

Tabela 15 - Resultado de classificação usando U-PLS-DA com 5 variáveis latentes para o grupo de Teste. (Dados de óleo vegetal comestível) Valores em percentual (%).

U-PLS-DA		**Classe predita**		
Banco de dados	**Classe real**	Soja	Milho	Girassol
Óleo vegetal comestível	Soja	100	-	-
	Milho	-	100	-
	Girassol	-	25	75

5.3.5 NFE (No feature extraction)

Na **Tabela 16** são apresentadas as classes preditas das amostras de teste dos dados de óleo vegetal comestível. Como pode ser observado, apenas a classe de soja apresentou taxa de classificação acima de 83%. O maior número de erros foi encontrado nas classes de milho e girassol, pode ser resultado da sobreposição ou adequação do cálculo de distância utilizado. Os perfis dos dados desdobrados são apresentados na **Figura 40**.

Tabela 16- Resultado de classificação usando NFE para o grupo de Teste. (Dados de óleo vegetal comestível) Valores em percentual (%).

NFE		**Classe predita**		
Banco de dados	**Classe real**	Soja	Milho	Girassol
Óleo vegetal comestível	Soja	100	-	-
	Milho	17	83	-
	Girassol	25	25	50

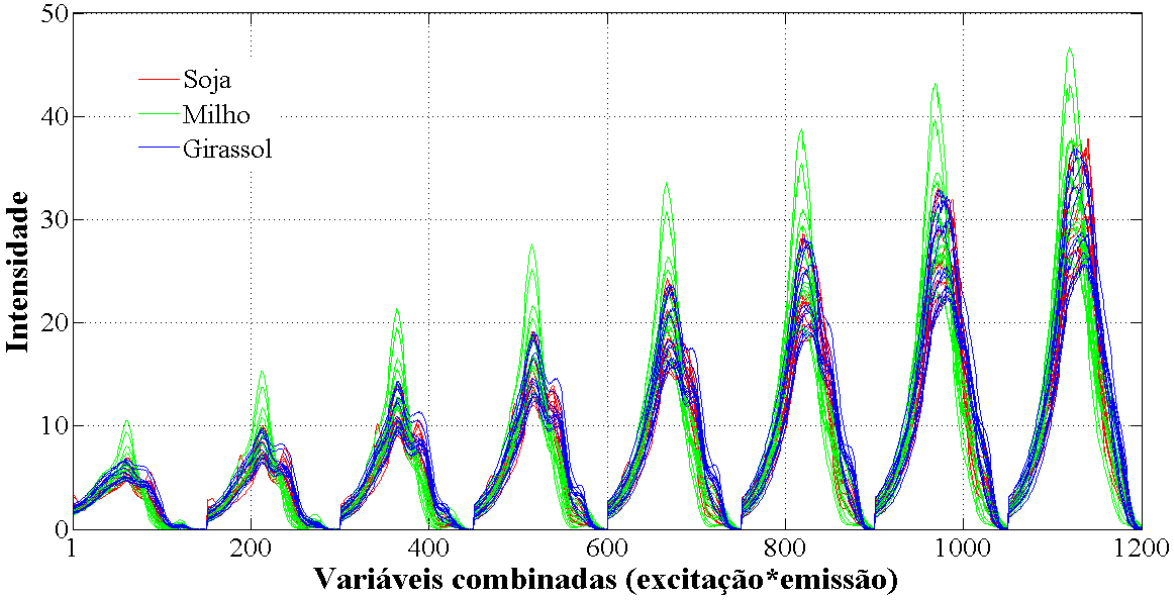


Figura 40 –Perfil de desdobramento das amostras das classes de treinamento do banco de dados e óleo vegetal comestível.

Na **Tabela 17**, apresentada a seguir, são disponibilizados os valores de taxa de classificação correta (%) das para as amostras do grupo de teste usando: 2D-LDA, PARAFAC-LDA, TUCKER-3-LDA, U-PLS-DA e NFE.

Tabela 17-Resultado de classificação do grupo de Teste usando diferentes estratégias de modelagem.(Dados de óleo vegetal comestível) Valores em percentual (%).

Dados	2D-LDA	PARAFAC-LDA	TUCKER-3-LDA	U-PLS-DA	NFE
Óleo vegetal comestível	100 (4)	92(5)	100(4) 92(5)	92 (5)	77

Na **Tabela 17** é possível identificar que apenas os modelos 2D-LDA e Tucker-3-LDA apresentaram uma classificação correta para todas as amostras de teste.Vale salientar que para o uso de Tucker-3-LDA (4 fatores), a interpretação química dos dados é inviável pois o número de fatores da matrizes de escores e das matrizes de loadings foram bem diferentes, como apresentado na **Figura 38**. Os erros do modelo PARAFAC-LDA podem ser associados a pouca adequação a dados não trilineares. Para os modelos U-PLS-DA e NFE, os erros podem está relacionados a sobreposição dos dados quando desdobrados, em especial para o NFE que não usa nenhuma estratégia para extração de propriedades discriminantes, seja a partir de vetores ou variáveis latentes.

Na **Tabela 18** é apresentado um resumo dos resultados encontrados com os modelos de classificação para cada um dos bancos de dados estudados. A partir desses resultados é possível inferir que os modelos 2D-LDA apresentaram, para todos os bancos de dados, os maiores valores de taxa de classificação correta, tornando os modelos com poder de discriminante melhor ou comparável com os demais modelos apresentados na literatura que foram usados para comparação.

Tabela 18-Resultados comparativos: Taxas de classificação corretas obtidas nos conjuntos de teste para os conjuntos de dados estudados. O número de vetores de projeção, fatores ou variáveis latentes empregadas em cada modelo é indicado entre parênteses.Valores em percentual (%).

Dados	2D-LDA	PARAFAC-LDA	TUCKER-3-LDA	U-PLS-DA	NFE
Simulado I	100 (1)	100 (2)	100 (2)	100 (3)	100
Simulado II	100 (1)	100 (4)	100 (4)	100 (4)	70
Presunto de Parma Curado a seco	86 (5)	86 (5) 76 (6)	86 (5)	81 (7)	76
Óleo vegetal comestível	100 (4)	92 (5)	100 (4) 92 (5)	92 (5)	77

CAPÍTULO 6

CONCLUSÃO

6. CONCLUSÃO

Neste trabalho foi proposto o uso de análise discriminante linear em duas dimensões (2D-LDA), até então usado em estudos de reconhecimento de faces em imagem, como uma nova estratégia para classificar dados químicos de segunda ordem.

Para avaliar o desempenho do 2D-LDA e a capacidade discriminante do modelos gerados, foram utilizados quatro conjuntos de dados que contemplassem três condições: Dados trilineares (auto fluorescência de superfície para identificação do nível de maturação de presunto de Parma curado a seco), dados não trilineares (2 conjuntos de simulados de matriz de excitação-emissão); e dados não bilineares/trilineares (Fluorescência sincrônica total para identificação da matéria prima de óleos comestíveis).

A principal resposta do algoritmo 2D-LDA são os vetores de características, os quais são usados na etapa de classificação das amostras. Foi possível verificar que a partir desses vetores é possível fazer inferências sobre perfis dos analitos ou comprimentos de onda que são responsáveis pelas características discriminantes dos modelos. Mais especificamente, para o conjunto de dados simulados I foram obtidos vetores de características com perfil semelhante ao de emissão dos analitos A1, A2 e A4. Para os demais conjuntos de dados foi possível verificar que os vetores de características auxiliaram na verificação de quais comprimentos de onda são responsáveis pela discriminação das amostras.

Em geral, os resultados com os modelos de classificação usando 2D-LDA foram melhores ou comparáveis aos obtidos com as seguintes estratégias: NFE (procedimento de classificação baseado em calculo de distância em dados desdobrados sem extração de características), U-PLS-DA (Análise Discriminante usando mínimos quadrados parciais nos dados desdobrados), PARAFAC-LDA (LDA com escores PARAFAC) e TUCKER-3-LDA

(LDA com escores TUCKER-3). Mais especificamente, os modelos 2D-LDA apresentaram valores de taxa de classificação correta de 100% para os conjuntos de dados simulados e de óleo vegetal comestível. Para o banco de dados de presunto de Parma, o 2D-LDA alcançou 86% de acerto na classificação das amostras de teste, sendo este o maior valor encontrado entre os diferentes modelos de classificação avaliados.

Os resultados alcançados indicam que o 2D-LDA é de fato uma estratégia promissora para a classificação baseada em dados químicos de segunda ordem.

6.1 Propostas futuras

Como propostas futuras planeja-se a implementação e disponibilização do algoritmo do 2D-LDA em forma de interface do Matlab® e adaptação do algoritmo de classificação 2D-LDA com as técnicas de seleção de variáveis e de intervalos.

REFERÊNCIAS

[1] A. K. Smilde, R. Bro, P. Geladi. *Multi-way Analysis. In: Applications in the Chemical Sciences*. **Wiley Publisher**, England (2004) 396 p.

[2] A. Evrim, Y. Bülent. *Unsupervised Multiway Data Analysis: A Literature Survey.* **IEEE Transactions On Knowledge And Data Engineering**, 21 (2009) 6-20.

[3] A. C. Olivieri, G. M. Escandar, H. C. Goicoechea, A. M. DE La Peña. *Fundamentals and analytical applications of multiway calibration*, **Elsevier,** USA (2015) 618 p.

[4] A. C. Olivieri. *Recent advances in analytical calibration with multi-way data*. **Anal. Methods,** 4 (2012) 1876-1886.

[5] K. S. Booksh, B. Bronk, J. Czege. *Three-Way Calibration*. **Comprehensive Chem.**, 3 (2009) 379-412.

[6] A. C. Olivieri. *Analytical Figures of Merit: From Univariate to Multiway Calibration*. **Chem. Reviews**, 114 (2014) 5358-5378.

[7] A. C. Olivieri, G. Escandar. *Practical three-way calibration*, **Elsevier**, USA (2014) 330 p.

[8] C. Durante, R. Bro, M. Cocchi. A *classification tool for N-way array based on SIMCA methodology*. **Chemometr Intell Lab Syst .**, 106 (2011) 73-85.

[9] R. M. Callejón, J. M. Amigo, E. Pairo, S. Garmón, J. A. Ocana, M. L. Morales. *Classification of Sherry vinegars by combining multidimensional fluorescence, parafac and different classification approaches*. **Talanta**, 88 (2012) 456-462.

[10] L. Lenhardt, R. Bro, I. Zekovic, T. Dramicanin, M. D. Dramicanin. *Fluorescence spectroscopy coupled with PARAFAC and PLS DA for characterization and classification of honey*. **Food Chem.**, 175 (2015) 284-291.

[11] M. Valcárcel, S. Cárdenas. *Qualitative Analysis*. **Encyclopedia of Analytical Science** (2005) 2 ed.

[12] R. MUÑOZ-OLIVAS. *Screening analysis: an overview of methods applied to environmental, clinical and food analyses*. **TRAC**, 23 (2004) 203-216.

[13] F. Guimet, R. Boqué, J. Ferré. *Application of non-negative matrix factorization combined with Fisher's linear discriminant analysis for classification of olive oil excitation–emission fluorescence spectra*. **Chemometr Intell Lab Syst .**, 81 (2006) 94-106.

[14] S. M. Azcarate, A. A. Gomes, M. R. Alcaraz, M. C. U. Araújo, J. M. Camiña, H. C. Goicoechea.*Modeling excitation–emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety*. **Food Chem.,** 184 (2015) 214-219.

[15] R. Henrion. *N-way principal component analysis theory, algorithms and applications*. **Chemometr Intell Lab Syst .**, 25 (1994) 1-23.

[16] R. Bro. *PARAFAC. Tutorial and applications*. **Chemometr Intell Lab Syst .**, 38 (1997) 149-171.

[17] L. R. Tucker. *Some mathematical notes on three-mode factor analysis*. **Psychometrika**, 31 (1966) 279-311.

[18] R. A. Fisher. *The use of multiple measurements in taxonomic problems*. **Annals Of Eugenics**, 7 (1936) 179-188.

[19] M. Barker, W. Rayens. *Partial least squares for discrimination*. **J. Chem.**, 17 (2003) 166-173.

[20] S. S. Ouertani, G. Mazerolles, J. Boccard, S. Rudaz, M. Hanifi. *Multi-way PLS for discrimination: Compact form equivalent to the tri-linear PLS2 procedure and its monotony convergence*. **Chemometr Intell Lab Syst .**, 133 (2014) 25-32.

[21] M. Li, B. Yuan, *2D-LDA: A statistical linear discriminant analysis for image matrix*, **Pattern Recogn. Lett.,** 26 (2005) 527-532.

[22] Z. Liang, Y. Li, P. Shi.*A note on two-dimensional linear discriminant analysis*. **Pattern Recogn. Lett.** 29 (2008) 2122-2128.

[23] D. U. Cho, U.D. Chang, B. H. Kim, S. H. Lee, Y. L. J. Bae, S. C. Ha. *2D Direct LDA Algorithm for Face Recognition*, **Fourth International Conference on Software Engineering Research**, **Management and Applications (SERA'06)** 2006.

[24] A. Rozza, G. Lombardi, E. Casiraghi , P. Campadelli. *Novel Fisher discriminant classifiers*.**Pattern Recog**., 45 (2012) 3725-3737.

[25] Y. Xu, J. Yang, Z. Jin. *A novel method for Fisher discriminant analysis*. **Pattern Recog.,** 37 (2004) 381-384.

[26] Z. Ji, P. Jing, T. Yu, Y. Su, C. Liu.*Ranking Fisher discriminant analysis*. **Neurocomputing,** 120 (2013) 54-60.

[27] K. H. Esbensen, P. Geladi. *2.13 – Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History,Practice*. **Comprehensive Chem.,** (2009) 211–226.

[28] S. Wold, M, Sjöström, L. Eriksson. *PLS-regression: a basic tool of chemometrics*. **Chemometr Intell Lab Syst .**, 58 (2001) 109-130.

[29] H. Martens, T. Naes. *Multivariate Calibration*, **Wiley** (1992) 438 p.

[30] A. Folch-Fortuny, J. M. Prats-Montalbán, S. Cubero, J. Blasco, A. Ferrer. *VIS/NIR hyperspectral imaging and N-way PLS-DA models for detection of decay lesions in citrus fruits*. **Chemometr Intell Lab Syst .**, 156 (2016) 241-248.

[31] S. E. G. Porter, D. R. Stoll, S. C. Rutan, P. W. Carr, J. D. Cohen. *Analysis of Four-Way Two-Dimensional Liquid Chromatography-Diode Array Data: Application to Metabolomics*. **Anal. Chem.**, 78 (2006) 5559-5569.

[32] H. C. Goicoechea, S. YU, A. C. Olivieri, A. D. Campiglia. *Four-Way Data Coupled to Parallel Factor Model Applied to Environmental Analysis: Determination of 2,3,7,8-Tetrachloro-dibenzo-para-dioxin in Highly Contaminated Waters by Solid−Liquid Extraction Laser-Excited Time-Resolved Shpol'skii Spectroscopy*. **Anal. Chem.**, 77 (2005) 2608-2616.

[33] A. Xia, H. Wu, S. Li, R. Yu. *Alternating penalty quadrilinear decomposition algorithm for an analysis of four-way data arrays*. **J. Chemometrics**, 21 (2007) 133-144.

[34] A. A. Gomes. *Algoritmo das projeções sucessivas para seleção de variáveis em calibração de segunda ordem*. **Tese (Doutorado)**, João Pessoa (2015), 126 p.

[35] D. L. Massart. Handbook of Chemometrics and Qualimetrics: Part A. **Elsevier Science**, Amsterdam (1997), 867 p.

[36] B. E. Wilson, B.R. Kowalski. *QuantitativeAnalysis in the Presence of Spectral Interferents Using Second-Order Nonbilinear Data*. **Anal. Chem**., 61 (1989) 2277-2284.

[37] C. G. Zampronio, S. P. Gurden, L. A. B. Moraes, M.N. Eberlin, A. K. Smilde,R. J. Poppi.*Direct sampling tandem mass spectrometry (MS/MS) and multiway calibration for isomer quantitation*. **Analist**, 127 (2002) 1054-1060.

[38] M. M. Sena, M. G. Trevisan, R. J. Poppi. PARAFAC: *Uma ferramenta quimiométrica*

para tratamento de dados multidimensionais. Aplicações na determinação direta de fármacos em plasma humano por espectrofluorimetria. **Química Nova**, 28 (2005) 910-920.

[39] A. Baum, P. W. Hansen, A. S. Meyer, J. D. Mikkelsen. *Simultaneous measurement of two enzyme activities using infrared spectroscopy: A comparative evaluation of PARAFAC, TUCKER and N-PLS modeling*. **Anal. Chim. Acta,** 790 (2013) 14-23.

[40] A. Graham. *Kronecker Products and Matrix Calculus with Applications*. **Wiley**, New York (1981) 130 p.

[41] E. K. Gnang, Y. Filmus. *On the spectra of hypermatrix direct sum and Kronecker products constructions*. **Linear Algebra and its Applications**, 519 (2017) 238-277.

[42] *An interactive introduction to The Tucker3 model in Chemometrics*. Disponível em: <http://www.models.life.ku.dk/~courses/tucker/tindex.htm>. Data de acesso: 05/04/2017.

[43] R. A. Harshman. *Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-modal factor analysis*, **UCLA Working Pap. Phonetics**, 16 (1970) 1–84.

[44] J. D. Carroll, Jih-Jie Chang. *Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition*. **Psychometrika,** 35 (1970) 283-319.

[45] C. A. Andersson, R. Bro. *The N-way Toolbox for MATLAB*. **Chem. Int. Lab. Sys.**, 52 (2000) 1-4.

[46] R. Bro. *Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications*. **Tese (Doutorado)**, Holanda (1998) 290 p.

[47] M. M. Barnard. *The secular variations of skull characters in four series of Egyptian skulls*. **Ann. Eugenics**, 6 (1935) 352–371.

[48] M. OTTO. *Chemometrics : statistics and computer application in analytical chemistry*. **Wiley-VCH**, New York (1999) 314 p.

[49] U. T. C. P. Souto, M. F. Barbosa, H. V. Dantas, A. S. Pontes, W. S. Lyra, P. H. G. D. Diniz, M. C. U. Araújo, E. C. Silva.*Identification of adulteration in ground roasted coffees using UV–Vis spectroscopy and SPA-LDA*. **Food Sci. Tec.**, 63 (2015) 1037-1041.

[50] A. C. Silva, L. F. B. L. Pontes, M. F. Pimentel, M. J. C. Pontes. *Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods*. **Talanta**, 93 (2012) 129-134.

[51] A. C. Silva, J. E. M. Paz, L. F.B. L. Pontes, S. G.Lemos, M. J. C. Pontes. *An electroanalytical method to detect adulteration of ethanol fuel by using multivariate analysis*. **Electrochim. Acta**, 111 (2013) 160-164.

[52] F. Gosetti, U. Chiuminatto, E. Mazzucco,R. Mastroianni, E. Marengo. *Ultra-high-performance liquid chromatography/tandem high-resolution mass spectrometry analysis of sixteen red beverages containing carminic acid: Identification of degradation products by using principal component analysis/discriminant analysis*. **Food Chem.**, 167 (2015) 454-462.

[53] S. Rezzi, D. E. Axelson, K. Héberger, F. Reniero, C. Mariani, C. Guillou. *Classification of olive oils using high throughput flow 1H NMR fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks*. **Anal. Chim. Acta**, 552 (2005) 13-24.

[54] N. Kumar, A. Bansal, G.S. Sarma, R. K. Rawal. *Chemometrics tools used in analytical chemistry: An overview*. **Talanta**, 123 (2014) 186-199.

[55] S. F. C. Soares, A. A. Gomes, M. C. U. Araujo, A. R. G. F., R. K. H. Galvão. *The*

successive projections algorithm. **TrAC**, 42 (2013) 84-98.

[56] L. A. Ribeiro, A. S. Soares, T. W.Lima,C. A. C. Jorge, R. M. Costa, R. L. Salvini, C. J. Coelho, F. M. Federson, P. H.R. Gabriel. *Multi-objective Genetic Algorithm for Variable Selection in Multivariate Classification Problems: A Case Study in Verification of Biodiesel Adulteration*. **Procedia Comput Sci**, 51 (2015) 346-355.

[57] R. G. Brereton, G. R.Lloyd. *Partial least squares discriminant analysis: taking the magic away*. **J. Chemometrics**, 28 (2014) 213-225.

[58] D. Ballabio, V. Consonni. *Classification tools in chemistry. Part 1: linear models. PLS-DA*. **Anal. Methods**, 5 (2013) 3790-3798.

[59] R. Bro. *Multilinear PLS*. Disponível em:<http://www.models.life.ku.dk/~rasmus/presentations/Npls_sugar/npls.htm>.Acesso em: 22/05/2016.

[60] K. V. Branden, M. Hubert.*Robust classification in high dimensions based on the SIMCA Method*. **Chemometr Intell Lab Syst**, 79 (2005) 10-21.

[61] F. Toldrá. *Dry-cured meat products*. **FOOD & NUTRITION PRESS**, USA (2002) 239 p.

[62] L. Bolzoni, G. Barbieri, R. Virgili.*Changes in volatile compounds of Parma ham during maturation*, **Meat Sci.**, 43 (1996) 301-310.

[63] A. K. Agarwal, *Business and intellectual property: protect your ideas*, **Random House**, India (2016) 168p.

[64] J. K. Møller, G. Parolari, L. Gabba, J. Christensen, L. H. Skibsted. *Monitoring Chemical Changes of Dry-Cured Parma Ham during Processing by Surface Autofluorescence Spectroscopy*, **J. Agric. Food Chem.**, 51 (2003) 1224-1230.

[65] R. D. O'brien. *Fats and Oils: Formulating and Processing for Applications*. (third ed.) **CRC Press**, Florida (2008) 680 p.

[66] S. C. Savva, A. Kafatos, *Vegetable oils: Dietary importance*. **Encyclopedia of food and health** (2016) 365-372.

[67] C. E. T. da Silva, V. L. Filardi, I. M. Pepe, M. A. Chaves, C. M. S. Santos. *Classification of food vegetable oils by fluorimetry and artificial neural networks*. **Food Control**, 47 (2015) 86-91.

[68] Y. G. Martín, J. L. P. Pavón, B. M. Cordero, C. G. Pinto. *Classification of vegetable oils by linear discriminant analysis of Electronic Nose data*. **Anal. Chim. Acta**, 384 (1999) 83-94.

[69] A. S. Luna, A. P. Silva, J. Ferré, R. Boqué. *Classification of edible oils and modeling of their physico-chemical properties by chemometric methods using mid-IR spectroscopy*. **Spectrochim. Acta A**, 100 (2013) 109-114.

[70] F. F. Gambarra-Neto, G. Marino, M. C. U. Araújo, R. K. H. Galvão, M. J. C. Pontes, E. P. de Medeiros, R. S. Lima.*Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis*. **Talanta**, 77 (2009) 1660-1666.

[71] K. Poulli, G. Mousdis, C. Georgiou. *Rapid synchronous fluorescence method for virgin olive oil adulteration assessment*. **Food Chem.**, 105 (2007) 369-375.

[72] K. I. Poulli, G. A. Mousdis, C. A. Georgiou. *Classification of edible and lampante virgin olive oil based on synchronous fluorescence and total luminescence spectroscopy*. **Anal. Chim. Acta** 542 (2005) 151-156.

[73] E. Sikorska, T. Górecki, I. V. Khmelinskii, M. Sikorski, J. Kozioł. *Classification of edible oils using synchronous scanning fluorescence spectroscopy*, **Food Chem.**, 89

(2005) 217-225.

[74] M. Insausti, A. A. Gomes, F. V. Cruz, M. F. Pistonesi, M. C. U. Araujo, R. K. H. Galvão, C. F. Pereira, B. S. F. Band. *Screening analysis of biodiesel feedstock using UV–vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm*. **Talanta**, 97 (2012) 579-583.

[75] J. Tan , R. Li , Z. Jiang, S. Tang, Y. Wang , M. Shi , Y.Xiao , B. Jia , T.Lu , H. Wang. *Synchronous front-face fluorescence spectroscopy for authentication of the adulteration of edible vegetable oil with refined used frying oil*. **Food Chem.**, 217 (2017) 274-280.

Apêndice 1 – Algoritmo 2D-LDA (Validação Externa)

```
function [DistTotal1,errors,Class_Test,Train_features,Test_features]=d2LDA
(Train,Group_Train,Test,Group_Test,mode1,mode2,nTrain,nTest)
%%
% Two-Dimension Linear Discriminant Analysis (2DLDA) for second order chemical data %
%
% Author: Adenilton Camilo da Silva - PPGQ/UFPB
% Orientador: Prof. Dr. Mário Cesar Ugulino de Araujo
% Colaboração: Prof. Dr. Roberto Kawakami (ITA - Instituto Tecnológico de
% Aeronáutica)
%              Dr. Sofacles Figueredo (DEQ, CT - UFPB)
% e-mail: adeniltoncamilo@gmail.com                              %
% August 4, 2015
%last modification: January 17, 2017

%INPUT DATA%

% Train => Training samples in unfold format IK x J ( I = samples, J = mode
% with smaller number of variables, K = mode with largest number of
% variables);

% Test => Test samples in unfold format IK x J ( I = samples, J = mode
% with smaller number of variables, K = mode with largest number of variables);

%Group_Train => Column vector with the class index (1, 2, ...) for each training sample;

%Group_Test => Column vector with the class index(1, 2, ...) for each Test sample;

%mode1=> Number of variables in mode K (larger number of variables);

%mode2=> Number of variables in mode J (smaller number of variables);

%nTrain=> Number of training samples;

%nTest=>Number of test samples;
%OUTPUT DATA

%Class_Test = Vector with the classes for Test samples;
%Train_features = Matrix with feature vectors to Training set;
%Test_features = Matrix with feature vectors to Test set;
%a_vectors= Matrix of autovectors (each column is a vector selected);
%errors = Number of classification errors to Test set;

%Start

%(Step one) Unfolding of the Training Matrix
a=1;
b=mode1;
vectorTrain=zeros(nTrain,[mode1*mode2]);
for i= 1:nTrain
vectorTrain(i,:)=reshape(Train(a:b,:),1,[mode1*mode2]); %vector form
a=a+mode1;
b=b+mode1;
```

```
end

% (Step two) Mean of training samples
meanTrain=mean(vectorTrain);% mean sample in vector form (unfold)
meanTrainm=reshape(meanTrain,mode1,mode2);%mean all samples in matrix form

% (Step three)Calculating the scattering matrices and feature vectors
Sb=zeros(mode2,mode2);
Sw=zeros(mode2,mode2);

%Calculating scattering between classes (Sb)
%description:
%Traini= Class i samples in unfold form;
%mean_Traini=Mean Sample of i class in unfold form;
%meanTrainm= mean all samples in matrix form;
%media_Trainim=Mean Sample of i class in matrix form;
%Sb=scattering between classes;
%sampleim=Sample of the class i;

for i= 1:max(Group_Train)
indexTrain{i} = find(Group_Train==i);
   Traini = vectorTrain(indexTrain{i},:);
mean_Traini{i} = mean(Traini);
media_Trainim{i}=reshape(mean(Traini),mode1,mode2);
    [m1,n1]=size(Traini);
    Sb=Sb+m1*(media_Trainim{i}-meanTrainm)'*(media_Trainim{i}-meanTrainm);

%Calculating scattering intraclass (Sw)
for j = 1:m1
samplei=Traini(j,:);
sampleim=reshape(samplei,mode1,mode2);
       Sw=Sw+(sampleim-media_Trainim{i})'*(sampleim-media_Trainim{i});
end
end

%(Step Four)Calculating the projetion vectors
[U,S]=eig(pinv(Sw)*Sb);
B=diag(S);
[a_values,IX]=sort(B,'descend');
eigval=a_values;%autovalor values
a_vectors=U(:,IX);%Sequence of eigenvectors
a_values2=(a_values/sum(a_values))*100; %autovalor Normalization
a1=0;
for i=1:mode2
a1=[a1+a_values2(i,:)];
p_a_values(i,:)=a1;%Additive efect on autovalor
end
plot(a_values);
figure,plot([p_a_values]);
disp('Número de autovetores para classificação')
n_a_vectors = input('Opcao: ');
eigvec=a_vectors(:,1:n_a_vectors);%Manual selection of autovector number

% (Step five)Calculating the feature vectors
% features vectors for training samples
features=[];
```

```
Train_features=[];
a=[];
for i=1:n_a_vectors
c=1;
d=mode1;
for j= 1:nTrain
a=Train(c:d,:)*a_vectors(:,i);
features(j,:)=a';
c=c+mode1;
d=d+mode1;
end
Train_features{i}=features;

end

%feature vectors for test samples
Test_features=[];
for i=1:n_a_vectors
a=1;
b=mode1;
for j= 1:nTest
Test_features{j,i}=[Test(a:b,:)*a_vectors(:,i)]';
a=a+mode1;
b=b+mode1;
end
end

% (Step six)Classification

DistTotal1=zeros(nTrain,nTest);
for i=1:n_a_vectors
for j=1:nTest
f_Test=Test_features{j,i};
f_Train=Train_features{i};
reptest = repmat(f_Test,nTrain,1);
   Dif = reptest - f_Train;
   Dif2 = Dif.^2;
dist = sum(Dif2,2);
dist1 = sqrt(dist);
dist2(:,j)=dist1/max(dist1); %normalização dos eixos
end
Dist{i}=dist2;
DistTotal1=[DistTotal1+(dist2.^2)];
end
DistTotal=(DistTotal1.^0.5);
for i=1:nTest
   Dist_testi=DistTotal(:,i);
for j=1:max(Group_Train)
indexTrain{j} = find(Group_Train==j);
Dist_testic(j,i)=mean(Dist_testi(indexTrain{j},:));
end
end
   [~,Class_Test]=min(Dist_testic);
Class_Test=Class_Test';
  Class_dif=Group_Test-Class_Test;
   [~,b]=find(Class_dif==0);
   CC=sum(b);
```

```
errors=nTest-CC;
   CCR=(CC/nTest)*100;

disp(['Número de acertos de Teste:   ' num2str(CC)])
disp(['Taxa de Classificação Correta:   ' num2str(CCR)])
end
```

Apêndice 2 – Artigo publicado (Apresentação do 2D-LDA)

Analytica Chimica Acta 938 (2016) 53–62

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Two-dimensional linear discriminant analysis for classification of three-way chemical data

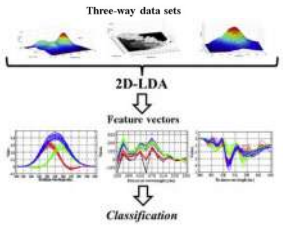
Adenilton C. da Silva [a], Sófacles F.C. Soares [a, b], Matías Insausti [c], Roberto K.H. Galvão [d], Beatriz S.F. Band [c], Mário César U. de Araújo [a, *]

[a] Universidade Federal da Paraíba, Departamento de Química, Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), Caixa Postal 5093, CEP 58051-970, João Pessoa, PB, Brazil
[b] Departamento de Engenharia Química, Centro de Tecnologia (CT), Universidade Federal da Paraíba, 58051-900, João Pessoa, PB, Brazil
[c] FIA Laboratory, Analytical Chemistry Section, INQUISUR (UNS-CONICET), Av. Alem 1253, B8000CPB, Bahía Blanca, Buenos Aires, Argentina
[d] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, 12228-900, São José dos Campos, SP, Brazil

HIGHLIGHTS

- Use of 2D-LDA for extraction of classification features from three-way chemical data.
- Case studies involving simulated data and real-life data sets: Parma ham and edible vegetable oils.
- Use of surface autofluorescence and total synchronous fluorescence spectrometries.
- Better results compared with the use of spectral data with no feature extraction.
- Better results compared with PLS Discriminant Analysis applied to the unfolded data, as well as PARAFAC-LDA and TUCKER3-LDA.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form
15 July 2016
Accepted 4 August 2016
Available online 20 August 2016

Keywords:
Two-dimensional linear discriminant analysis
PARAFAC-LDA
TUCKER3-LDA
Three-way fluorescence data
Dry-cured Parma ham
Edible vegetable oil

ABSTRACT

The two-dimensional linear discriminant analysis (2D-LDA) algorithm was originally proposed in the context of face image processing for the extraction of features with maximal discriminant power. However, despite its promising performance in image processing tasks, the 2D-LDA algorithm has not yet been used in applications involving chemical data. The present paper bridges this gap by investigating the use of 2D-LDA in classification problems involving three-way spectral data. The investigation was concerned with simulated data, as well as real-life data sets involving the classification of dry-cured Parma ham according to ageing by surface autofluorescence spectrometry and the classification of edible vegetable oils according to feedstock using total synchronous fluorescence spectrometry. The results were compared with those obtained by using the spectral data with no feature extraction, U-PLS-DA (Partial Least Squares Discriminant Analysis applied to the unfolded data), and LDA employing TUCKER-3 or PARAFAC scores. In the simulated data set, all methods yielded a correct classification rate of 100%. However, in the Parma ham and vegetable oil data sets, better classification rates were obtained

* Corresponding author.
E-mail address: laqa@quimica.ufpb.br (M.C.U. Araújo).

http://dx.doi.org/10.1016/j.aca.2016.08.009
0003-2670/Published by Elsevier B.V.

Apêndice 3 – Artigo publicado (Aplicação do 2D-LDA)

Talanta xxx (2017) xxx-xxx

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com

Differentiation of cumin seeds using a metal-oxide based gas sensor array in tandem with chemometric tools

Mahdi Ghasemi-Varnamkhasti[a,*], Zahra Safari Amiri[a], Mojtaba Tohidi[a], Majid Dowlati[b], Seyed Saeid Mohtasebi[c], Adenilton C. Silva[d], David D.S. Fernandes[d], Mário C.U. Araujo[d]

[a] Department of Mechanical Engineering of Biosystems, Faculty of Agriculture, Shahrekord University, Shahrekord, Iran
[b] Department of Mechanical Engineering of Biosystems, Faculty of Agriculture, University of Jiroft, Jiroft, Iran
[c] Department of Agricultural Machinery Engineering, Faculty of Agricultural Engineering and Technology, University of Tehran, Karaj, Iran
[d] Universidade Federal da Paraíba, Departamento de Química, Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), Caixa Postal 5093, 58051-970 João Pessoa, PB, Brazil

ARTICLE INFO

Keywords:
Cumin seeds
Electronic nose
Classification
2D-LDA
U-PLS-DA

ABSTRACT

Cumin is a plant of the Apiaceae family (umbelliferae) which has been used since ancient times as a medicinal plant and as a spice. The difference in the percentage of aromatic compounds in cumin obtained from different locations has led to differentiation of some species of cumin from other species. The quality and price of cumin vary according to the specie and may be an incentive for the adulteration of high value samples with low quality cultivars. An electronic nose simulates the human olfactory sense by using an array of sensors to distinguish complex smells. This makes it an alternative for the identification and classification of cumin species. The data, however, may have a complex structure, difficult to interpret. Given this, chemometric tools can be used to manipulate data with two-dimensional structure (sensor responses in time) obtained by using electronic nose sensors. In this study, an electronic nose based on eight metal oxide semiconductor sensors (MOS) and 2D-LDA (two-dimensional linear discriminant analysis), U-PLS-DA (Partial least square discriminant analysis applied to the unfolded data) and PARAFAC-LDA (Parallel factor analysis with linear discriminant analysis) algorithms were used in order to identify and classify different varieties of both cultivated and wild black caraway and cumin. The proposed methodology presented a correct classification rate of 87.1% for PARAFAC-LDA and 100% for 2D-LDA and U-PLS-DA, indicating a promising strategy for the classification different varieties of cumin, caraway and other seeds.

1. Introduction

The cumin flowering plant from the Apiaceae family (umbelliferae) is considered as one of the most important spices in the world and is the principal compound of mixed spices and curries [1,2]. Two types of consumed cumin in the world include cumin with scientific name of Bunium Persicum Boisswhich, so called black caraway, and cumin with the scientific name of Cuminum cyminum L. [3]. Seeds of cumin are used in the food industry as flavoring in bread, cheese, sweets, meat products, sauces, and beverages and the essence of this aromatic plant are also used in cosmetics and toiletries, toothpaste, chewing gum, and pharmaceuticals [4,5]. Black caraways seeds are used as well for sterilization of tissues of the body which have undergone surgery and for the production of some agricultural and veterinary drugs.

Nowadays Iran is one of the major exporters of black caraway, providing 20–40% of the world's market needs [6]. Green cumin is another Iranian economically important medicinal plants with health benefits such as anti-seizure, anti-epileptic, stomach re-enforcement, anti-flatulence, and anti-indigestion and has antioxidant and anti-cancer properties [7]. Iran prepares about 40% of the world's cumin consumption. Black caraways are produced in central and western Asia and southeastern Europe and cumin in the Mediterranean and South-West and Central Asia.

Cumin seeds obtained from different regions have different proportions of volatile organic compounds (VOCs) in their ingredients. Due to the limited areas with appropriate conditions for the growth of cumin

* Corresponding author.
Email address: ghasemymahdi@gmail.com (M. Ghasemi-Varnamkhasti)

https://doi.org/10.1016/j.talanta.2017.08.024
Received 11 May 2017; Received in revised form 3 August 2017; Accepted 6 August 2017
Available online xxx
0039-9140/

Anexo -1

Corcondia:

$$\text{CORCONDIA} = 100 * \left(1 - \frac{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}(g_{def} - h_{def})^2}{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}h_{def}^2}\right)$$

onde g_{def} é o elemento da matriz central calculada com o Tucker-3 a partir dos pesos do PARAFAC, h_{def} é o elemento de um tensor binário contendo valores um na superdiagonal e zero nas demais posições e F é o número de fatores do modelo.

PRODUTO DE KRONECKER:

O produto de Kronecker, também conhecido como produto tensorial e representado pelo símbolo "Ä", foi proposto pelo alemão L. Kronecker, no século XIX. Caracteriza-se como um operador matricial binário, transformando duas matrizes de dimensões arbitrárias em uma matriz de dimensão maior, com uma estrutura especial de bloco. Dadas as matrizes A n x m, de dimensões n x m, e B p x q, de dimensões p x q:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{bmatrix}_{nxm} \qquad \mathbf{B} = \begin{bmatrix} b_{1,1} & \cdots & b_{1,q} \\ \vdots & \ddots & \vdots \\ b_{p,1} & \cdots & b_{p,q} \end{bmatrix}_{pxq}$$

O produto de Kronecker, definido por **A** Ä **B,** é uma matriz de dimensões np x mq, com a estrutura de bloco dada por:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & \cdots & a_{n,m}\mathbf{B} \end{bmatrix}_{npxmq}$$

Equações de acordo com a referência:M. M. Sena, M. G. Trevisan, R. J. Poppi.PARAFAC: *Uma ferramenta quimiométrica para tratamento de dados multidimensionais. Aplicações na determinação direta de fármacos em plasma humano por espectrofluorimetria*. **Química Nova**, 28 (2005) 910-920.

Anexo -2

1- Participação em Eventos científicos:

- 37° Reunião Anual da Sociedade Brasileira de Química, 26 a 29 de Maio de 2014, Natal - RN. (Participação e apresentação de trabalho).
- 18° ENQA | Encontro Nacional de Química Analítica, 18 a 21 de Setembro de 2016, Florianópolis – SC.(Participação, apresentação de trabalho e miniconferência).

2- Atividades em outras instituições:

- Curso sobre Hyperspectral Image Analysis, 35 h, UFPE. Período: 17 a 21 de Março de 2014.
- Complementação do tema de pesquisa do doutorado no ITA (Instituto tecnológico da Aeronáutica). Supervisão: Professor Dr. Roberto Kawakami Harrop Galvão (Div. Engenharia eletrônica). Período: 11 de Novembro a 4 de Dezembro 2015.

3- Produções cientifícas:

K. D. T. M. Milanez, A. C. Silva, J. E. M. Paz, E. P. Medeiros, M. J. C. Pontes. S*tandardization of NIR data to identify adulteration in ethanol fuel*.**Microchemical Journal**, 124(2016) 121–126.

A. C. Silva, S. F. C. Soares, M. Insausti, R. K. H. Galvao, B. S. F. Band, M. C. U. Araujo. *Two-dimensional linear discriminant analysis for classification of three-way chemical data*. **Analytica Chimica Acta**, 938 (2016), 53-62.

M. Ghasemi-Varnamkhasti, Z. S. Amiri, M. Tohidi, M. Dowlate, S. S. Mohtasebi, A. C. Silva, D. D. S. Fernandes, M. C. U. Araújo, *Differentiation of cumin seeds using a metal-oxide based gas sensor array in tandem with chemometric tools*, **Talanta** (in press) 2017.