UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

ALGORITMO INSPIRADO NOS MORCEGOS PARA SELEÇÃO DE VARIÁVEIS EM PROBLEMAS DE CLASSIFICAÇÃO

JULIANA DA CRUZ SOUZA

João Pessoa – PB – Brasil
Março / 2023

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

ALGORITMO INSPIRADO NOS MORCEGOS PARA SELEÇÃO DE VARIÁVEIS EM PROBLEMAS DE CLASSIFICAÇÃO

JULIANA DA CRUZ SOUZA*

Tese de Doutorado apresentada como requisito para obtenção do título de Doutora em Ciências pela Universidade Federal da Paraíba. **Área de concentração:** Química Analítica.

Orientador: Prof. Dr. Edvan Cirino da Silva.

Co-orientador: Prof. Dr. Sófacles Figueredo Carreiro Soares.

***Bolsista (CAPES)**

João Pessoa – PB – Brasil
Março / 2023

Catalogação na publicação
Seção de Catalogação e Classificação

S719a Souza, Juliana da Cruz.
Algoritmo inspirado nos morcegos para seleção de variáveis em problemas de classificação / Juliana da Cruz Souza. - João Pessoa, 2023.
145 f. : il.

Orientação: Edvan Cirino da Silva.
Coorientação: Sófacles Figueredo Carreiro Soares.
Tese (Doutorado) - UFPB/CCEN.

1. Química analítica. 2. Bioinspiração. 3. Análise discriminante linear. 4. Classificação multivariada. I. Silva, Edvan Cirino da. II. Soares, Sófacles Figueredo Carreiro. III. Título.

UFPB/BC CDU 543(043)

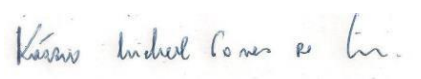
Elaborado por GRACILENE BARBOSA FIGUEIREDO - CRB-15/794

Algoritmo inspirado nos morcegos para seleção de variáveis em problemas de classificação.

Tese de Doutorado apresentada pela aluna Juliana da Cruz Souza e aprovada pela banca examinadora em 03 de março de 2023.

Prof. Dr. Edvan Cirino da Sila
DQ/UFPB
Orientador

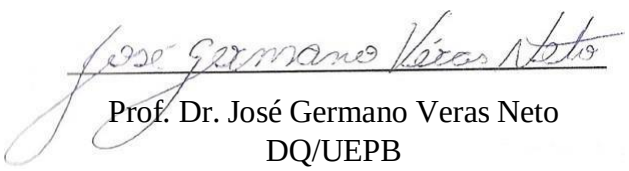
Prof. Dr. Sófacles Figueredo Carreiro Soares
CT/UFPB
Co-orientador



Prof. Dr. Kássio Michell Gomes de Lima
IQ/UFRN
Examinador

Prof. Dr. Clarimar José Coelho
Escola Politécnica/PUC-GO
Examinador

Prof. Dr. Sherlan Guimarães Lemos
DQ/UFPB
Examinador



Prof. Dr. José Germano Veras Neto
DQ/UEPB
Examinador

Dedico este trabalho à minha mãe Luzia.

AGRADECIMENTOS

A Deus pela vida, saúde e força para seguir em frente nos momentos mais difíceis.

À minha mãe Luzia Cruz, minha irmã Jucélia Cruz, minha prima Jéssica Itaiane e meu noivo Raul Alves pelo incentivo, carinho e apoio em todos os momentos.

Aos meus amigos do LAQA (Jainny, Ana Rosa, Wallis, Thyago, Daniella, Kelly, Laila e Carla) por me receberem carinhosamente e pela amizade.

Aos meus orientadores Prof. Dr. Edvan Cirino da Silva e Dr. Sófacles Figueredo Carreiro Soares por todo apoio, conhecimento compartilhado, orientação e esforços aplicados no desenvolvimento do trabalho.

Aos professores Mário Ugulino (UFPB) e Clarimar José Coelho (Pontifícia Universidade Católica de Goiás) pelas contribuições para publicação do artigo.

Ao laboratório de Automação e Instrumentação em Química Analítica e Quimiometria (LAQA).

Às professoras Drª Andréa Monteiro e Drª Elaine Cristina da Unidade Acadêmica de Serra Talhada/Universidade Federal Rural de Pernambuco, por me incentivarem e apoiarem na vida acadêmica.

À Kevla por disponibilizar os dados de cafés.

Ao programa de pós-graduação em química da Universidade Federal da Paraíba.

À CAPES pela bolsa concedida.

"*The mind that opens to new ideas never returns to its original size*"

Albert Einstein.

RESUMO

Título: **"Algoritmo Inspirado nos Morcegos para Seleção de Variáveis em Problemas de Classificação"**

Autora: **Juliana da Cruz Souza**

O uso da Análise Discriminante Linear (LDA) em modelagem de classificação multivariada permite a construção de modelos no domínio dos dados originais, o que possibilita a realização de inferência química direta dos resultados. Entretanto, essa técnica requer uma baixa dimensionalidade dos dados e produz modelos com problemas de generalização quando existe uma alta multicolinearidade entre as variáveis. Para superar esses problemas, o uso de algoritmos de seleção de variáveis tem se mostrado muito eficiente especialmente quando dados UV-Vis, NIR, etc, são usados. Nesse contexto, o uso de algoritmos bio-inspirados (a exemplo do algoritmo genético-GA) tem permitido a realização bem-sucedida de seleção de variáveis. No presente trabalho, propõe-se o algoritmo inspirado no comportamento dos morcegos (Bat Algorithm-BA) para a seleção de variáveis em modelagem via LDA. O algoritmo proposto, denominado aqui BA-LDA, utiliza uma função de custo associada ao risco médio de classificação incorreta (G_{cost}), a qual foi implementada no código do seu programa escrito em Matlab. O desempenho do BA-LDA foi avaliado em quatro estudos de caso, envolvendo o emprego de dados espectrométricos de massas (MS), NIR, UV-Vis e em dados com informação simulada. Para cada conjunto de dados analisados, os parâmetros do BA-LDA foram otimizados usando um planejamento fatorial fracionário 2^{4-1}. Os dados MS foram provenientes de análises de 216 amostras de soro de pacientes com e sem câncer de ovário. Os dados NIR foram obtidos na análise de 60 amostras de cafés pertencentes a duas classes (*gourmet* e tradicionais). Para obtenção de dados UV-Vis, foram registrados espectros de amostras de óleos vegetais pertencentes a quatro classes, a saber: soja, canola, milho e girassol. Para o estudo com uma classe de amostras simuladas, foram empregados dados NIR de diesel. O desempenho do BA-LDA foi comparado ao obtido com os algoritmos GA-LDA e SPA-LDA usados para seleção de variáveis e com as técnicas de análise discriminante por mínimos quadrados parciais (PLS-DA) e modelagem independente e flexível por analogia de classe (SIMCA). O algoritmo proposto selecionou 11, 3, 7 e 9 variáveis e obteve as taxas de classificação correta (TCC) de 93, 100, 100 e 100 % na classificação baseada nos dados de, respectivamente, MS, NIR, UV-Vis e da classe simulada (NIR). No conjunto de dados MS, o BA-LDA superou o desempenho do SPA-LDA (79,1 % de TCC) e GA-LDA (88.4 % de TCC), porém foi inferior ao do algoritmo PLS-DA que apresentou 98% de TCC. Para os demais conjuntos de dados, a performance do BA-LDA foi comparável ao desempenho dos algoritmos clássicos. Em todos os estudos de caso, o BA-LDA superou o desempenho do SIMCA. Ademais, o BA-LDA se mostrou menos susceptível ao ruído adicionado aos espectros das amostras de teste do conjunto de dados simulados. Visto que o BA-LDA é estocástico, seu principal diferencial é a convergência e robustez que demonstrou em todos os conjuntos de dados, nos quais as variáveis selecionadas possibilitaram uma interpretação química segura.

Palavras-chave: Bioinspiração. Classificação Multivariada. Análise Discriminante Linear.

ABSTRACT

Title: "Bat-Inspired Algorithm for Variable Selection in Classification Problems"

Author: Juliana da Cruz Souza

The use of Linear Discriminant Analysis (LDA) in multivariate classification modeling allows the construction of models in the domain of the original data, in which a direct chemical inference of the results may be accomplished. However, this technique requires a low dimensionality of the data and produces models with generalization problems when there is a high multicollinearity among the variables. To overcome these drawbacks, the use of variable selection algorithms has proved to be very efficient especially when UV-Vis, NIR, etc, data are used. In this context, bio-inspired algorithms (such as the genetic algorithm-GA) have allowed the successful selection of variables. In the present work, a bat-inspired algorithm (BA) for selection variables in modeling via LDA is proposed. This algorithm, here named BA-LDA, uses a cost function associated with the average risk of misclassification (G_{cost}), which was implemented in its code written in Matlab. The performance of BA-LDA was evaluated in four case studies, involving the use of mass spectrometric (MS), NIR, and UV-Vis data, as well as a dataset with simulated information. For each analyzed dataset, the BA-LDA parameters were optimized using a 2^{4-1} fractional factorial design. MS data were resulting of analyzes of 216 serum samples from patients with and without ovarian cancer. The NIR data were acquired in analysis of 60 coffee samples belonging to two classes (gourmet and traditional). UV-Vis data were obtained from recorded spectra of vegetable oil samples belonging to four classes, namely: soybean, canola, corn and sunflower. For the study with a class of simulated samples, diesel NIR data were employed. The performance of BA-LDA was compared to those obtained with the GA-LDA and SPA-LDA algorithms used for variable selection; it was also compared to the partial least squares discriminant analysis (PLS-DA) and independent and flexible modeling by class analogy (SIMCA). The proposed algorithm selected 11, 3, 7 and 9 variables and obtained correct classification rates (TCC %) of 93, 100, 100 and 100% in the classification based on data from MS, NIR, UV-Vis and of the simulated class (NIR). In the case of MS data, BA-LDA outperformed SPA-LDA (79.1% TCC) and GA-LDA (88.4% TCC), but was lower than the PLS-DA algorithm that showed a TCC of 98%. For the other datasets, the BA-LDA performance was comparable to the classical algorithms. In all case studies, BA-LDA outperformed SIMCA. Furthermore, the BA-LDA proved to be less susceptible to noise added to the spectra of the test samples from the simulated dataset. Since the BA-LDA is stochastic, its main differential is the convergence and robustness that it demonstrated in all data sets, in which the selected variables allowed a safe chemical interpretation.

Keywords: Biospiration. Multivariate Classification. Linear Discriminant Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BA – "*Bat Algorithm"* - Algoritmo dos morcegos

LDA – "*Linear Discriminant Analysis"* - Análise discriminante linear

G_{cost} – "*Average risk of misclassification*" - Risco médio de uma classificação incorreta pela LDA

g_k– "*Risk of misclassification of object xk of the kth validation sample*" - Risco de uma classificação incorreta do objeto x_k da k-ésima amostra de validação

GA – "*Genetic Algorithm*" - Algoritmo genético

NIR – "*Near Infrared*" - Região do infravermelho próximo

TCC – Taxa de classificação correta

UV-Vis – "*Visible Ultraviolet*" - Região ultravioleta visível

MS– "*Mass Spectrometry*" - Espectrometria de massas

PCA – "*Principal Component Analysis*" - Análise de Componentes Principais

PCs – "*Principal Component*" - Componentes principais

MLR – "*Multiple linear regression*" - Regressão linear múltipla

PLS – "*Partial least squares regression*" - Regressão por mínimos quadrados parciais

SIMCA – "*Soft Independent Modeling of Class Analogy*" - Modelagem independente e flexível por analogia de classe

ACO – "*Ant Colony Algorithm*" - Algoritmo colônia de formigas

HS – "*Harmonic Search Algorithm*" - Algoritmo de busca harmônica

IWO – "*Invasive weed optimization algorithm*" - Algoritmo de otimização invasiva de ervas daninhas

SPA – "*Successive Projection Algorithm*" - Algoritmo das projeções sucessivas

SUMÁRIO

INTRODUÇÃO

Capítulo 1

CAPÍTULO 1: INTRODUÇÃO

1. Apresentação da problemática e proposta de solução

A busca pela distinção entre classes ou grupos de amostras é um problema frequente em diversas áreas do conhecimento. Artigos relatando a investigação de adulteração e/ou falsificação de alimentos e bebidas (JAMWAL *et al.*, 2020, 2021; VALINGER *et al.*, 2021; VISCONTI; RODRÍGUEZ; ANIBAL, 2020; WENG *et al.*, 2020), verificação de falsificações de documentos e outras aplicações voltadas para problemas de classificação forenses (FARID *et al.*, 2021; MAJDA *et al.*, 2018; SHARMA, V. *et al.*, 2018), distinção de soro de pacientes com e sem câncer (XU *et al.*, 2017), classificação de solos (CHAUHAN *et al.*, 2021), entre outros, são amplamente reportados na literatura. Dentre as técnicas empregadas para realizar análises com o objetivo de classificar amostras, se destacam, com elevada aplicabilidade, as espectroanalíticas, como as espectrometrias de Infravermelho Próximo (NIR) e Médio (MID), Ultravioleta Visível (UV-Vis), Raman e Imagens Hiperespectrais (AMJAD *et al.*, 2018; CEBI *et al.*, 2019; JAMWAL *et al.*, 2020, 2021; JOLAYEMI; AJATTA; ADEGEYE, 2018; KARUNATHILAKA *et al.*, 2019; PAOLETTI *et al.*, 2018; VALINGER *et al.*, 2021; VISCONTI; RODRÍGUEZ; ANIBAL, 2020; WENG *et al.*, 2020; YVES *et al.*, 2021).

Em geral, as técnicas espectroanalíticas, principalmente a espectroscopia de infravermelho, apresentam vantagens como o baixo custo, análises rápidas e não invasivas e/ou destrutivas (LOHUMI *et al.*, 2015). Todavia, esses dados costumam apresentar alta dimensionalidade e multicolinearidade. Para lidar com dados de natureza multivariada em problemas de classificação, técnicas quimiométricas baseadas na compressão de dados - por exemplo, a análise discriminante de mínimos quadrados parciais bem estabelecida (*Partial least squares regression*, PLS-DA) (AZCARATE, Silvana Mariela *et al.*, 2013; BONIFAZI *et al.*, 2021; SILVA, N. C. D. *et al.*, 2015) - foram empregadas. No entanto, os modelos PLS-DA resultantes não permitem uma interpretação química mais fácil e direta porque as variáveis são transformadas. Nesse contexto, a análise discriminante linear (*Linear Discriminant Analysis*-LDA) (CAI; LIU, W., 2011; SAFO; AHN, 2016; WITTEN; TIBSHIRANI, 2011) é uma ferramenta de classificação alternativa que opera no domínio de dados originais.

Em suma, a LDA tem por objetivo encontrar funções lineares das variáveis que podem ser usadas para classificação (VARMUZA; FILZMOSER, 2009). Apesar de suas vantagens, especialmente no que concerne aos cálculos e à simplificação dos resultados de classificação, a LDA é restrita a problemas de baixa dimensão. Ademais, essa técnica requer que o número de

amostras seja superior ao número de variáveis a serem incluídas no modelo (VARMUZA; FILZMOSER, 2009). Em face desses problemas, estudos têm sido realizados a fim de desenvolver metodologias quimiométricas para seleção de variáveis em modelagem de classificação multivariada (NEMA; THAKUR, 2015; PONTES, A. S. *et al.*, 2020; PONTES, C. *et al.*, 2005; SEM, 2021). Esses trabalhos apresentaram, quanto aos resultados mais expressivos, uma melhora significativa na classificação de amostras com a realização da seleção de variáveis. Sendo assim, reforçam a necessidade de desenvolvimento de algoritmos e estratégias para seleção de variáveis em modelagem de classificação baseada em LDA. Nesse contexto, a seleção de variáveis possibilita a redução da dimensionalidade e a multicolinearidade presentes nos dados no domínio original. Como resultado, torna-se possível a construção de modelos quimiométricos mais simples e fáceis de interpretar do ponto de vista químico e/ou físico.

Algoritmos bioinspirados tem sido desenvolvidos e aplicados com sucesso para resolver diferentes problemas de otimização e realizar a seleção de variáveis. Alguns algoritmos inspirados na natureza são: O algoritmo de enxame de partículas (PSO), proposto por Kennedy e Eberhart para otimização (KENNEDY; EBERHART, 1995) e adaptado para o problema de seleção de variáveis para classificação usando análise discriminante linear (LDA) (LIN, S.; CHEN, S., 2009); O algoritmo genético (GA) desenvolvido por John Holland para solucionar problemas complexos de otimização (FILHO, A. C.; POPPI, R.J., 1999) e utilizado em diferentes problemas de seleção de variáveis (CHENG, J.; SUN; PU, 2016; RIBEIRO, L. A. *et al.*, 2015; SHEYKHIZADEH; NASERI, 2018); O algoritmo colônia de formigas (ACO) desenvolvido por Dorigo e Stutzle (DORIGO; STÜTZLE, 2004) para otimização e adaptado para seleção de variáveis em calibração usando regressão linear múltipla (MLR) (ZHANG, Y. *et al.*, 2019) e classificação usando LDA (PONTES, A. S. *et al.*, 2020) ; E o algoritmo de ervas daninha (IWO) desenvolvido para seleção de variáveis em modelagem de calibração usando MLR e classificação usando LDA (SHEYKHIZADEH; NASERI, 2018).

Um algoritmo bioinspirado que tem se destacado para solucionar problemas complexos de otimização é o algoritmo dos morcegos (BA- *Bat Algorithm*). Esse algoritmo simula o mecanismo natural de ecolocalização utilizado pelos morcegos ao se moverem em busca de presas e desviarem dos objetos ao redor. Desde sua formulação (YANG, Xin-She, 2010a), várias versões do BA foram propostas, aprimoradas e utilizadas para solucionar problemas diversos (LIU, Q. *et al.*, 2018; MESSAOUDI; KAMEL, 2019; NADERI; KHAMEHCHI; KARIMI, 2019; NIU *et al.*, 2018; YANG, Q.; DONG; ZHANG, J., 2021; YUE; ZHANG, H.,

2020). Os algoritmos inspirados nos morcegos, em geral, são simples e de fácil implementação (ZHU, B. *et al.*, 2016). A principal característica dos BAs desenvolvidos é uma rápida convergência para soluções ótimas (MESSAOUDI; KAMEL, 2019; ZHU, B. *et al.*, 2016) e desempenho superior, em problemas de otimização (MIRJALILI; MOHAMMAD, 2014; NADERI; KHAMEHCHI; KARIMI, 2019; WANG, G.; GUO, L., 2013), quando comparado a outros algoritmos estocásticos como o GA e PSO (MIRJALILI; MOHAMMAD, 2014; WANG, G.; GUO, L., 2013). Além destas vantagens o BA foi formulado com um mecanismo de busca local que possibilita a procura por melhores soluções quando o morcego virtual se depara com uma solução que não é desejável. Assim, o BA é um algoritmo promissor para solucionar problemas de naturezas distintas desde que seja corretamente adaptado.

1.1 Apresentação do trabalho e composição do texto da tese

No presente trabalho, propõe-se uma modificação do algoritmo inspirado no comportamento dos morcegos – concebido por X. YANG, em 2010 – de modo a ser explorado, pela primeira vez, para seleção de variáveis em modelagem de classificação via LDA. Assim, o algoritmo modificado é aqui denominado BA-LDA, que é o acrônimo da expressão em inglês "*Bat Algorithm - Linear Discriminant Analysis*". Esse algoritmo utiliza uma função de custo associada ao risco médio G_{cost} de classificação incorreta em LDA, a qual foi adotada e implementada no código do seu programa.

A multicolinearidade e alta dimensionalidade – que são, geralmente, encontradas nas variáveis associadas a dados multivariados espectrométricos e que prejudicam a eficiência e capacidade de generalização da LDA – foram superadas com o BA-LDA. Para esse propósito, os morcegos virtuais usados nesse algoritmo possibilitam a minimização da multicolinearidade mediante a seleção das variáveis que levem a um menor valor para G_{cost}. Ademais, o algoritmo proposto reduz a dimensionalidade, importante para o melhor condicionamento da matriz dos dados em LDA, eliminando as variáveis redundantes ou não informativas.

O texto da presente tese apresenta, no **Capítulo 2**, a fundamentação do trabalho na qual são descritas as principais técnicas quimiométricas de tratamentos de dados; são também descritas técnicas de classificação baseadas em seleção de variáveis envolvendo algoritmos estocásticos, cujos fundamentos são abordados com o apoio da literatura. O **Capítulo 3** descreve o algoritmo dos morcegos original e a reformulação proposta nesta tese e o **Capítulo 4** apresenta a metodologia adotada no desenvolvimento do trabalho. O BA-LDA

proposto foi avaliado em quatro estudos de caso, cujos resultados são apresentados e discutidos no **Capítulo 5**. Os quatro estudos de caso são descritos, sucintamente, nos parágrafos a seguir.

O primeiro estudo de caso permitiu avaliar a capacidade do BA-LDA em distinguir amostras de soro de pacientes com e sem câncer de ovário. Os dados foram provenientes de um espectrômetro de massas de alta resolução. A finalidade desse estudo foi testar o desempenho do BA-LDA frente aos dados espectrométricos de massas.

No segundo estudo de caso, o BA-LDA foi aplicado para discriminar amostras reais pertencentes a duas classes de cafés, a saber: o *gourmet* e o tradicional utilizando a técnica espectrométrica de infravermelho próximo (NIR). O objetivo deste estudo foi testar a capacidade do BA-LDA de lidar com espectros de refletância NIR muito semelhantes e seleção de variáveis em um espaço com dimensionalidade elevada (1301 variáveis).

No terceiro estudo de caso, o BA-LDA foi avaliado a partir de dados espectrométricos UV-Vis, para discriminar amostras de óleos vegetais pertencentes a quatro classes distintas, a saber, girassol, canola, milho e soja. Nesta aplicação, o papel do BA-LDA foi avaliado em termos de sua capacidade de lidar com espectros de baixa resolução e forte sobreposição resultante de bandas de absorção largas nas regiões de UV-Vis.

No quarto estudo de caso, o BA-LDA foi aplicado em dados NIR de amostras de diesel, assim como em dados com informação simulada em que toda a variabilidade (inclusive envolvendo a linha de base e ruído) foi mantida. A finalidade deste estudo foi testar a capacidade do BA-LDA em distinguir as amostras inteiramente reais das amostras com a inclusão da banda espectral de baixa intensidade. Ainda no **Capítulo 5**, apresenta-se o resultado de um estudo da inserção de ruído nos dados de diesel com a informação simulada a fim de avaliar o desempenho do BA-LDA.

Para todos esses conjuntos de dados o desempenho de classificação do BA-LDA foi comparado com o dos métodos tradicionais: Modelagem independente e flexível por analogia de classe (Soft Independent Modeling of Class Analogy – SIMCA) e PLS-DA; comparou-se também com o de algoritmos de seleção de variáveis, sendo um determinístico SPA-LDA e um estocástico GA-LDA. Outro aspecto interessante deste trabalho diz respeito à avaliação da robustez do algoritmo BA-LDA, envolvendo os conjuntos de dados acima mencionados. Para efeito de comparação, a robustez do GA-LDA também foi avaliada.

1.2 Objetivos

1.2.1 Geral

Reformular o algoritmo dos morcegos para seleção de variáveis em classificação multivariada via LDA.

1.2.2 Específicos

(i) Reformulação matemática do algoritmo dos morcegos de modo a introduzir a função custo apropriada para avaliar o risco médio do erro de classificação e guiar o processo de seleção de variáveis em análise discriminante linear (LDA).

(ii) Implementar o código do programa do algoritmo proposto BA-LDA em linguagem MatLab ® 2011b.

(iii) Otimizar os parâmetros de entrada do BA-LDA.

(iv) Demonstrar a eficiência do BA-LDA em estudos de casos de dados reais provenientes de diferentes técnicas (dados espectrométricos nas regiões NIR, UV-Vis e dados de espectrometria de massas) envolvidos na classificação de amostras de:

- Cafés gourmet e tradicionais (NIR);
- Óleos vegetais (soja, canola, girassol e milho) (UV-Vis);
- Soro de pacientes com e sem câncer ovário (MS);

(v) Demonstrar a eficiência do BA-LDA em um estudo de caso com amostras reais e simuladas. Esse estudo envolveu a classificação de:

- Amostras reais de diesel e amostras com banda espectral simulada.

(vi) Comparar o desempenho do BA-LDA com a classificação SIMCA, com o algoritmo PLS-DA e com os algoritmos de seleção de variáveis SPA-LDA e GA-LDA, popularmente conhecidos no campo da análise multivariada.

(vii) Avaliar a sensibilidade do BA-LDA ao ruído introduzido nos dados em um estudo com os dados de diesel e dados simulados.

(viii) Avaliar a robustez do BA-LDA e comparar com o GA-LDA.

FUNDAMENTAÇÃO

Capítulo 2

CAPÍTULO 2: FUNDAMENTAÇÃO

2. Fundamentação Teórica

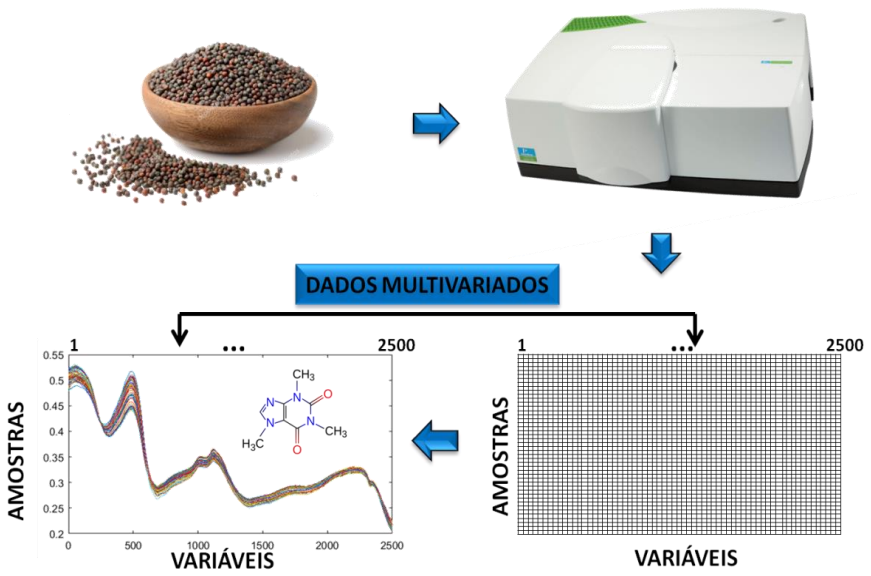
Neste capítulo, serão descritos os fundamentos que embasam o trabalho iniciando-se pela explanação sobre as técnicas e análises que proporcionam os dados multivariados. Nesse contexto, descreve-se o papel da aplicação das técnicas quimiométricas usadas e, posteriormente, dos principais tratamentos de dados assim como a seleção de amostras. Ainda neste capítulo, são descritos os fundamentos da classificação multivariada, as principais métricas usadas para classificação e a seleção de variáveis. Apresenta-se também uma breve revisão bibliográfica sobre algoritmos estocásticos empregados no contexto da seleção de variáveis e uma descrição detalhada do algoritmo genético acoplado a LDA.

2.1 Análise multivariada e a quimiometria

A análise multivariada pode ser definida como o estudo de múltiplas variáveis aleatórias que são relacionadas entre si ou que possuem conjuntos de relações de modo que os diferentes efeitos dessas variáveis não podem ser interpretados separadamente (HAIR *et al.*, 2009). Na atualidade esse tipo de estudo é fundamental, tendo em vista que com desenvolvimento da tecnologia a obtenção de dados em dimensões cada vez mais elevadas se tornou frequente em diversas áreas do conhecimento (BAQUETA *et al.*, 2021; HAASE; ARROYO; TREJOS, 2020; JANNAT *et al.*, 2018; VALINGER *et al.*, 2021).

Na química analítica, técnicas de natureza multivariada como as espectroanalíticas são amplamente empregadas (ARAÚJO, T. K. L. *et al.*, 2021; LÓPEZ-MAESTRESALAS *et al.*, 2018; SHENG; MIAW; MARTINS, M.; *et al.*, 2018; SHENG; MIAW; SENA; *et al.*, 2018). A **Figura 1** mostra uma representação de medidas espectrométricas de infravermelho e a obtenção dos espectros que apresentam vasta quantidade de informações (2500 variáveis) por cada amostra analisada.

Figura 1-Ilustração da análise de amostras por espectrometria de infravermelho e a obtenção de dados multivariados.



Fonte: (própria).

No exemplo da **Figura 1**, os dados após coletados pelo equipamento, são convertidos em uma matriz **X** (*I* x *J*) contendo um arranjo ordenado de linhas *I* e colunas *J*. Cada linha representa uma amostra *I* analisada, e cada amostra está associada a um vetor linha que contém os valores das *J* medidas feitas. Esses valores correspondem as variáveis, que são representadas pelas colunas da matriz, ou seja, cada coluna *J* contém os valores de uma medida para todas as amostras. Para o exemplo ilustrado com a técnica espectrométrica supracitada, cada variável está associada a um comprimento de onda ou número de onda. Assim, para este exemplo, as medidas realizadas levaram a obtenção de uma matriz de dados contendo duas mil e quinhentas variáveis. A obtenção dessa vasta quantidade de informações é comum com o emprego de diferentes técnicas espectroanalíticas. Porém, a matriz **X** com os vários sinais numéricos ou mesmo as bandas e picos presentes no espectro não indicam diretamente as informações necessárias para realizar associações entre as amostras e/ou entre as propriedades de interesse destas. Para extrair tais informações, pode-se recorrer as técnicas estatísticas multivariadas disponíveis na quimiometria (FERREIRA, 2015).

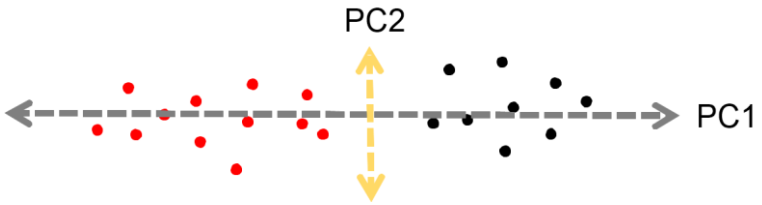
A quimiometria pode ser definida como uma disciplina da química que emprega técnicas estatísticas e matemáticas para extrair as informações (físicas e/ou químicas) mais relevantes presentes nos dados e selecionar procedimentos e experimentos ótimos (VARMUZA; FILZMOSER, 2009). O desenvolvimento das ferramentas quimiométricas veio para suprir a dificuldade da interpretação dos dados multivariados. Ao mesmo passo, com

auxilio dessas ferramentas ampliou-se o uso de tecnologias cada vez mais informativas (ARAÚJO, T. K. L. *et al.*, 2021; LIN, S.; CHEN, S., 2009; PONTES, A. S. *et al.*, 2020; PONTES, C. *et al.*, 2005; SHEYKHIZADEH; NASERI, 2018). Um dos grandes campos de estudo da química analítica envolvendo técnicas multivariadas e aplicação de ferramentas quimiométricas é a classificação multivariada. Neste campo, as ferramentas quimiométricas são usadas para atribuir amostras à determinadas classes ou para formar limites entre as classes de amostras e possibilitar a discriminação entre elas. Essas ferramentas quimiométricas podem auxiliar desvendando relações ocultas entre os sinais analíticos o que favorece a discriminação das classes de amostras (PANCHUK *et al.*, 2018). Na **Seção 2.3** os fundamentos da classificação multivariada são abordados, mas antes disto será apresentada a Análise de Componentes Principais (PCA). A PCA será discutida antes, pois, ela é usada na modelagem de classificação Modelagem SIMCA (apresentada na **Seção 2.3**), mas sozinha não se caracteriza como um método de classificação, e sim como um método de reconhecimento de padrões não supervisionado.

2.2 Análise de Componentes Principais

A Análise de Componentes Principais (*Principal Componente Analysis*, PCA) é um método de reconhecimento de padrões não supervisionado que realiza a redução de dimensionalidade dos dados (VARMUZA; FILZMOSER, 2009). Para isso, as informações semelhantes são agrupadas em novas variáveis (ou PCs) a partir de combinações lineares das variáveis originais. As novas variáveis (PCs) são projetadas de maneira ortogonal entre si, assim não portam informações redundantes. A projeção das informações em PCs, ocorre de acordo com a variabilidade dos dados originais. Assim, a primeira (PC1) é projetada na direção de maior variabilidade dos dados. A PC2 é projetada de maneira ortogonal à PC1 e na direção de segunda maior variância dos dados. As outras PCs seguem o mesmo padrão, sendo projetadas de maneira ortogonal as anteriores e descrevendo as máximas variâncias restantes. Dessa forma, as informações contidas nos dados podem ser completamente descritas em poucas PCs (FERREIRA, Márcia M. C., 2015). A **Figura 2**, representa a projetação da PC1 e PC2 conforme procedimento descrito.

Figura 2-Ilustração da direção de projeção da PC1 e PC2 nos dados.



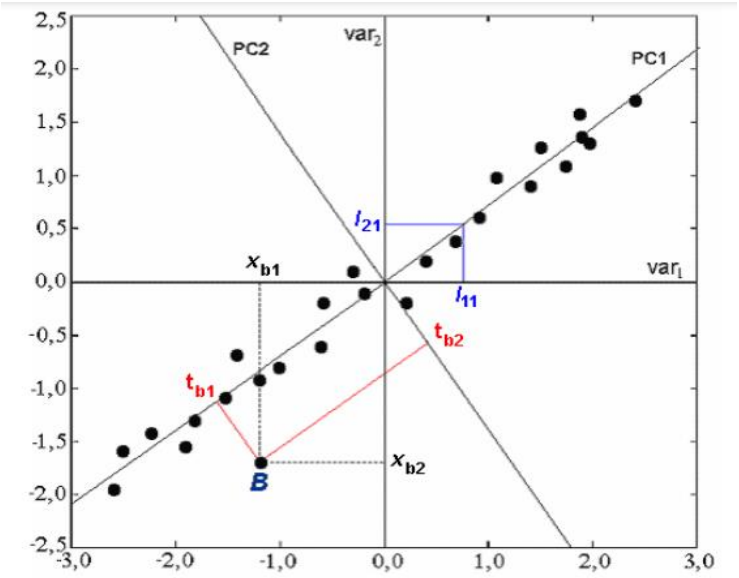
Fonte: (própria)

Para as amostras representadas na **Figura 2**, a direção que descreve a maior variabilidade, e, portanto, a PC1, é representada pela seta em cinza. A partir da projeção das amostras nesta PC é simples diferenciar as amostras vermelhas das pretas. Já na PC2, imaginando que todas as amostras são "arrastadas" em direção a seta amarela, as amostras vermelhas e pretas estariam juntas com algumas sobrepostas. Isso significa que a PC1 porta informações capaz de discriminar as amostras enquanto a PC2 traz outro tipo de variabilidade presente nos dados (FERREIRA, Márcia M. C., 2015).

Matematicamente, a PCA pode ser descrita como a decomposição da matriz de dados **X** (*I* x *J*) em duas, uma de pesos **L** e uma de escores **T,** resultando em $\mathbf{X}=\mathbf{T}\mathbf{L}^{T}$. Sendo a matriz de escores relacionadas às amostras e a matriz de pesos às contribuições das variáveis (FERREIRA, Márcia M. C., 2015). A partir das colunas de **L**, as direções dos eixos das PCs são traçadas e na sequência as amostras são projetadas nesses eixos. Essa projeção das amostras é descrita pela matriz de escores **T** que é representada por **T**=**XR**. Dessa forma, **R** é a matriz de transformação que converte o domínio original dos dados para o domínio das PCs. Uma vez que $\mathbf{X}=\mathbf{T}\,\mathbf{L}^{T}$ (ou **T=X L**) e **T**=**X R**, pode-se inferir que a matriz de pesos (**L**) é a própria matriz de transformação (**R**).

Como as colunas de **L** projetam as direções das PCs e nem todas essas direções precisam ser usadas para representar os dados, evita-se a correlação entre as variáveis (entre essas colunas). Ferreira (2015), descreve essas primeiras PCs como *A* que representa a dimensionalidade intrínseca dos dados. Assim, tendo *A* (componentes principais significativas), as demais PCs irão descrever os resíduos (**E**) e uma nova equação representará os dados, sendo esta, $\mathbf{X}=\widehat{\boldsymbol{X}}+\mathbf{E}$ com $\widehat{\mathbf{X}} = \mathbf{T}_A\,\mathbf{L}_A^{T}$. Assim, as matrizes de escores e pesos são, respectivamente, $\mathbf{T}_A$ (*I* x *A*) e $\mathbf{L}_A$ (*J* x *A*) (FERREIRA, Márcia M. C., 2015). Cada coluna de $\mathbf{L}_A$ representa, o peso (ou a contribuição) de cada variável na formação da PC. Os pesos são cossenos dos ângulos entre o eixo da PC e os eixos originais das variáveis. A representação da transformação pode ser verificada na **Figura 3**.

Figura 3-Ilustração da projeção das amostras nas PCs.



Fonte: (FERREIRA, Márcia M. C., 2008).

onde, var_1 e var_2 são os eixos das variáveis originais e x_{b1} e x_{b2} as coordenadas da amostra B nesses eixos; t_{b1} e t_{b2} são as novas coordenadas de B no eixos das PCs, ou seja, são os escores. Já l_{11} e l_{21} são os pesos das variáveis 1 e 2 na formação da PC1.

Como as primeiras PCs agregam quase toda a informação presente nos dados, deixando de fora as informações redundantes, a PCA consegue reduzir a dimensionalidade dos dados. Na **Seção 2.3** serão abordados os fundamentos da classificação multivariada, dentre estes o uso da PCA na metodologia de classificação SIMCA.

2.3 Fundamentos da classificação multivariada

A classificação multivariada, também conhecida como método de reconhecimento de padrões supervisionado, é caracterizada por empregar um conjunto de amostras com propriedades distintas (grupos ou classes) conhecidas, a fim de construir modelos matemáticos para realizar a classificação de amostras desconhecidas (FERREIRA, Márcia M. C., 2015). Para essa finalidade, funções que separam as amostras (ou classificadores) de acordo com as semelhanças entre elas, são utilizadas. Assim, quanto mais semelhantes são as amostras em relação as variáveis medidas mais próximas elas se encontram no espaço multidimensional (FERREIRA, Márcia M. C., 2015). Cada grupo composto por amostras semelhantes pode ser categorizado por um índice de classe com o intuito de distingui-los. Com isso, amostras desconhecidas podem ou não serem associadas aos grupos previamente estabelecidos.

Em geral, as técnicas de classificação utilizam as amostras conhecidas dividindo-as em dois subconjuntos, isto é, o de treinamento e validação que são usados, respectivamente, na construção e validação dos modelos. A avaliação do modelo é então realizada por meio de testes e conceitos estatísticos que são estabelecidos para discriminação dos grupos. Com isso, é possível atestar se a atribuição das classes das amostras externas foi realizada de forma correta, devendo apresentar o menor número possível de classificações incorretas para confiabilidade dos testes estatísticos (VARMUZA; FILZMOSER, 2009). Dentre as técnicas mais comuns na química para classificação encontram-se a modelagem SIMCA e duas técnicas de análise discriminantes, a Análise Discriminante de Quadrados Mínimos Parciais (*Partial least squares regression*- PLS-DA) e Análise Discriminante Linear (LDA). Antes da descrição de tais técnicas é importante conhecer as principais métricas da classificação.

2.3.1 Métricas de classificação

- *Matriz de confusão*

A matriz de confusão é uma tabela que mostra os resultados da classificação das amostras considerando as classes verdadeiras (que ficam dispostas nas colunas) e as classes estimadas pelo modelo (que ficam dispostas nas linhas). Os valores contidos nestas tabelas são o número de amostras de cada classe dispostas no local onde foram classificadas (FERREIRA, Márcia M. C., 2015). Uma representação de uma matriz de confusão para duas classes de amostras pode ser vista na **Figura 4**.

Figura 4-Representação da matriz de confusão para duas classes de amostras.

		Classe verdadeira	
		A	**B**
Classe prevista	**A**	AA (VP)	BA (FP)
	B	AB (FN)	BB (VN)
	Nenhuma		

Fonte: (própria).

Para duas classes hipotéticas (A e B), AA indica o número de amostras que são da classe A e que foram previstas na classe A. Da mesma forma, o quadrante com BB indica o

número de amostras que são verdadeiramente da classe B e que foram previstas na classe B. Assim, a diagonal da matriz de confusão apresenta as amostras que foram corretamente classificadas. Quando observa-se a primeira linha das atribuições, onde temos BA, será adicionado o número de amostras da classe B que foram incorretamente previstas como classe A. Na linha seguinte temos o quadrante com AB, que apresenta o número das amostras da classe A incorretamente classificadas como classe B. A última linha da matriz de confusão é usada quando existem amostras que não foram atribuídas a nenhuma das classes (FERREIRA, Márcia M. C., 2015). Essa matriz de confusão é usada principalmente para interpretação dos resultados obtidos pelo método SIMCA, que possibilita observar todas essas situações descritas.

Ainda observando a **Figura 4**, quando a amostra B é classificada incorretamente como A (primeira linha- quadrante BA) esse erro é conhecido como ***tipo I*** e indica um falso positivo (FP). Quando essa amostra é classificada corretamente em sua classe (BB), temos um verdadeiro negativo (VN). Quando a amostra A é incorretamente classificada como B (segunda linha – quadrante AB) é considerado erro do ***tipo II***, e este pode ser definido como falso negativo (FN). Já quando a amostra A é classificada como A (primeira linha – AA) temos um verdadeiro positivo (VP) (FERREIRA, Márcia M. C., 2015).

Após definidos esses parâmetros (FP, VP, VN e FN), é possível calcular as seguintes métricas de classificação: Exatidão ou Taxa de classificação correta, Sensitividade e a Seletividade.

- *Taxa de classificação correta (TCC)*

A taxa de classificação correta ou exatidão é uma das principais métricas para avaliar o desempenho da classificação e é definida como uma razão entre as amostras classificadas corretamente e o número total de amostras, conforme **Equação 1 e Figura 4**.

$$TCC = \frac{AA+BB}{AA+AB+BA+BB}100 = \frac{VP+VN}{VP+FN+FP+VN}100 \qquad (1)$$

(FERREIRA, Márcia M. C., 2015).

- *Sensitividade*

A Sensitividade (SEN) também conhecida como taxa de verdadeiros positivos é expressa em porcentagem conforme **Equação 2**.

$$SEN = \frac{AA}{AA+AB}100 = \frac{VP}{VP+FN}100 \quad (2)$$

onde AA são as amostras da classe A classificadas corretamente e AB são as amostras da classe A classificadas incorretamente como pertencentes a classe B (**Figura 4**).

- *Seletividade*

A Seletividade (SEL) também denominada de especificidade, representa as amostras que não pertencem a classe A e que foram classificadas corretamente como não pertencentes a classe A, **Equação 3**.

$$SEL = \frac{BB}{BA+BB}100 = \frac{VN}{FP+VN}100 \quad (3)$$

Assim, a partir da matriz de confusão é possível calcular as principais métricas de classificação. A **Seção 2.3.2** descreve o método de classificação SIMCA.

2.3.2 Modelagem independente e flexível por analogia de classe

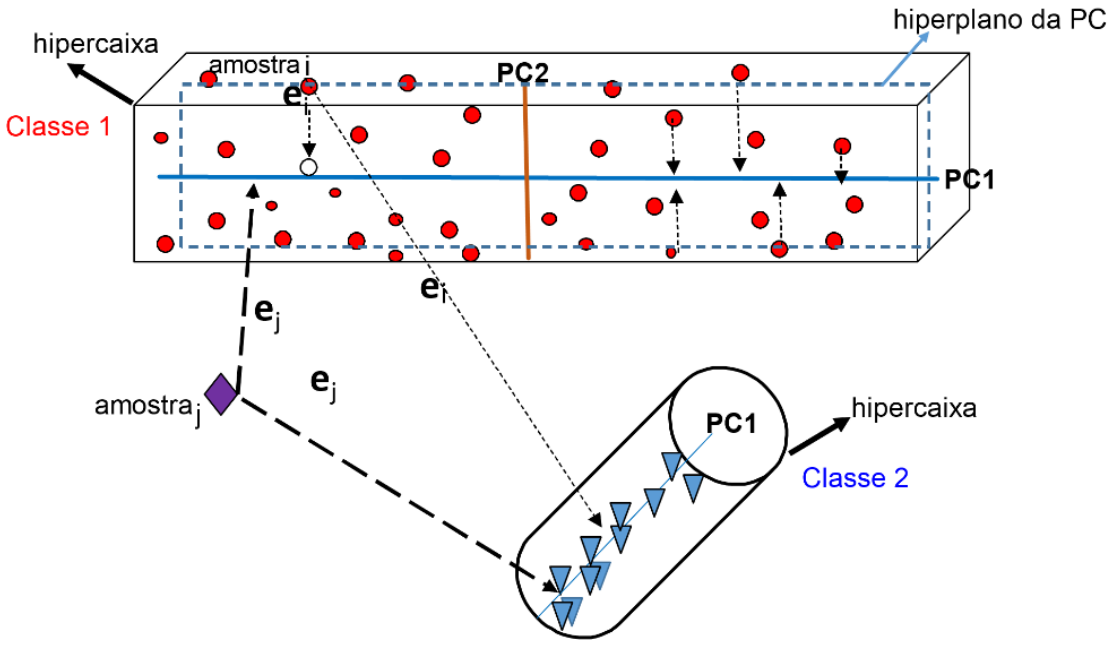
O método de classificação Modelagem independente e flexível por analogia de classe-(*Soft Independent Modelling of Class Analogies*, SIMCA) usa PCA para modelar e descrever a estrutura de cada grupo dos dados multivariados em um espaço de menor dimensão (VARMUZA; FILZMOSER, 2009). Assim, para cada grupo ou classe de amostras do conjunto de treinamento, a PCA é aplicada separadamente com um número de PCs determinado também por grupo, para isso, pode-se realizar a validação cruzada nos dados de treinamento (VARMUZA; FILZMOSER, 2009). Dessa forma, uma classe pode necessitar, por exemplo, de duas componentes principais enquanto a outra pode ser descrita com apenas uma.

Para cada grupo, a PCA é delimitada em uma forma (podendo ser retangular, redonda, quadrada, entre outras) que representa a projeção (nas PCs) das amostras pertencentes ao grupo. Na **Figura 5,** para a classe 1 as PCs (PC1 e PC2) formam um retângulo. Os limites da forma ou hiperplano gerado pelas PCs podem ser definidos pelo desvio padrão dos escores nas PCs, mas outros métodos também são relatados (FERREIRA, Márcia M. C., 2015). Para cada hiperplano das PCs, uma "caixa" com todas as amostras da classe é definida (**Figura 5**). Ou seja, o espaço dos resíduos **E** (informações não modeladas nas PCs) é definido. A classificação SIMCA de uma amostra de treinamento ocorre comparando a variância residual dessa amostra

com a variância residual média da classe. Com um *teste-F* é possível definir uma valor limite da variância residual da classe em questão, delimitando uma "hipercaixa" no espaço complementar ao das PCs (FERREIRA, Márcia M. C., 2015). Na **Figura 5** a hipercaixa para classe 1 é representada na forma de um paralelepípedo e para a classe 2 na forma de um cilindro.

Para atribuição de amostras externas à classe, avalia-se a projeção da amostra no espaço das PCs e a distância das fronteiras da classe (delimitadas pela "hipercaixa"). Isso é feito para todas as classes e após a comparação a amostra vai ser associada estatisticamente a uma classe, a outra classe, em ambas as classes ou a nenhuma delas. O último caso ocorre quando a variância residual da amostra é superior a todas as variâncias residuais das amostras de treinamento delimitadas nas classes (FERREIRA, Márcia M. C., 2015). A **Figura 5** representa esse processo descrito, sendo uma amostra desconhecia *j* representada na cor roxa e sua respectiva variância residual representada por e_j para ambas as classes apresentadas, assim, esta amostra não é classificada em nenhuma das classes. Ainda na **Figura 5**, para a amostra *i* (círculo vermelho), é possível verificar que a variância residual da mesma está contida nos limites da classe 1 e é muito superior quando comparada ao modelo da classe 2. Assim, a amostra *i* é classificada como pertencente a classe 1.

Figura 5-Ilustração da modelagem SIMCA.



Fonte: (própria).

Matematicamente, um *teste-F* é usado para definir se a amostra está distante do hiperplano das PCs (FERREIRA, Márcia M. C., 2015). O valor $F_{crítico}$ é o valor máximo de *F*

para que a amostra seja classificada dentro da classe. Assim, primeiramente calcula-se a variância residual da amostra S_i (**Equação 4**) e a variância residual média da classe S_0 **(Equação 5**).

$$S_i^c = \sqrt{\frac{\sum_{j=1}^{J}(e_j^c)^2}{J-A_c}} \tag{4}$$

$$S_0^c = \sqrt{\frac{\sum_{i=1}^{Nc}\sum_{j=1}^{J}(e_{ij}^c)^2}{(N_c-A_c-1).(J-A_c)}} \tag{5}$$

onde c representa a classe, *i* as amostras e *j* as variáveis, e_{ij} é o resíduo da *i-ésima* amostra na *j-ésima* variável, com N_c sendo o número de amostras de treinamento da classe *c* e A_c o número de componentes principais (FERREIRA, Márcia M. C., 2015; PONTES, M. J. C., 2009).

A soma na **Equação 4** é realizada apenas para as variáveis, enquanto na **Equação 5** ela é aplicada também as amostras. Na sequência, emprega-se o *teste-F* conforme **Equação 6**, e verifica se o desvio padrão residual da amostra é significativamente maior que o desvio padrão residual médio da classe.

$$F_{cal} = \frac{(S_i^c)^2}{(S_0^c)^2} \cdot \frac{N_c}{N_c-A_c-1} \tag{6}$$

Para um determinado nível de significância o valor *F* calculado pode ser comparado ao tabelado. Assim, quando se tem um F_{cal} menor que o $F_{crítico}$, a amostra pode ser associada a classe em estudo. Para uma visão mais aprofundada da classificação SIMCA a bibliografia pode ser consultada (FERREIRA, Márcia M. C., 2015; VARMUZA; FILZMOSER, 2009).

2.3.3 Análise Discriminante de Mínimos Quadrados Parciais

A análise discriminante pelo método dos Quadrados Mínimos Parciais (*Partial least squares regression*, PLS-DA) é uma técnica tradicionalmente empregada em problemas de classificação (AZCARATE, Silvana Mariela *et al.*, 2013; BONIFAZI *et al.*, 2021; SANTANA *et al.*, 2020; SILVA, N. C. D. *et al.*, 2015). Para usar o PLS-DA é necessário definir a matriz de variáveis dependentes **Y** (com vetores ***y***), que contenha os índices das classes. Após definida a matriz de variáveis dependentes o modelo de regressão do PLS é construído. O modelo PLS

usa variáveis latentes para descrever a variabilidade dos dados e relaciona os escores da matriz de dados (**X**) com a matriz dos índices das classes (FERREIRA, Márcia M. C. *et al.*, 1999). Assim, as variáveis latentes são traçadas nas direções que maximizam a separação entre as classes. Existem duas técnicas PLS-DA, a PLS1-DA e a PLS2-DA (SANTANA *et al.*, 2020). Na técnica PLS1-DA, um modelo é construído para cada classe. Assim, para classe A constrói-se uma matriz **Y** com valores iguais a 1 e para as demais classes usa-se o valor 0. Para classe B, outro modelo é construído, usando 1 para as amostras da classe e 0 para as demais classes e assim esse processo continua para todas as classes. No PLS2-DA, um único modelo é calculado obtendo um único conjunto de escores e pesos para todas as classes, assim, usa-se o mesmo número de variáveis latentes na modelagem das classes (SANTANA *et al.*, 2020). Aqui descreveremos apenas o PLS2-DA.

No PLS2-DA, o algoritmo NIPALS (*non linear iterative partial least squares*) pode ser usado para decomposição das matrizes **X** e **Y** em escores (**T** e **U**) e pesos (**P** e **Q**), respectivamente (SANTANA *et al.*, 2020). Sendo $\mathbf{R}_x$ e $\mathbf{R}_y$, os resíduos de **X** e **Y** respectivamente, devido as informações não modeladas nas variáveis latentes. As **Equações 7** e **8** mostram tal decomposição.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{R_x} \tag{7}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{R_y} \tag{8}$$

Proporcional a covariância entre **X** e **Y**, são calculados os pesos (**W**) do PLS (**Equação 9**), onde as colunas de **W** representam as direções das variáveis latentes, e calcula-se a matriz de escores **T** a partir da combinação linear de **X** e **W** (**Equação 10**)(SANTANA *et al.*, 2020).

$$\mathbf{W} = \frac{\mathbf{X^T U}}{||\mathbf{X^T U}||} \tag{9}$$

$$\mathbf{T} = \mathbf{XW} \tag{10}$$

Os pesos **P** e **Q** e os escores **U** são calculados conforme as **Equações 11**, **12** e **13** (SANTANA *et al.*, 2020):

$$\mathbf{Q} = \frac{\mathbf{Y^T T}}{||\mathbf{U^T T}||} \tag{11}$$

$$\mathbf{P} = \frac{\mathbf{X^T T}}{||\mathbf{T^T T}||} \tag{12}$$

$$\mathbf{U} = \mathbf{YQ} \tag{13}$$

Com isso, as matrizes dos resíduos podem ser estimadas. Para estimativa dos coeficientes de regressão do PLS utiliza-se a **Equação 14** e finalmente calcula-se os valores de $\mathbf{Y}_{previsto}$ ($\widehat{\mathbf{Y}}$), que designam a classe à qual pertence a amostra, conforme **Equação 15** (SANTANA *et al.*, 2020):

$$\mathbf{B}_{\mathrm{PLS}} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{Q}^{\mathrm{T}} \tag{14}$$

$$\widehat{\mathbf{Y}} = \mathbf{X}.\mathbf{B}_{\mathrm{PLS}} \tag{15}$$

Os valores dos índices de classes estimados pelo modelo são aproximados aos valores inicialmente definidos. Assim, para duas classes inicialmente codificadas como 1 para a classe de interesse (classe 1) e 0 para classe 2, estimativas próximas a 1 serão atribuídas a classe 1 e estimativas próximas a 0 serão atribuídas a classe 2 (FERREIRA, Márcia M. C., 2015). Conforme Ferreira (2015), o limite entre as classes pode ser estimado a partir de uma combinação entre as funções de densidade de probabilidade e a teoria Bayesiana. Para uma classe C, uma amostra *i* tem sua estimativa pelo modelo PLS como $\hat{y}_i$ e a **Equação 16** demonstra a probabilidade P(C|$\hat{y}_i$) da amostra *i* pertencer a classe C. A probabilidade da amostra pertencer à todas as outras classes, representada pela classe D, é dada por P(D|$\hat{y}_i$).

$$P(C|\hat{y}_i) = \frac{p(\hat{y}_i|C) \times P(C)}{p\left(\hat{y}_i\middle|C\right) \times P(C) + p(\hat{y}_i|D) \times P(D)\big)};$$

$$P(D|\hat{y}_i) = \frac{p(\hat{y}_i|D) \times P(D)}{p(\hat{y}_i|C)\times P(C)+p(\hat{y}_i|D)\times P(D)}. \tag{16}$$

onde P(C) e P(D), são as probabilidades de ocorrência das amostras pertencerem às classes C e D, respectivamente, conforme **Equações 17** e **18** (FERREIRA, Márcia M. C., 2015).

$$P(C) = \frac{I_C}{I_C+I_D} \tag{17}$$

$$P(D) = \frac{I_D}{I_C+I_D} \tag{18}$$

Para cada classe uma distribuição normal é ajustada e a média ($\bar{\hat{y}}$) e o desvio padrão (*s*) dos valores previstos pode ser usado para estimar as funções de densidade de probabilidade para classe C ($p(\hat{y}_i|C)$) e para classe D ($p(\hat{y}_i|D)$), conforme **Equação 19**.

$$p(\hat{y}_i|C) = \frac{1}{s_C\sqrt{2\pi}}\, e^{\frac{-1}{2}\left(\frac{\hat{y}_i-\bar{\hat{y}}}{s_C}\right)^2};$$

$$p(\hat{y}_i|D) = \frac{1}{s_D\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{\hat{y}_i - \bar{\hat{y}}}{s_D}\right)^2}. \quad (19)$$

Se P(C|$\hat{y}_i$) > P(D|$\hat{y}_i$), então a amostra é da classe C. Se P(D|$\hat{y}_i$) > P(C|$\hat{y}_i$), então a amostra é da classe D (FERREIRA, Márcia M. C., 2015). A soma das probabilidades é ajustada para 1 conforme denominador da **Equação 16**.

Uma simplificação demonstrada por Ferreira (2015) foi: se $p(\hat{y}_i|C) \times P(C) > p(\hat{y}_i|D) \times P(D)$ a amostra pertence à classe C; caso contrário a amostra pertence à classe D. O limiar de classificação é obtido pela **Equação 20**.

$$p(\hat{y}|C) \times P(C) = p(\hat{y}|D) \times P(D) \quad (20)$$

Para determinar o número ideal de variáveis latentes (VL) no PLS-DA usa-se a validação cruzada no conjunto de treinamento. Assim, os modelos são construídos com diferentes números de VL e as amostras separadas para validação são testadas nestes modelos. Os resultados são armazenados até que todas as amostras sejam testadas. Ferreira (2015) indica que uma maneira de escolher o número de VL é observar a sensitividade e seletividade na validação cruzada, sendo o número ótimo de VL aquele que levar as maiores sensitividades e seletividades. Santana *et al*. (2020) usa a análise da porcentagem de amostras classificadas corretamente nas amostras de validação cruzada, assim, o número de VL definido é o do modelo que levou a melhor classificação (SANTANA *et al.*, 2020). Após determinado o número de VL, o modelo é aplicado na classificação de amostras externas, cujas classes devem ser previamente conhecidas para avaliar o desempenho do modelo.

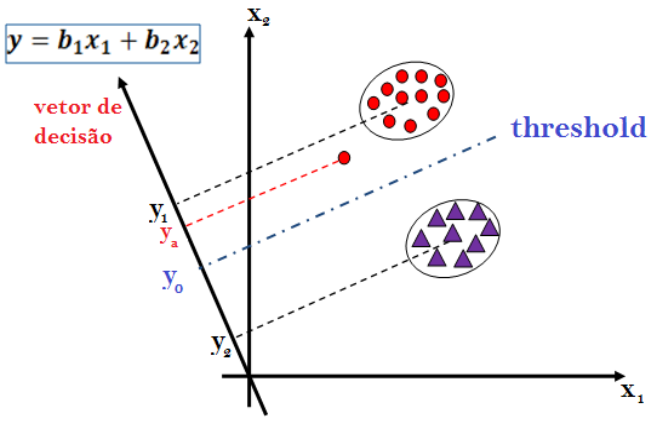
Apesar de bem estabelecida, a técnica PLS-DA gera modelos que não permitem uma interpretação química mais fácil porque as variáveis são transformadas. Nesse contexto, a análise discriminante linear (LDA) (CAI; LIU, W., 2011; SAFO; AHN, 2016; WITTEN; TIBSHIRANI, 2011) é uma ferramenta de classificação alternativa que opera no domínio original dos dados. Obviamente, a LDA permite uma interpretação mais fácil e direta dos resultados.

2.3.4 Análise Discriminante Linear - LDA

Dentre os métodos de reconhecimento de padrões supervisionados, a Análise Discriminante Linear (LDA) é uma modelagem tradicionalmente empregada na classificação de dados químicos (LIN, S.; CHEN, S., 2009; PONTES, A. S. *et al.*, 2020; PONTES, C. *et al.*,

2005; RIBEIRO, L. A. *et al.*, 2015; SOUTO *et al.*, 2010). A LDA usa funções discriminantes lineares que visam maximizar a variância entre as classes e minimizar a variância dentro da classe (PONTES, M. J. C., 2009). Para discriminação dos grupos ou classes de amostras, a análise discriminante linear emprega, em geral, a abordagem de *Fisher*. Segundo esta abordagem, os dados químicos de natureza multivariada são projetados em escores de forma a separar os grupos transformados o máximo possível (VARMUZA; FILZMOSER, 2009). Para duas classes, os objetos dos grupos são projetados na variável na direção de maior separação (vetor de decisão **y**) que fornecerá os escores discriminantes (y_1, y_2,..., y_n) para as classes 1 e 2, como ilustrado na **Figura 6**.

Figura 6- Representação do princípio da LDA para duas classes de objetos ou amostras.



Fonte: (própria).

O critério para separação dos grupos é dado pela diferença das médias dos escores de cada classe dividido pela raiz quadrada da variância agrupada (que corresponde à soma ponderada das variâncias das classes 1 e 2 conforme as **Equações 21** e **22**.

$$\frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \rightarrow max \tag{21}$$

$$s_y^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2} \tag{22}$$

onde $\bar{y}_1$ e $\bar{y}_2$ são médias aritméticas dos escores discriminantes de cada grupo e S_y é a raiz quadrada da variância agrupada.

O vetor de separação, que maximiza a separação das classes, foi dado por *Fisher* como:

$$\boldsymbol{b}_{FISHER} = \boldsymbol{S}_P^{-1}(\overline{\boldsymbol{x}}_{\boldsymbol{1}} - \overline{\boldsymbol{x}}_{\boldsymbol{2}}) \tag{23}$$

onde $\overline{\boldsymbol{x}}_{\boldsymbol{1}}$ e $\overline{\boldsymbol{x}}_{\boldsymbol{2}}$ são os vetores médios aritméticos dos dados dos grupos 1 e 2 e $\mathbf{S}_P$ é a matriz de covariância agrupada dado pela **Equação 24** (VARMUZA; FILZMOSER, 2009).

$$\boldsymbol{S}_P = \frac{(n_1-1)\,\boldsymbol{S}_1+\cdots+(n_k-1)\,\boldsymbol{S}_k}{n_1+\cdots+n_k-k} \tag{24}$$

Dessa forma, uma amostra desconhecida x_a é classificada a partir do seu escore discriminante (y_a na **Figura 6**) por meio da projeção na direção definida pelo vetor de decisão, usando a **Equação 25**. O escore y_a é comparado com um limiar de classificação (*threshold* – **Figura 6**) que é o centro entre as médias dos grupos determinado pela **Equação 26**. Portanto a amostra x_a é classificada, de acordo com a ilustração na **Figura 6**, como pertencente à classe 1.

$$\boldsymbol{y}_a = \boldsymbol{b}_{FISHER}^T \boldsymbol{x}_a \tag{25}$$

$$y_0 = \frac{\boldsymbol{b}_{FISHER}^T \overline{\boldsymbol{x}}_1 + \boldsymbol{b}_{FISHER}^T \overline{\boldsymbol{x}}_2}{2} \tag{26}$$

Apesar de simples, a LDA é restrita a problemas de baixa dimensão e que apresentam número de amostras superior ao número de variáveis a serem incluídas no modelo (VARMUZA; FILZMOSER, 2009). Quando essas características não estão presentes nos dados, a matriz de covariância (**S**) torna-se singular não sendo possível calcular a matriz inversa (**Equação 23**). Desta forma, os modelos LDA não podem ser diretamente aplicados para os dados multivariados como os espectrométricos que, usualmente, apresentam alta dimensionalidade.

Além disso, a LDA pode ser prejudicada quando ocorre o problema da multicolinearidade (FERREIRA, Márcia M. C., 2015), usualmente presente em dados espectrométricos. A multicolinearidade pode ser explicada como uma inter-relação entre as variáveis independentes (COIMBRA *et al.*, 2005), assim, tais variáveis podem fornecer informações redundantes e prejudicar o modelo LDA. A solução para estes problemas (alta dimensionalidade e multicolinearidade) requer a aplicação de técnicas ou algoritmos de seleção de variáveis.

2.4 Seleção de variáveis

A seleção de variáveis pode ser definida como um processo de escolha de um subconjunto de elementos apropriados para a construção de modelos (GALVÃO, R. K. H.; ARAÚJO, M. C. U., 2009). As metodologias de seleção de variáveis em conjuntos de dados multivariados foram desenvolvidas pela necessidade de redução de regiões não informativas, ruidosas e que portavam informações redundantes (ANDERSEN; BRO, 2010). Essas características são comuns das técnicas mais utilizadas atualmente como as espectrométricas. Assim, a seleção de variáveis fornece os meios adequados de descartar tais informações, identificando em meio aos dados completos as variáveis que contribuem para os modelos. Ademais, a seleção de variáveis favorece a interpretação química dos espectros por operar no domínio dos dados originais, bem como possibilita o desenvolvimento, a partir de variáveis selecionadas, de instrumentos portáteis para serem usados em campo.

Diferentes metodologias são descritas na literatura para realizar a seleção de variáveis, sendo que as mais recentes têm demonstrado melhores desempenhos quando empregam algoritmos (PAULA, DE *et al.*, 2019; SHEYKHIZADEH; NASERI, 2018; ZHANG, D. *et al.*, 2021; ZHANG, Y. *et al.*, 2019) para essa finalidade. Os algoritmos apresentam vantagens para selecionar variáveis, pois são usualmente guiados por uma função-custo baseados em critérios e subconjuntos de variáveis que possibilitam a otimização dos modelos. Caso não promovam resultados satisfatórios, outro subconjunto de variáveis é testado até alcançar um critério de parada adotado no procedimento. Assim, os algoritmos são definidos como uma sequência de etapas para executar tarefas e encontrar soluções ótimas dos problemas (CORMEN, 2017) e podem apresentar uma natureza determinística ou estocástica.

2.4.1 Algoritmos determinísticos

Os algoritmos determinísticos são caracterizados por apresentarem sempre as mesmas soluções (subconjuntos de variáveis) após diferentes execuções, desde que aplicados sempre as mesmas condições iniciais (GOMES, A. A. *et al.*, 2021). Vários trabalhos foram reportados na literatura envolvendo a aplicação de algoritmos determinísticos para seleção de variáveis (CENTNER *et al.*, 1996; MEHMOOD *et al.*, 2011; PIERNA *et al.*, 2009; PONTES, C. *et al.*, 2005; RINNAN *et al.*, 2013). A reprodutibilidade do subconjunto de variáveis selecionado é uma vantagem destes algoritmos em relação aos estocásticos. Todavia, conforme Yang (2010), apesar de serem muito eficientes em encontrar ótimos locais, esses algoritmos podem ficar

presos nestas regiões principalmente quando se trata de funções multimodais (com muitos mínimos locais) (YANG, Xin-She, 2010b). A exemplo, pode-se destacar o uso da seleção de variáveis na busca por um subconjunto que leve a um melhor modelo a partir de dados multivariados.

Uma maneira de amenizar este problema de ótimos locais é introduzir no algoritmo um componente estocástico, tornando o algoritmo probabilístico. Assim, diferentes subconjuntos de variáveis podem ser avaliados de acordo com a heurística de cada estocástico na busca de variáveis que levem a soluções melhores para as modelagens. Além disso, conforme Yang (2010b), os algoritmos estocásticos podem produzir resultados finais que geralmente convergem para as mesmas soluções ótimas com uma dada precisão (YANG, Xin-She, 2010b).

Um algoritmo determinístico que desde a sua formulação (PONTES, C. *et al.*, 2005) tem se destacado com elevada aplicabilidade na classificação multivariada é o Algoritmo das Projeções Sucessivas SPA-LDA (BARBOSA *et al.*, 2018; CHEN, H.; TAN; LIN, Z., 2020; KHANMOHAMMADI; GARMARUDI; GUARDIA, 2013; SEM, 2021; SOARES, S. F. C. *et al.*, 2013). Nestes trabalhos, o SPA-LDA demonstrou como resultados mais expressivos, excelentes desempenhos de classificação. Assim, o SPA-LDA foi usado, nesta Tese, como método comparativo ao algoritmo proposto. Sendo assim, a **Seção 2.4.1.1** apresenta uma breve descrição do SPA-LDA.

2.4.1.1 Descrição do algoritmo determinístico usado: SPA-LDA

O Algoritmo das Projeções Sucessivas-SPA foi fundamentado tendo como objetivo a redução da multicolinearidade para realizar a seleção de variáveis. Para isso, usa-se a matriz das amostras de treinamento **X** (k_c x j) e realiza-se uma série de operações de projeções de vetores que minimizam a colinearidade. Conforme Pontes (2009), o algoritmo inicia pela *i-ésima* coluna da matriz de treinamento centrada na média ou auto-escalonada. Após definido o vetor inicial, em cada iteração o próximo vetor selecionado será projetado no subespaço ortogonal, que apresenta menor multicolinearidade em relação as variáveis já incluídas (PONTES, M. J. C., 2009).

Para construção dos subconjuntos de variáveis selecionadas, Pontes (2009) descreve seis etapas (PONTES, M. J. C., 2009):

- **Passo 1-** Inicialização:

$\mathbf{z}^1 = \mathbf{x}_j$

$\mathbf{x}^1_\mathbf{k} = x_k,\ \ k = 1,...,J$

$L\ (1, j) = j$

- **Passo 2 –** Cálculo da matriz de projeção P no subespaço ortogonal a z^1:

$$P^i = I - \frac{z^i (z^i)^T}{(z^i)^T z^i} \qquad (27)$$

com I sendo a matriz de identidade, com dimensões apropriadas.

- **Passo 3 -** Cálculo dos vetores projetados $\mathbf{x}_k^{i+1}$:

$$\mathbf{x}_k^{I+1} = \mathbf{P}^I \mathbf{x}_k^I \qquad (28)$$

para todos os k=1,... J.

- **Passo 4** – Determinar o índice k* do vetor de maior projeção e armazena-lo na matriz **VARSEL**.

$$\mathbf{k}^* = \arg\max_{k=1,\dots,J} \left|\left|\mathbf{x}_k^{I+1}\right|\right| \qquad (29)$$

$$\mathbf{VARSEL}(\mathbf{i}+1, \mathbf{j}) = \boldsymbol{k}^* \qquad (30)$$

- **Passo 5-** Fazer $\mathbf{z}^{i+1} = \mathbf{x}_{k^*}^{k+1}$ (vetor que define as operações de projeção para a iteração seguinte.

- **Passo 6 -** Fazer i=i+1. Se i<M volte para o **Passo 2**.

Sendo **M** as variáveis em cada subconjunto, cadeia de variáveis construídas de acordo com essas operações que foram descritas.

Após a geração dos subconjuntos de variáveis selecionadas pelo SPA, utiliza-se a função custo G_{cost} (**Equação 40** discutida na **Seção 3.3**) para escolher o melhor subconjunto, este será o que leva ao menor risco médio de classificação incorreta pela LDA (PONTES, M. J. C., 2009).

Tendo em vista que esta Tese teve o intuito de apresentar um novo algoritmo estocástico desenvolvido para seleção de variáveis para classificação LDA, a **Seção 2.4.2** apresenta uma

descrição dos algoritmos estocásticos e uma breve revisão literária sobre alguns algoritmos reportados.

2.4.2 Algoritmos estocásticos

Os algoritmos estocásticos podem ser definidos como processos que evoluem de maneira aleatória (TAYLOR; KARLIN, 1998) e fazem parte das técnicas de inteligência artificial que podem simular comportamentos naturais para o processamento de informações. Esses algoritmos são importantes para seleção de variáveis em matrizes multivariadas uma vez que permitem a interpretação dos dados complexos e a geração de resultados úteis para os modelos em curto espaço de tempo. As principais vantagens (YANG, Xin-She, 2010b) dos algoritmos estocásticos são destacadas abaixo:

- Não necessitam de conhecimento prévio do problema;
- Não há restrições para o ponto de partida no domínio do problema;
- Não são facilmente presos em mínimos locais;
- Armazenam as soluções e comparam com iterações seguintes para escolha da solução mais próxima da ideal;
- Fornecem maneiras gerais de procurar uma boa solução dentro de um período razoável de tempo (SHEYKHIZADEH; NASERI, 2018).

Embora sejam amplamente utilizados em problemas de otimização com funções complexas, esses algoritmos ainda encontram uso relativamente incipiente na química analítica, especialmente na seleção de variáveis em modelagem de classificação multivariada (PONTES, A. S. *et al.*, 2020; RIBEIRO, L. A. *et al.*, 2015; SHEYKHIZADEH; NASERI, 2018). Esse aspecto torna esse campo de pesquisa muito atrativo e com grande potencial para abordagem desse tipo de problema.

2.4.3 Algoritmos estocásticos e revisão da literatura

Algoritmos probabilísticos inspirados na natureza tem sido desenvolvidos e utilizados com frequência para otimização de problemas reais, visto que, partem de parâmetros aleatórios e comparam as soluções possíveis em busca da solução ótima. Na seleção de variáveis, o desenvolvimento de pesquisas com algoritmos estocásticos tem se mostrado crescente, especialmente para os problemas envolvendo a calibração multivariada (ATTIA *et al.*, 2016; PAULA, DE *et al.*, 2017, 2019; SHEYKHIZADEH; NASERI, 2018; YANG, M. *et al.*, 2017;

ZHANG, Y. *et al.*, 2019). Para esse propósito, algoritmos bioinspirados como o algoritmo dos vagalumes (FA) (ATTIA *et al.*, 2016), algoritmo dos vagalumes multi-objetivo (MOFA) (PAULA, DE *et al.*, 2017), enxame de partículas (PSO) (HU *et al.*, 2019) e colônia de formigas (ACO) (XIAOWEI *et al.*, 2014) tem demonstrado excelentes desempenhos principalmente quando comparados ao tradicional algoritmo genético (GA) (ATTIA *et al.*, 2016; PAULA, DE *et al.*, 2017; XIAOWEI *et al.*, 2014). No contexto da classificação multivariada, a utilização de algoritmos estocásticos para seleção de variáveis ainda é incipiente. A seguir serão descritos alguns algoritmos estocásticos que foram aplicados para seleção de variáveis nesse contexto.

2.4.3.1 Algoritmo de enxame de partículas- PSO

O algoritmo de otimização de enxame de partículas (PSO) é um meta-heurístico que utiliza uma população (enxame) de indivíduos (partículas) para otimizar funções no espaço multidimensional. A atualização das posições (soluções) de cada partícula é realizada considerando a melhor posição anterior da partícula e a melhor posição global (posição de todos os outros membros da população) (NEMA; THAKUR, 2015). Esse algoritmo foi proposto para seleção de variáveis na classificação (PSO-LDA) no trabalho de Lin e Chen (2009) (LIN, S.; CHEN, S., 2009).

As principais etapas do PSO para seleção de variáveis são descritas a seguir.

- *Inicialização*. O algoritmo PSO-LDA parte de uma população onde cada partícula possui uma velocidade determinada aleatoriamente e encontra-se em uma posição também aleatória (LIN, S.; CHEN, S., 2009).
- *Avaliação das aptidões*. A aptidão de cada partícula é avaliada, ou seja, a adequação ao LDA. O vetor posição é salvo para a melhor partícula comparando as aptidões de cada uma com a melhor aptidão até então obtida. Se a partícula apresenta aptidão superior a aptidão global, esta última será atualizada (LIN, S.; CHEN, S., 2009).
- *Atualização.* A velocidade e posição são atualizadas até atingir um critério de parada (LIN, S.; CHEN, S., 2009).

Desde sua formulação para seleção de variáveis, o PSO tem sido utilizado e adaptado para os problemas de classificação multivariada (CHENG, R.; JIN, 2015; GU; CHENG, R.; JIN, 2016; NEMA; THAKUR, 2015; XUE; ZHANG, M.; BROWNE, 2014). No trabalho de Xue *et al.* (2014), foram adicionadas três novas estratégias de inicialização da população

(baseando-se nos mecanismos *forward* e *backward*) no algoritmo e três novos mecanismos de atualização da posição da melhor partícula e da melhor solução global (XUE; ZHANG, M.; BROWNE, 2014). Para atualização das posições das partículas, os autores levaram em consideração a minimização do número de variáveis selecionadas. Com isso, os resultados do PSO modificado mostraram um desempenho superior na classificação e na minimização do número de variáveis quando comparado ao PSO tradicional (XUE; ZHANG, M.; BROWNE, 2014).

Nema e Thakur (2015) modificaram a equação de atualização da velocidade das partículas no PSO (NEMA; THAKUR, 2015). Para isso, os autores variaram o fator de aprendizado cognitivo (*C1*) e o fator de aprendizado social (*C2*), consideradas constantes no PSO tradicional (LIN, S.; CHEN, S., 2009). Consequentemente, o *C1* diminui e *C2* aumenta exponencialmente com o tempo. A redução no valor de *C1* indica a diminuição da busca local de cada indivíduo. O *C2* está relacionado à direção de busca do ponto ótimo global, assim o aumento pode levar a maior probabilidade de convergência no ponto global. Como resultado, o novo PSO convergiu mais rapidamente e apresentou taxa média de classificação melhor que os métodos LDA e PSO-LDA tradicionais (NEMA; THAKUR, 2015).

No trabalho de Gu, Cheng e Jin (2016), foi proposto o uso de uma variante do PSO, o CSO (*Competitive Swarm Optimizer*) para o problema de seleção de variáveis em alta escala (GU; CHENG, R.; JIN, 2016). Neste algoritmo, ao invés da melhor posição atual e global, as partículas aprendem com concorrentes selecionados aleatoriamente. Para isso, o enxame era dividido aleatoriamente em dois grupos, sendo que, em cada grupo, as partículas competiam aos pares. A partícula vencedora passava para a próxima iteração, enquanto a que perdeu atualizava sua posição e velocidade conforme a partícula vencedora. Neste trabalho, os autores utilizaram um mecanismo para registro dos valores dos subconjuntos de variáveis selecionados. Caso o subconjunto já tivesse sido avaliado anteriormente, sua aptidão era atribuída à nova partícula sem a necessidade de reavaliação. Se o subconjunto da nova partícula não foi avaliado nas iterações anteriores, então o modelo de classificação era construído e avaliado. Os autores apresentam essa estratégia como uma vantagem no tempo de busca das partículas pelas soluções, uma vez que muitas partículas apresentaram posições semelhantes quando o enxame estava convergindo. Os resultados do trabalho demonstraram que o algoritmo proposto superou o método convencional baseado em PCA, no PSO-KMN e variantes do PSO (GU; CHENG, R.; JIN, 2016).

2.4.3.2 Algoritmo de Otimização por Colônias de Formigas - ACO

O Algoritmo de Otimização por Colônias de Formigas (ACO) é um algoritmo meta-heurístico inspirado no comportamento coletivo das formigas ao buscarem alimentos (SHAMSIPUR *et al.*, 2007). Ao saírem do ninho em busca de fontes de alimentos, as formigas depositam feromônios pelo trajeto. A formiga que faz o trajeto mais curto até o alimento retorna para o ninho mais rápido, indicando que a rota é a melhor pela maior concentração de feromônios deixados. Por outro lado, as formigas que fazem trajetos mais longos contribuem para o aumento da porcentagem de evaporação do feromônio naquela rota. Então, as formigas dão preferência a rotas com maior concentração de feromônios otimizando o caminho da busca pelos alimentos (SHAMSIPUR *et al.*, 2007).

A partir do entendimento desse comportamento, o ACO foi desenvolvido por Dorigo e Stützle (2004) (DORIGO; STÜTZLE, 2004) e aplicado para o problema do caixeiro viajante. Nos problemas de otimização envolvendo modelagem multivariada, o ACO foi explorado para selecionar variáveis para modelos de calibração (SHAMSIPUR *et al.*, 2007). Recentemente, o ACO foi reformulado para seleção de variáveis no contexto da classificação via LDA (PONTES, A. S. *et al.*, 2020).

Para a seleção de variáveis, assume-se que cada formiga representa uma solução para o problema, onde é considerado um vetor feromônio contendo um número de elementos igual ao número de variáveis do domínio original. O ACO gera aleatoriamente uma colônia inicial de formigas e, para codificar a posição de cada elemento do vetor feromônio, adota o valor “1” para uma variável selecionada e “0” para variável não selecionada. Dessa forma, a aptidão de cada formiga é avaliada de acordo com a função objetivo pertinente ao problema (SHAMSIPUR *et al.*, 2007). As variáveis com maior quantidade de feromônios apresentam uma maior probabilidade de serem selecionadas, evidenciando um comportamento cooperativo apresentado pelas formigas da colônia em uma dada iteração do ACO. No final das iterações, o algoritmo seleciona a formiga que leva ao subconjunto das melhores variáveis para o modelo.

No contexto da classificação, Pontes *et al.*, (2020) implementaram no algoritmo ACO-LDA uma geração aleatória de formigas cegas que impede um mínimo local (PONTES, A. S. *et al.*, 2020). Neste trabalho o ACO-LDA desenvolvido foi avaliado em dois estudos de caso: classificação de óleos vegetais comestíveis por meio de espectrometria ultravioleta-visível (UV-Vis) e classificação simultânea de amostras de chás em relação ao tipo e origem geográfica via espectrometria de infravermelho próximo (NIR). Nesse estudo o ACO-LDA apresentou

desempenho de classificação superior ao algoritmo genético (GA-LDA) e semelhante ao PLS-DA.

Um novo algoritmo de colônia de formigas foi utilizado no trabalho de Zhang *et al.*, (2018). Neste, o mecanismo de seleção de variáveis pelas formigas foi combinado com a relação de correlação entre variáveis visando minimizar a redundância. O algoritmo foi aplicado para seleção de variáveis na classificação de manchas típicas de casca de frutas cítricas que estão associados a sintomas de doenças com o objetivo de distingui-las de frutas saudáveis. O desempenho do ACO modificado foi comprado ao do algoritmo das projeções sucessivas (SPA). O ACO selecionou comprimentos de ondas em regiões características correspondente as substâncias associados as diferentes manchas cítricas e a classificação por SVM apresentou excelente desempenho (ZHANG, Y. *et al.*, 2018).

No trabalho de Sheykhizadeh e Naderi (2018) o ACO foi utilizado para seleção de variáveis envolvendo problemas de calibração e classificação multivariada. Esse algoritmo foi utilizado para comparação com um novo algoritmo proposto pelos autores, o algoritmo de otimização invasiva de ervas daninhas (IWO) discutido na **Seção 2.4.3.3.** Neste trabalho, o ACO apresentou desempenho semelhante aos demais algoritmos utilizados (IWO, GA e PSO) (SHEYKHIZADEH; NASERI, 2018).

Nos trabalhos discutidos o ACO realizou a seleção de variáveis associadas as informações químicas relevantes, características das substâncias em estudo, assim, apresentou-se como um algoritmo vantajoso para tais aplicações.

2.4.3.3 Otimização Invasiva de Ervas Daninhas - IWO

O algoritmo de Otimização Invasiva de Ervas Daninhas (IWO) trata-se de um recente desenvolvimento para seleção de variáveis (SHEYKHIZADEH; NASERI, 2018). O IWO, originalmente desenvolvido para problemas de otimização, foi implementado por esses autores no contexto da seleção de variáveis, tanto para calibração (IWO-PLS) quanto para classificação multivariada (IWO-LDA). A bioinspiração deste algoritmo foi fundamentada no mecanismo invasivo de colônias de ervas daninhas ao buscarem local apropriado para crescimento e reprodução.

O princípio básico deste algoritmo compreende as etapas a seguir (SHEYKHIZADEH; NASERI, 2018):

- *Inicialização*. Inicialização de uma população de ervas daninhas (possíveis soluções) gerada aleatoriamente.
- *Avaliação*. Avaliação da aptidão de cada erva daninha da população.
- *Reprodução.* As ervas daninhas produzem sementes de acordo com a aptidão de cada uma e da colônia. Assim, as ervas daninhas com menor aptidão produzem menos sementes (subconjuntos de variáveis) e as ervas daninhas mais aptas produzem mais sementes, gerando uma segunda população (pop 2).
- *Dispersão*. Esta etapa possibilita que as sementes geradas permaneçam nas proximidades da planta daninha dos pais a partir de uma distribuição aleatória normal e o desvio padrão será reduzido de um valor inicial para um valor final a cada iteração.
- *Mutação.* Na sequência, realiza-se a etapa da mutação para evitar a convergência prematura ou ficar preso em ótimos locais. Nessa etapa, são utilizados operadores de troca, inserção e reversão para todas as sementes da população 2, gerando assim uma população 3 (pop 3).
- *Função de afinidade*. Uma função de afinidade foi inserida com o intuito de possibilitar a geração de soluções altamente diversificadas. Além disso, verifica-se a porcentagem de soluções bem classificadas restantes em cada iteração.
- *Competitividade*. Nessa etapa as populações pop1, pop2 e pop3 são classificadas com base em seus valores de condicionamento físico. As ervas daninhas com pior condicionamento físico são removidas, resultando em "lacunas" no tamanho da colônia para que as melhores se repliquem. Nessa etapa, escolhe-se um determinado número de ervas daninhas. O algoritmo se repete até atingir um número máximo de iterações. A cada iteração é armazenado o maior valor de adequação da erva daninha considerando os valores das iterações anteriores. No final, a erva daninha da última iteração que apresentar os melhores valores de aptidão é selecionada como melhor solução (SHEYKHIZADEH; NASERI, 2018).

O algoritmo IWO apresenta algumas etapas semelhantes ao algoritmo genético (GA), como a geração da população inicial e a reprodução considerando a maior probabilidade para indivíduos mais aptos. No GA, que será descrito na **Seção 2.4.4**, quando se emprega o método da roleta os indivíduos mais aptos gerados pela população inicial apresentam maior fatia na roleta, ou seja, maior probabilidade de serem escolhidos para reprodução. Com isso, todos os indivíduos (até os menos aptos, porém, com menor probabilidade) podem ser selecionados. Da mesma forma, no IWO todas as ervas daninhas reproduzem sementes, porém as menos aptas

reproduzem a menor quantidade possível. A etapa de mutação também é realizada no algoritmo genético, todavia, diferentemente do IWO, apenas alguns cromossomos sofrem mutações no GA. No IWO, as etapas posteriores a mutação são fundamentais para o desempenho do algoritmo. O armazenamento do valor de adequação da melhor erva daninha, a cada iteração, pode ser considerado como uma vantagem nesta implementação que possibilitará o algoritmo convergir para melhores subconjuntos de variáveis.

As principais vantagens do IWO destacadas pelos autores são a facilidade de implementação, já que são utilizados recursos fáceis de programar, a estrutura simples do algoritmo e a boa robustez dos resultados (SHEYKHIZADEH; NASERI, 2018). No trabalho citado, o desempenho do IWO para classificação (IWO-LDA) foi comparado ao algoritmo genético (GA-LDA), ao algoritmo de enxame de partículas (PSO-LDA) e ao algoritmo colônia de formigas (ACO-LDA). A seleção de variáveis foi realizada para os seguintes problemas: Discriminação simultânea de fórmula infantil adulterada de não adulterada e classificação de melamina e ácido cianúrico usados, individual ou simultaneamente, em fórmula infantil, com o método de espectrometria de infravermelho; Classificação de 44 amostras de vinhos pela origem geográfica. Como resultado, o IWO demonstrou desempenho semelhante aos outros algoritmos usados, tornando-o promissor e eficaz para problemas de seleção de variáveis (SHEYKHIZADEH; NASERI, 2018).

A **Seção 2.4.4** descreve em detalhes o algoritmo genético usado para seleção de variáveis. O GA-LDA foi aplicado nesta Tese como método estocástico comparativo ao novo algoritmo proposto.

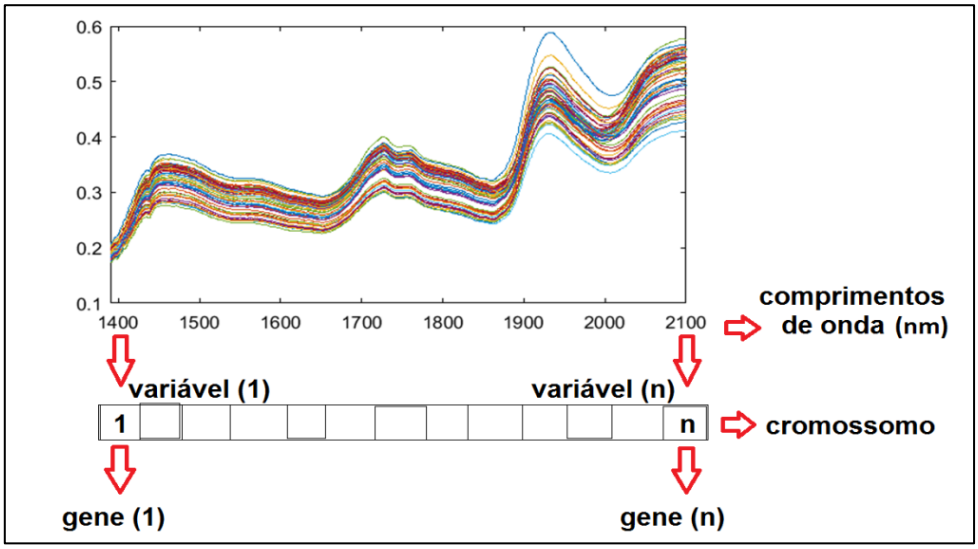
2.4.4 Descrição do algoritmo estocástico usado: GA-LDA

O algoritmo genético é uma técnica de busca estocástica desenvolvida por John H. Holland nos anos 60 para solucionar problemas de otimização complexos (FILHO, A. C.; POPPI, R.J., 1999). Esse algoritmo é fundamentado na teoria da evolução das espécies proposta por Darwin, segundo a qual indivíduos que se adaptam ao meio possuem maior probabilidade de sobreviver e gerar descendentes elevando sua população em relação aos que não se adaptaram (FILHO, A. C.; POPPI, R.J., 1999). Baseando-se neste princípio de geração de descendentes mais aptos (evoluídos), o algoritmo genético básico interpreta as informações dos sistemas como cromossomos genéticos e possui as seguintes etapas no contexto da seleção de variáveis:

Codificação das variáveis; Geração da população inicial; Avaliação das aptidões; Seleção; Cruzamento e *Mutação.*

- *Codificação das variáveis.* A codificação é realizada considerando cromossomos artificiais que contém as informações (genes) de cada parâmetro dos dados. No caso de problemas de seleção de variáveis em química analítica cada gene é indicado por uma variável (correspondente, usualmente, a um comprimento ou número de onda do espectro) e o conjunto das variáveis compõe o cromossomo, como ilustrado na **Figura 7**.

Figura 7-Demonstração da interpretação de variáveis espectrais como genes e composição dos cromossomos virtuais.

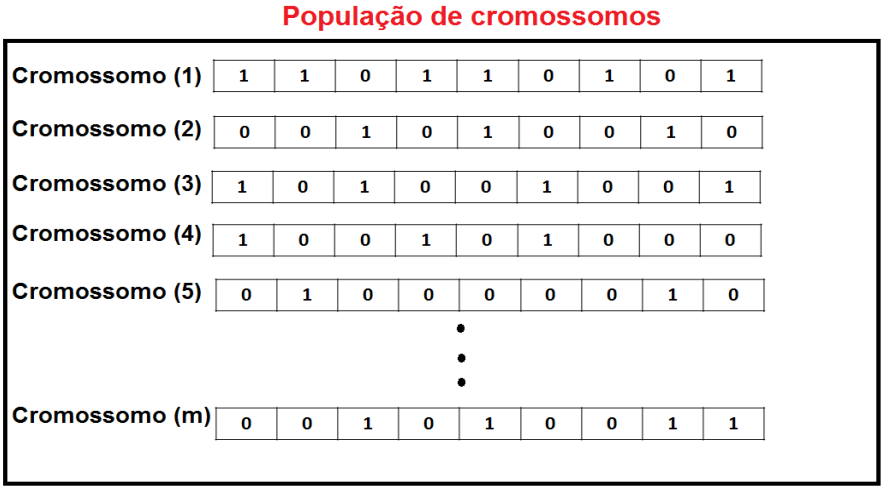


Fonte: (própria).

Para que o GA opere no domínio computacional, cada variável é convertida para o sistema binário. Assim, variáveis codificadas como "1" serão incluídas no modelo (variáveis selecionadas) enquanto as codificadas como "0", não. O percentual de variáveis selecionadas em cada cromossomo pode ser representado por um parâmetro "β" e o cromossomo é formado por todas as variáveis do espectro (KONZEN *et al.*, 2003).

- *Geração da população inicial*. Para que o algoritmo seja executado inicialmente é gerada, randomicamente, uma população de cromossomos (ou indivíduos) que representam possíveis soluções para o problema (**Figura 8**). Essa geração é completamente aleatória e cada indivíduo gerado é avaliado na próxima fase de execução do algoritmo.

Figura 8- Representação de uma população com *m* cromossomos virtuais e nove genes em cada cromossomo.

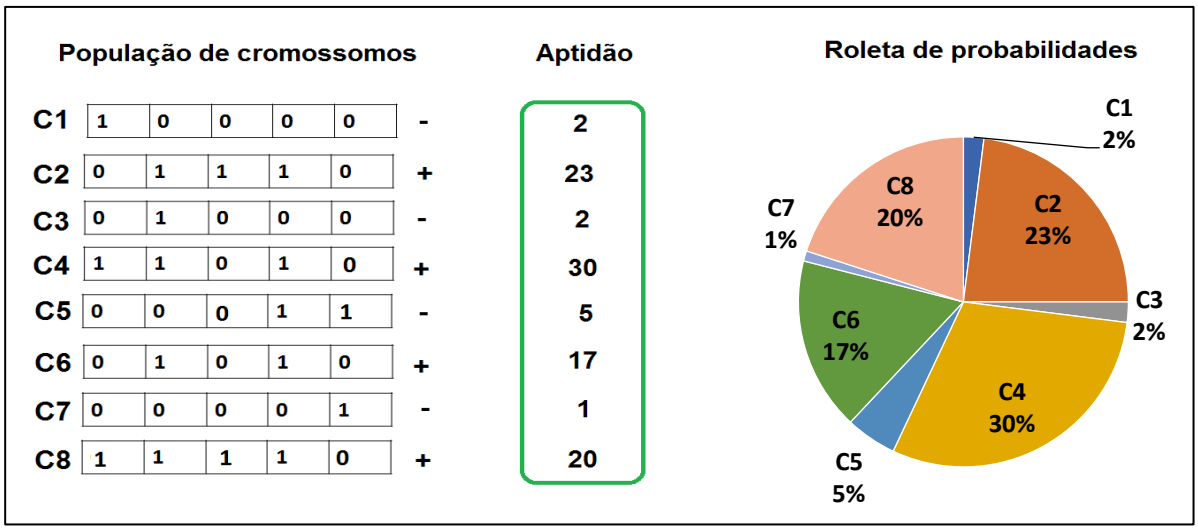


Fonte: (própria).

- *Avaliação da aptidão:* Cada cromossomo gerado aleatoriamente na população inicial contém um subconjunto de variáveis selecionadas (codificadas como “1”). A avaliação de cada subconjunto é realizada pela função objetivo da modelagem que está sendo utilizada. Os indivíduos que sobreviverão – e estão aptos para reproduzirem – serão aqueles que apresentarem a maior aptidão.
- *Cruzamento:* Nesta etapa os indivíduos começam a "acasalar" e "produzir descendentes”. Quando se faz o uso de elitismo (no caso do GA avançado) os indivíduos melhor avaliados na etapa anterior (geração da população inicial-n) são copiados diretamente para a próxima iteração (n+1). Os demais indivíduos da população (n) participarão do cruzamento e posterior avaliação, sendo que os melhores substituem a população inicial para próxima iteração onde todo o processo se repete até atingir um critério de parada (LEARDI, 2009). O problema em realizar dessa forma, a seleção dos cromossomos pais, consiste na geração de indivíduos cada vez mais semelhantes, o que impossibilitaria a saída da região se o algoritmo atingisse um ótimo local.

Outra forma mais interessante de selecionar os indivíduos para o cruzamento é por intermédio do método da roleta (LINDEN, 2008), onde cada indivíduo apresentará porcentagem (fatia) na roleta proporcional à sua aptidão (melhores respostas). Esse procedimento confere aos menos aptos uma chance de participarem do cruzamento, como exemplificado na **Figura 9**.

Figura 9- Método da roleta para seleção de cromossomos pais para o cruzamento e geração de descendentes.

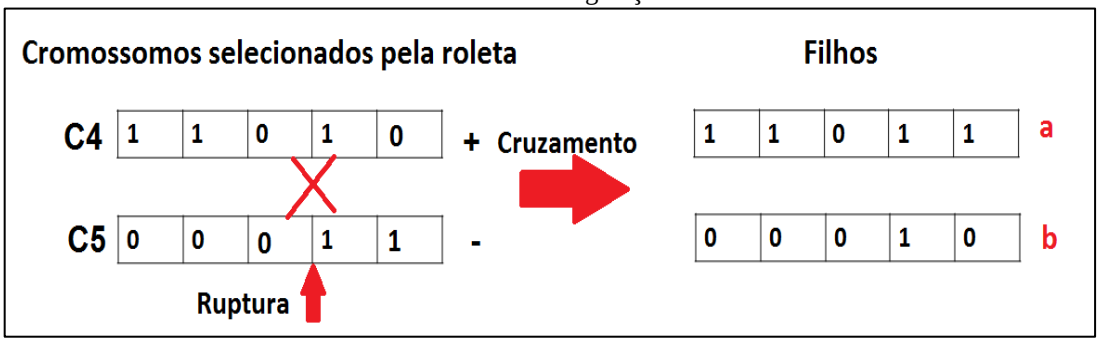


Fonte: (própria).

Como se pode notar na **Figura 9**, os indivíduos mais aptos possuem uma maior probabilidade de serem selecionados já que ganham um espaço maior na roleta que será girada aleatoriamente. Ao contrário, os menos aptos possuem menor fatia e, consequentemente, menor a probabilidade de serem selecionados. Analogamente, segundo a teoria de Darwin, os indivíduos que mais se adaptam ao ambiente (maior aptidão) possuem maior probabilidade de se reproduzirem e gerarem descendentes. Entretanto, os menos aptos também podem se reproduzir. A porcentagem de cruzamento, adotada usualmente, é de 60% (GALVÃO, R. K. H.; ARAÚJO, M. C. U., 2009).

A principal vantagem em utilizar este método é o fato de um indivíduo menos apto possuir um gene dominante (uma variável selecionada em determinada posição) que não aparece nos indivíduos mais aptos, o que pode levar no cruzamento a indivíduos com melhores características, como ilustrado na **Figura 10**.

Figura 10- Cruzamento entre cromossomos com maior (+) e menor (-) aptidão selecionados pelo método da roleta e geração de descentes.

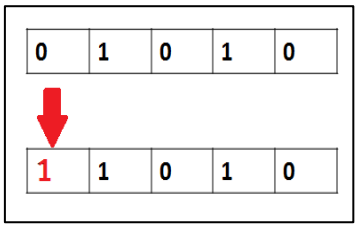


Fonte: (própria).

Neste exemplo, **Figura 10**, o filho (a) gerado pode apresentar uma melhor aptidão (devido à seleção da última variável) que não seria obtida pelo cruzamento entre os melhores cromossomos da população presente na **Figura 9**. Assim, no algoritmo genético torna-se fundamental a utilização de métodos como o da roleta na seleção dos cromossomos para o cruzamento.

- *Mutação*. A mutação pode ocorrer durante o processo de reprodução em uma parcela da população, sendo caracterizada por uma alteração aleatória de genes. Quando essas modificações nos genes promovem melhor adaptação do indivíduo ao ambiente, elas são transmitidas para os descendentes. Para esse operador genético, adota-se usualmente uma percentagem de mutação de 1 a 2% (LEARDI, 2009). Uma representação de mutação no código genético de cromossomos virtuais pode ser visualizada na **Figura 11**.

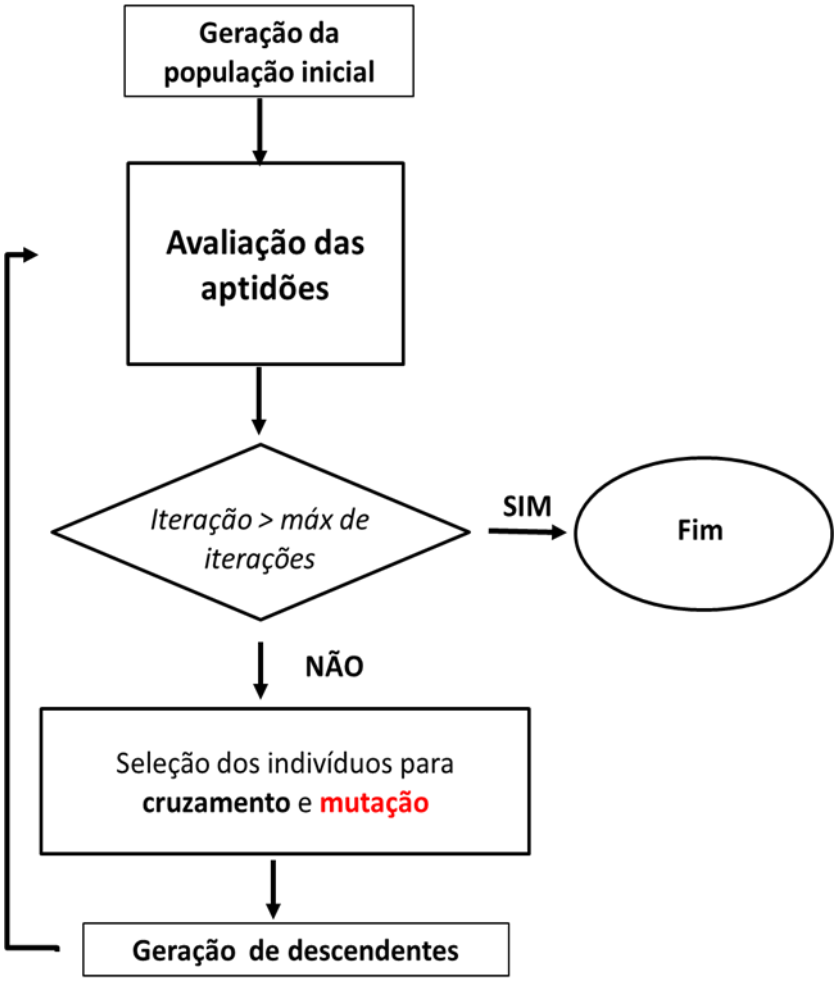
Figura 11- Representação de mutação no primeiro gene do cromossomo.



Fonte: (própria).

Um fluxograma envolvendo os operadores do algoritmo genético básico é esquematizado na **Figura 12**. O critério de parada do GA pode ser designado pelo número máximo de iterações, pela melhor aptidão ou quando o algoritmo convergir.

Figura 12- Fluxograma básico do GA.



Fonte: (própria).

No presente trabalho um algoritmo estocástico, inspirado na ecolocalização dos morcegos, é proposto para seleção de variáveis em modelagem LDA para fins de classificação, cuja descrição será apresentada no **Capítulo 3**.

ALGORITMO DOS MORCEGOS

Capítulo 3

CAPÍTULO 3: ALGORITMO DOS MORCEGOS

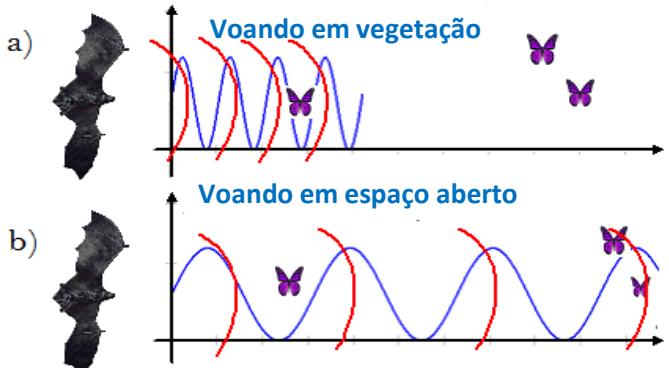
3. Algoritmo dos morcegos

O algoritmo dos morcegos (*Bat Algorithm*- BA) foi originalmente proposto por Yang (2010a) para solucionar problemas complexos de otimização (YANG, Xin-She, 2010a). Para formulação deste algoritmo Yang combinou vantagens de algoritmos como o PSO e o algoritmo de busca harmônica (HS). O BA foi concebido para simular o mecanismo de ecolocalização dos morcegos ao procurarem por presas. Esse mecanismo natural dos morcegos será abordado na **Seção 3.1**.

3.1 A ecolocalização dos morcegos naturais

Os morcegos são criaturas fascinantes que fazem uso de um sistema de emissão de sons de ondas ultrassônicas e captação do eco de retorno para se locomoverem. Esse mecanismo natural de movimentação dos morcegos a partir do eco é denominado ecolocalização e os permitem encontrar com precisão mesmo no escuro, presas, fendas de repouso e desviarem dos objetos circundantes (YANG, Xin-She, 2010a). A ecolocalização é utilizada por mais de 90% das espécies de morcegos e pode ser modulada dinamicamente (JAKOBSEN; BRINKLØV; SURLYKKE, 2013). Em geral, eles emitem sons, com frequência fixa variando o comprimento de onda, de alta intensidade e baixa taxa de emissão de pulsos ao procurarem por presas. Ao se aproximarem, diminuem a intensidade do som (para não espantar a presa) e elevam a taxa de emissão de pulsos aumentando assim a precisão da ecolocalização (YANG, Xin-She, 2010a). Entretanto, segundo Jakobsen *et al.*, (2013), os morcegos também alteram suas ecolocalizações em conformidade com o ambiente no qual caçam. De fato, os morcegos que voam dentro de uma vegetação emitem pulsos sonoros menos intensos, mais curtos e com maior frequência que os morcegos que voam em espaço aberto (JAKOBSEN; BRINKLØV; SURLYKKE, 2013). Dessa forma, em espaço aberto a ecolocalização tem um alcance maior apesar de ser menos precisa, conforme ilustrado na **Figura 13**.

Figura 13- A) Morcego emitindo pulsos sonoros com maior frequência (aumenta a precisão) e menor comprimento de onda (diminui o alcance da ecolocalização). B) Morcego emitindo pulsos sonoros com menor frequência (diminui a precisão) e maior comprimento de onda (aumenta o alcance da ecolocalização).



Fonte: (própria).

A frequência dos pulsos emitidos pelos morcegos encontra-se relacionada com os comprimentos das ondas ultrassônicas emitidas. A expressão matemática que representa esta relação é apresentada na **Equação 31**, onde *v* é a velocidade do som no ar é tipicamente de 340 m/s), λ o comprimento de onda do som de ondas ultrassônicas emitidas e f a frequência.

$$\lambda = v/f \tag{31}$$

Em geral, na natureza essa frequência dos pulsos emitidos é constante e encontra-se na faixa entre 25 e 150 kHz. Os pulsos duram apenas alguns milésimos de segundo e a taxa de emissão destes varia de 10 a 20 s^{-1}, podendo ser acelerada até 200 pulsos por segundo quando o morcego se aproxima da presa (YANG, Xin-She, 2010a).

Assim, a ecolocalização dos morcegos é governada, principalmente, pelos seguintes fatores:

I. Frequência dos pulsos emitidos que delimita o espaço de busca;
II. Amplitude ou volume do som que varia de um máximo quando o morcego está distante da presa para um mínimo quando se aproxima da presa;
III. Taxa de emissão de pulsos que variam de um mínimo quando o morcego está longe da presa para um máximo quando se aproxima.

A partir desse conhecimento, o algoritmo dos morcegos foi desenvolvido (YANG, Xin-She, 2010a).

3.2 Descrição do algoritmo básico dos morcegos virtuais

Inspirado nos princípios da ecolocalização, o algoritmo BA utiliza a frequência de pulsos sonoros emitidos durante a movimentação dos morcegos na busca por melhores soluções. Nesse algoritmo, as soluções buscadas são as posições ($\boldsymbol{x}_i$) dos morcegos virtuais. A população inicial é gerada e cada morcego tem sua própria taxa de emissão de pulsos (r_i), amplitude ou volume do pulso (A_i), frequência (f_i) e velocidade ($\boldsymbol{v}_i$), determinados aleatoriamente. A aptidão de cada morcego da população inicial é armazenada, sendo atribuída a posição $\boldsymbol{x}_*$ para o melhor morcego. Na sequência as posições ($\boldsymbol{x}_i$) dos morcegos são atualizadas pelas **Equações 32** a **34**, pelas mudanças na velocidade ($\boldsymbol{v}_i$) e frequência dos pulsos sonoros emitidos (fi).

$$fi = fmin + (fmax - fmin)\beta \tag{32}$$

$$\boldsymbol{v}_i^t = \boldsymbol{v}_i^{t-1} + (\boldsymbol{x}^t - \boldsymbol{x}_*)f_i \tag{33}$$

$$\boldsymbol{x}_i^t = \boldsymbol{x}_i^{t-1} + \boldsymbol{v}_i^t \tag{34}$$

onde f_{min} e f_{max} correspondem ao limite inferior e superior do espaço de busca, sendo $\beta \in [0,1]$ um escalar aleatório de uma distribuição normal gerado a cada iteração para cada morcego e t denota o número de iterações ($t \rightarrow \infty$).

Após atualização das posições dos morcegos, é verificada a taxa de emissão de pulsos sonoros (r_i) e comparada com um escalar gerado por uma função randômica. Quando ruído aleatório for maior que a taxa de emissão de pulsos ($r_i^t < rand$), isso indica que o morcego está distante da presa (da melhor solução, $\boldsymbol{x}_*$). Nesses casos, realiza-se uma busca local a partir de um pequeno deslocamento aleatório dado pelo valor de ε conforme representado na **Equação 35**, gerando uma nova posição para o morcego.

$$\boldsymbol{x}_{novo} = \boldsymbol{x}_{atual} + \varepsilon A^t \tag{35}$$

onde $\varepsilon \in [-1,1]$ é um número aleatório e A^t é a intensidade média (amplitude ou volume do pulso emitido) de todos os morcegos nesta iteração. Quanto menor a amplitude da população de morcegos, mais próximos da presa eles se encontram.

Na sequência avalia-se se a nova solução é melhor que a solução da iteração anterior e também compara-se a amplitude do pulso A_i com uma função randômica, caso $rand < A_i^t$ provavelmente o morcego está próximo da melhor solução. Sendo verdadeiras essas condições,

os morcegos atualizam suas posições. A taxa de emissão de pulsos e a amplitude também são atualizadas, conforme **Equações 36 e 37**.

$$r_i^{t+1} = r_i^0[1 - \exp(-\gamma t)] \tag{36}$$

$$A_i^{t+1} = \alpha A_i^t \tag{37}$$

onde γ e α são números reais adotados de acordo com os intervalos 0< α<1 e γ>0.

Como pode ser inferido a partir das **Equações 36** e **37**, a amplitude tende a diminuir conforme *t* tende a infinito e a taxa de emissão de pulsos tende a se aproximar de 1. Ou seja, A_i é reduzido e r_i elevado. O algoritmo se repete até atingir um critério de parada.

3.3 Descrição do algoritmo adaptado dos morcegos virtuais

Nesse trabalho foi proposta uma versão do BA algoritmo desenvolvida para uma nova aplicação voltada para seleção de variáveis em modelagem de classificação multivariada usando Análise Discriminante Linear (LDA). O algoritmo proposto, denominado BA-LDA, foi formulado com uma função de custo (G_{cost}) (PONTES, C. *et al.*, 2005) que calcula o risco médio do erro de classificação. Para isso, o G_{cost} é calculado conforme a **Equação 38** empregando um conjunto de amostras de validação.

$$G_{cost} = \frac{1}{K-L-C}\sum_{k=1}^{k_V} g_k \tag{38}$$

onde *K* é o número de amostras de treinamento, *L* é o número de variáveis selecionadas, *C* o número de classes alvo e g_k é o risco de uma amostra de validação ser classificada incorretamente. Para uma amostra desconhecida x_k da *k-ésima* amostra de validação, determinar-se o g_k pela **Equação 39**.

$$g_k = \frac{r^2(\boldsymbol{x}_k, \mu_{lk})}{min_{lj \neq lk}\, r^2(\boldsymbol{x}_k, \mu_{lj})} \tag{39}$$

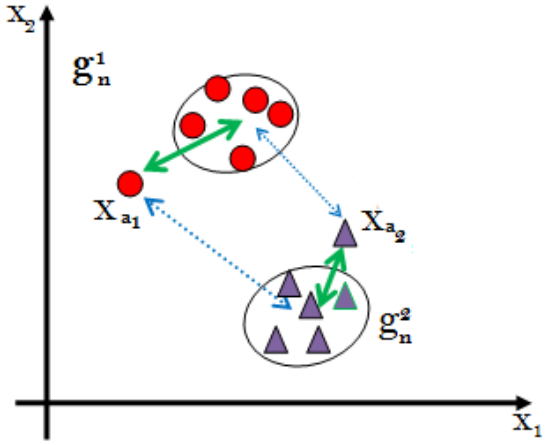
Na **Equação 39**, o numerador e o denominador correspondem, respectivamente, ao quadrado das distâncias de *Mahalanobis* entre o objeto $\boldsymbol{x}_k$ e a média de sua classe ($\boldsymbol{\mu}_{lk}$) e entre a média da classe errada mais próxima ($\boldsymbol{\mu}_{lj}$), conforme representado na **Figura 14**. Ambos os

quadrados das distâncias de *Mahalanobis*, $r^2(\boldsymbol{x}_k, \mu_{lk})$ e $r^2(\boldsymbol{x}_k, \mu_{lj})$, são encontrados usando as **Equações 40** e **41**.

$$r^2(\boldsymbol{x}_k, \mu_{lk}) = (\boldsymbol{x}_k - \mu_{lk})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \mu_{lk})^T \quad (40)$$

$$r^2(\boldsymbol{x}_k, \mu_{lj}) = (\boldsymbol{x}_k - \mu_{lj})\Sigma^{-1}(\boldsymbol{x}_k - \mu_{lj})^T \quad (41)$$

Figura 14- Ilustração do cálculo do risco do erro de classificação para uma amostra x_{a1}.



Fonte: (própria).

Como se pode observar na **Equação 39**, o risco de classificação incorreta (g_k) de uma amostra particular (x_{a1}) é baixo quando o valor da distância de Mahalanobis entre essa amostra e o centro da classe errada mais próxima (g_n^2) é alto (**Figura 14**). Isso significa que a referida amostra encontra-se próxima à sua classe verdadeira e distante da classe incorreta (**Figura 14**). Idealmente, uma amostra x_k do conjunto de validação deve estar o mais próximo do centro da sua classe verdadeira e o mais distante possível do centro das outras classes. Assim, o risco médio do erro de classificação (G_{cost}) dever ser obtido com o menor valor possível.

A função G_{cost} foi utilizada como guia para os morcegos virtuais de modo que suas posições sejam atualizadas apenas quando for obtido um menor valor para G_{cost}.

3.3.1 Fluxograma do algoritmo BA-LDA proposto

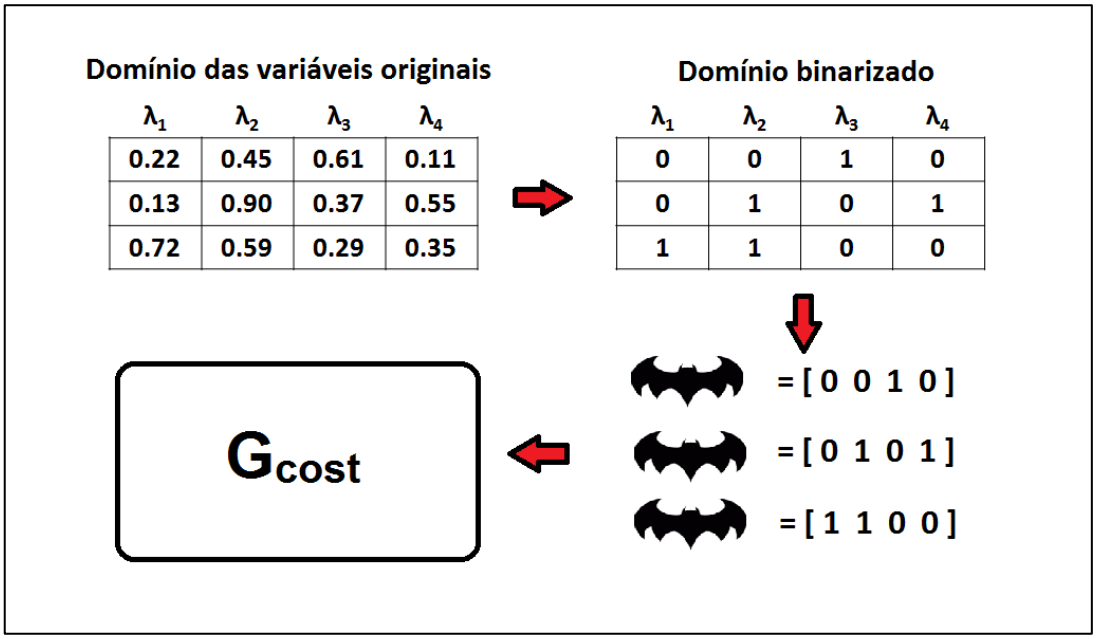
Para a inicialização do algoritmo proposto, os parâmetros de entrada de dados são: o número mínimo (*n_min*) e máximo (*n_max*) de variáveis selecionadas, as matrizes com as amostras de treinamento do modelo (*Trein*), validação (*Val*) e teste (*Teste*) e as matrizes com os índices de classe das amostras de treinamento (*Trein_group*), validação (*Val_group*) e teste (*Test_group*). Além disso, o número de morcegos virtuais (*mbats*) e o número de iterações (*N*) são definidos, bem como γ e α. Nesta Tese realizamos um planejamento fatorial para otimização destes parâmetros (*mbats, N,* γ e α) para cada conjunto de dados estudado (**Seção 4.7**).

Para um conjunto de dados robusto que é dividido por série de teste (amostras de treinamento, validação e teste), o número máximo (*n_max*) de variáveis a serem selecionadas não pode ser maior do que o número de amostras de treinamento menos o número de classes. Quando a validação interna é aplicada (amostras de treinamento e teste), *n_max* não deve ser maior que o número de amostras de treinamento menos o número de classes menos um grau de liberdade. Assim, o algoritmo seleciona um número de variáveis entre *n_min* e *n_max*.

O algoritmo proposto seleciona as variáveis dos dados multivariados, que são mais apropriadas para o modelo BA-LDA aplicado ao problema de classificação, por intermédio do melhor morcego virtual ($\boldsymbol{x}_*$) obtido no final do processo de busca. Para isso, são executados os seguintes passos: *inicialização*, *avaliação da função objetivo*, *atualização das posições*, *busca local*, *avaliação da função objetivo das novas posições* e *localização do melhor morcego*, descritos abaixo conforme Fluxograma apresentado na **Figura 16 (A)** e **(B)**.

- *Inicialização.* Essa etapa é representada na **Figura 16 (A)**, na qual é gerada uma matriz de posições que possui a mesma dimensão da matriz de dados originais. Essa matriz gerada aleatoriamente contém valores que variam de 0 a 1 e representa a população inicial de morcegos (**Figura 15)**. Para seleção de variáveis, que é um problema binário, a matriz de posições iniciais dos morcegos deve ser binarizada. Assim, conforme ilustração na **Figura 15**, valores maiores que 0,5 são codificados como 1 (representando as variáveis inicialmente incluídas no modelo) e valores iguais ou menores que 0,5 são codificados como 0 (indicando variáveis não incluídas). Assim, cada morcego possui um subconjunto de variáveis selecionadas na população inicial. Além disso, define-se aleatoriamente a taxa de emissão (r_i), a amplitude (A_i) e a frequência (f_i) dos pulsos, assim como a velocidade ($\boldsymbol{v}_i$) dos morcegos.

Figura 15- Binarização e seleção inicial de variáveis.



Fonte: (própria).

- *Avaliação da função objetivo.* Calcula-se o valor de G_{cost} correspondente a cada morcego da população inicial e armazenam-se as posições (variáveis selecionadas) dos morcegos virtuais. O menor valor obtido para G_{cost} determina a posição inicial do melhor morcego, $\boldsymbol{x}_*$. Esta etapa conclui o primeiro bloco do fluxograma de execução do programa BA-LDA (**Figura 16 (A)**).

- *Atualização das posições.* Por meio das **Equações (32)**, **(33)** e **(34)** novas posições ($\boldsymbol{x_i}$) são geradas para cada morcego a partir de uma busca aleatória. A atualização só ocorre se a taxa de emissão dos pulsos estiver aumentando $r_i^t \rightarrow r_i^0$ (indicando um aumento na precisão de procura por presas pelos morcegos naturais).

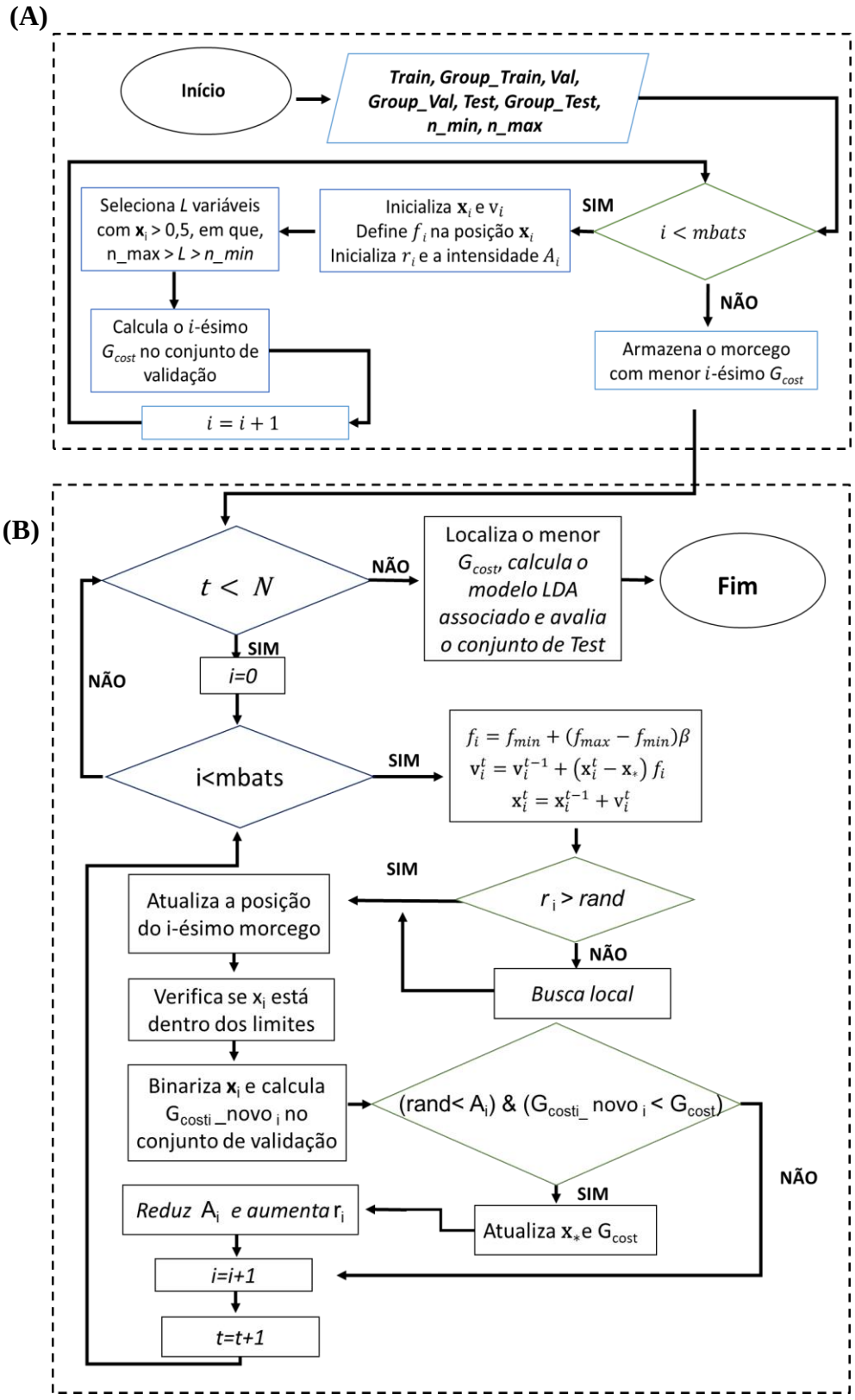
- 4) *Busca local.* Quando a taxa de emissão de pulsos for menor que o ruído aleatório ($r_i^t < rand$), isso indica que o morcego encontra-se distante de $\boldsymbol{x}_*$(melhor posição). Nesses casos, realiza-se uma busca local que permite atualizar a posição desse morcego, obtendo-se uma nova posição ($\boldsymbol{x_{novo}}$), a partir de um pequeno deslocamento aleatório em torno da posição atual ($\boldsymbol{x_{atual}}$) conforme representado na **Equação (35)**.
- *Avaliação da nova função objetivo das novas posições.* As novas posições dos morcegos somente deverão ser aceitas se:

- Se o valor (*rand*), obtido com uma função randômica, for menor que a amplitude do pulso (A_i). Neste caso, o *i*-ésimo morcego encontra-se perto da melhor posição;
- Se o custo, calculado para as novas posições dos morcegos (G_{costi_novo}), for menor que o G_{cost} da iteração anterior. Isso indica que as variáveis selecionadas pelos morcegos, nessas novas posições para cada iteração, levam a um menor erro de classificação.

Se essas condições são satisfeitas, então a taxa de emissão de pulsos (r_i) e a amplitude (A_i) são atualizados de acordo com as **Equações (36)** e (**37)**, indicando que os morcegos estão se aproximando da presa. Após esta etapa, o algoritmo executa a próxima iteração em que novas posições são propostas e avaliadas para cada morcego. Essas etapas (**Figura 16**) são repetidas para todos os morcegos (*mbats*) da população inicial até que todas as iterações (*N*) tenham sido executadas.

- *Localização do melhor morcego.* Ao final, localiza-se o melhor morcego ($\boldsymbol{x}_*$) que levou ao menor G_{cost} entre todas as iterações executadas e utilizam-se as variáveis selecionadas por ele na construção do modelo BA-LDA, conforme fluxograma da **Figura (16)**. Dessa forma, o algoritmo proposto possibilita a convergência para os melhores subconjuntos de variáveis que podem levar a melhor discriminação de classes de amostras.

Figura 16- Fluxograma do BA-LDA para seleção de variáveis.



Fonte: (própria).

O próximo capítulo apresenta a metodologia envolvida no trabalho.

METODOLOGIA

Capítulo 4

CAPÍTULO 4: METODOLOGIA

4. Metodologia

Nesta Tese o desempenho do algoritmo proposto (BA-LDA) foi avaliado em quatro estudos de caso, a saber, na classificação de dados (MS) referentes a amostras de soro de pacientes com e sem câncer de ovário, dados (NIR) reais referentes a amostras de cafés, dados (UV-Vis) referentes a amostras de óleos vegetais e dados de diesel (NIR) com informação simulada. As **Seções 4.1**, **4.2**, **4.3** e **4.4** descrevem esses dados.

4.1 Dados MS: Estudo de caso – Classificação de soro de pacientes com e sem câncer de ovário

Para este estudo de caso, os dados referentes as amostras de soro de pacientes com e sem câncer de ovário foram obtidos diretamente no *software* MatLab® 2017b. Para acessa-los bastou-se digitar o comando "*load ovariancancer*". Após isso, duas matrizes foram carregadas no *Worksapce*. A matriz "*obs*" contendo os sinais proteômicos de 216 amostras com 4000 variáveis, obtidos em um espectrômetro de massas de alta resolução. A matriz "*grp*" mostra os índices das classes, sendo 121 amostras pertencentes a classe de pacientes com câncer de ovário -PC e 95 amostras de controle (pacientes sem câncer de ovário -PS). Os dados foram divididos pelo KS em amostras de treinamento (130 amostras), validação (43 amostras) e teste (43 amostras) e aplicados aos diferentes métodos de classificação. A **Tabela 1** resume a divisão de todos os conjuntos de dados com o algoritmo KS.

As amostras de soro foram obtidas no Programa Nacional de Detecção Antecipada do Câncer de Ovário (NOCEDP) e na clínica de oncologia *gyne cologic da Northwestern University* (Chicago, IL, EUA). As amostras de controle foram adquiridas a partir de mulheres inscritas no NOCEDP que não tiveram nenhuma evidência de câncer por 5 anos. As amostras de pacientes com câncer foram obtidas de mulheres no estado pré-operatório e que apresentavam carcinoma epitelial de ovário. A discussão sobre o preparo das amostras até a obtenção dos sinais é apresentada no trabalho de Conrads *et al.,* (2004) (CONRADS *et al.*, 2004).

4.2 Dados NIR: Estudo de caso – Classificação de cafés

O segundo conjunto de dados analisado consistiu em espectros de reflectância no infravermelho próximo (NIR) de 60 amostras de cafés moídos pertencentes a duas classes (30 *gourmet* e 30 tradicionais). Os dados foram adquiridos na região entre 4907 e 6160 cm^{-1} (2037,9 e 1623 nm) com resolução de 1 cm^{-1}, utilizando um Espectrofotômetro FT-NIR, Modelo *Analect Diamond* 20, acoplado ao Acessorio de Reflectância Difusa, da *Applied Instrument Technologies*®. Inicialmente, os dados foram divididos em treinamento (40 amostras) e teste (20 amostras), utilizando o algoritmo KS. As amostras de treinamento também foram usadas para validação interna na modelagem LDA para classificação, enquanto as amostras de teste foram usadas apenas na avaliação final dos modelos. Os dados foram obtidos e disponibilizados por Araújo (ARAÚJO, T. K. L. *et al.*, 2021).

4.3 Dados UV-Vis: Estudo de caso – Classificação de óleos vegetais

O terceiro conjunto de dados analisado consistiu em espectros de ultravioleta visível (UV-Vis) de 119 amostras de óleos vegetais pertencentes a quatro classes distintas (milho: 29, soja: 30, canola: 29 e girassol: 31). Os espectros foram registrados de 220 a 400 nm, com resolução de 1 nm. Neste trabalho, removemos a região não informativa (339 - 400 nm) e usamos os dados entre 220 e 338 nm. Para avaliação com algoritmos BA-LDA, GA-LDA, SPA-LDA, PLS-DA e na classificação SIMCA, os dados foram divididos em amostras de treinamento (69 amostras), validação (25 amostras) e teste (25 amostras), utilizando o algoritmo KS. O conjunto de dados foi proveniente do trabalho de Pontes (PONTES, M. J. C., 2009). Para comparação com os resultados obtidos por Pontes (2009), a mesma divisão dos dados pelo KS e os espectros completos foram usados na segunda aplicação do BA-LDA (resultados descritos na **Seção 5.3.5** e apresentados na **Tabela 14**).

4.4 Dados NIR: Estudo de caso - Diesel e informação simulada

A simulação foi realizada a partir de um conjunto de dados de amostras reais, mantendo-se a variabilidade dos dados reais e acrescentando novas informações simuladas para designar uma classe diferente de amostras. Para isso, utilizou-se os dados NIR de amostras de diesel disponíveis em "http://www.eigenvector.com/data/SWRI/index.html". Esses dados são referentes apenas a uma única classe de amostras do combustível e foram disponibilizados para testes com algoritmos de seleção de variáveis para calibração multivariada.

Para a criação de uma nova classe, inicialmente, foi construída uma matriz apenas com as duzentas primeiras amostras e trezentos e cinquenta variáveis do conjunto de dados reais. As cem primeiras amostras foram designadas como pertencentes a classe 1 e as outras cem amostras foram modificadas com a inserção de informações que simulavam uma banda espectral entre as variáveis 84 e 108. A banda espectral foi simulada com uma intensidade pequena com o propósito de verificar a capacidade do algoritmo em distinguir essas amostras como pertencentes a uma nova classe. Para esses dados, um estudo de adição de ruído também foi realizado para verificar a sensibilidade dos métodos de classificação usados. O conjunto de dados foi divido em amostras de treinamento (120 amostras), validação (40 amostras) e teste (40 amostras), usando o algoritmo KS, a **Tabela 1** mostra a divisão dos conjuntos para todos os estudos de casos analisados.

Tabela 1: Divisão dos conjuntos de dados pelo KS.

		Conjuntos		
Amostras	Classes	Treinamento	Validação	Teste
Diesel com informação simulada (NIR)	Reais	60	20	20
	Simuladas	60	20	20
	Total	120	40	40
Cafés (NIR) CV	*Gourmet*	20	0	10
	Tradicionais	20	0	10
	Total	40	0	20
Óleos vegetais (UV-Vis)	Canola	17	6	6
	Milho	17	6	6
	Soja	18	6	6
	Girassol	17	7	7
	Total	69	25	25
Soro de pacientes com e sem câncer de ovário (espectro de massas)	PC	73	24	24
	PS	57	19	19
	Total	130	43	43

PC- pacientes com câncer. PS- pacientes sem câncer. CV- Validação Cruzada.

Fonte: (própria).

4.5 Softwares e procedimentos quimiométricos

Neste trabalho, o código BA-LDA foi desenvolvido e implementado no Matlab® 2011b. O SIMCA foi usado empregando o programa *The Unscrambler X* vs 10.4. Os algoritmos BA-LDA, GA-LDA, SPA-LDA e PLS-DA foram executados no Matlab® 2017b. O GA-LDA é um algoritmo estocástico tradicionalmente usado como método de comparação com novos algoritmos (ATTIA *et al.*, 2016; PAULA, DE *et al.*, 2017; SHEYKHIZADEH; NASERI, 2018), por esse motivo, foi utilizado neste estudo. Como mencionado, o PLS-DA também é um classificador tradicional e, portanto, também foi usado para comparar com o BA-LDA. O algoritmo KS, foi usado para separar as amostras por série de testes (treinamento, validação e amostras de teste) e por validação interna ou validação cruzada (treinamento e amostras de teste). Este algoritmo maximiza a distância euclidiana entre os vetores de resposta instrumental para as amostras selecionadas. O código KS também foi usado no Matlab® 2017b. Para realizar o planejamento dos experimentos e determinar os parâmetros ideais do BA-LDA, o software STATISTICA 12 foi empregado.

4.6 Parâmetros

Para determinação dos parâmetros ideais do BA-LDA, foi realizado um planejamento fatorial fracionado 2^{4-1} com uma repetição para cada um dos conjuntos de dados estudados. Os resultados foram tratados usando o programa STATISTICA 12. Os parâmetros e níveis estudados podem ser vistos na **Tabela 2**.

Tabela 2: Fatores estudados no planejamento fatorial fracionado2^{4-1}.

Fatores	+	-
α	0,9	0,5
γ	0,9	0,4
mbats	50	30
N	500	200

Fonte: (própria).

A **Tabela 3** mostra a matriz de delineamento, gerada com 2^{4-1} fatores. Essa tabela foi usada para cada um dos conjuntos de dados e os experimentos foram executados para avaliação da combinação de parâmetros que levariam ao melhor desempenho do algoritmo.

Tabela 3: Matriz de delineamento gerada com 2^{4-1} fatores.

Replicadas	**α**	**γ**	***mbats***	***N***
1	-	-	-	-
1	+	-	-	+
1	-	+	-	+
1	+	+	-	-
1	-	-	+	+
1	+	-	+	-
1	-	+	+	-
1	+	+	+	+
2	-	-	-	-
2	+	-	-	+
2	-	+	-	+
2	+	+	-	-
2	-	-	+	+
2	+	-	+	-
2	-	+	+	-
2	+	+	+	+

Fonte: (própria).

4.7 Robustez

Para o estudo da robustez o desempenho do BA-LDA foi comparado ao do GA-LDA, por ser um algoritmo estocástico. Para cada conjunto de dados, cada algoritmo foi executado cem vezes e foram construídos histogramas de frequência das variáveis selecionadas. O objetivo do estudo foi verificar se os algoritmos convergiam para seleção de variáveis em regiões específicas. Os próximos capítulos apresentam os resultados obtidos para cada conjunto de dados.

RESULTADOS E DISCUSSÃO

Capítulo 5

CAPÍTULO 5: RESULTADOS E DISCUSSÃO

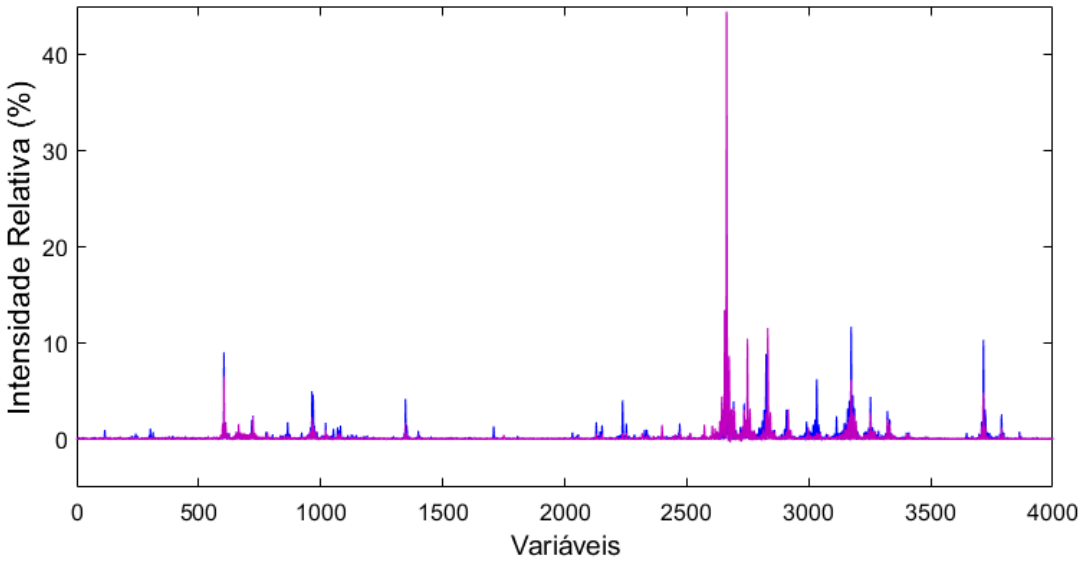
5. Resultados e Discussão

As **Seções 5.1** a **5.4** apresentam os resultados obtidos nesta tese para os quatro conjuntos de dados avaliados. A otimização dos parâmetros do algoritmo proposto (BA-LDA) e a comparação do desempenho deste com outros métodos de classificação foi discutida. Assim, a **Seção 5.1** aborda o primeiro conjunto de dados avaliado.

5.1 Estudo de caso envolvendo a classificação de soro de pacientes com e sem câncer de ovário baseada em dados espectrométricos de massas

A **Figura 17** mostra os perfis proteômicos das 216 amostras de soro de pacientes com e sem câncer de ovário. Os espectros em azul (**Figura 17**) correspondem as amostras dos pacientes com câncer e os espectros em lilás são referentes as amostras dos pacientes sem câncer de ovário. Devido à grande variabilidade biológica, a variabilidades nas análises e ao ruído instrumental, a discriminação entre essas classes de amostras é um problema complexo. Conforme Conrads *et al.*, (2004) a principal fonte de variação dos dados no estudo foi determinada como relacionada ao instrumento e quanto mais pré-processamentos (suavização, filtragem) necessários, maior a probabilidade de suprimir características reais com informações de diagnóstico (CONRADS *et al.*, 2004). Assim, estudamos os sinais sem pré-processamento. Conrads et al. (2004) acreditam que características altamente discriminatórias podem estar muito próximas do ruído de fundo e por isso não recomendam o pré-processamento.

Figura 17- Espectros de massas de soro de pacientes com câncer de ovário e pacientes sem a doença.



Fonte: (própria).

Para todos os conjuntos de dados aplicados nesta Tese, inicialmente foi realizado um planejamento fatorial fracionado 2^{4-1} para determinar os parâmetros ideias do BA-LDA. Assim, a **Seção 5.1.1** apresenta os resultados obtidos para tal planejamento.

5.1.1 Otimização dos parâmetros do BA-LDA

Para otimização dos parâmetros do BA-LDA foram avaliados a significância dos fatores (α, γ, *mbats* e *N*) e suas respectivas interações. O intuído deste experimento foi determinar os parâmetros significantes e que levam as melhores respostas de taxa de classificação correta (TCC).

A **Tabela 4** mostra as respostas obtidas em termos de %TCC para cada um dos experimentos executados no planejamento fracionado 2^{4-1} com uma repetição.

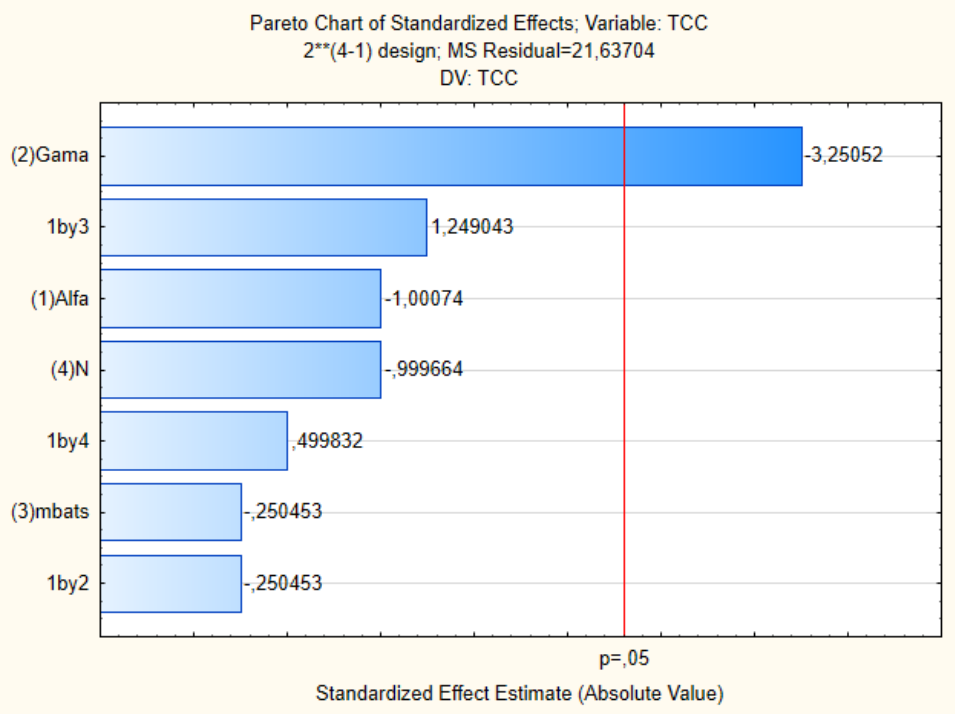
Tabela 4: Matriz de respostas no planejamento fatorial fracionado 2^{4-1}.

Replicadas	α	γ	*mbats*	*N*	*%TCC*
1	0,5	0,4	30	200	90,70
1	0,9	0,4	30	500	81,40
1	0,5	0,9	30	500	79,07
1	0,9	0,9	30	200	74,42
1	0,5	0,4	50	500	86,05
1	0,9	0,4	50	200	88,37
1	0,5	0,9	50	200	79,07
1	0,9	0,9	50	500	76,74
2	0,5	0,4	30	200	83,72
2	0,9	0,4	30	500	79,07
2	0,5	0,9	30	500	74,42
2	0,9	0,9	30	200	72,09
2	0,5	0,4	50	500	74,42
2	0,9	0,4	50	200	79,07
2	0,5	0,9	50	200	74,42
2	0,9	0,9	50	500	72,09

Fonte: (própria).

A partir destas respostas (TCC), os efeitos e suas interações foram avaliadas. O gráfico de Pareto dos efeitos é mostrado na **Figura 18**.

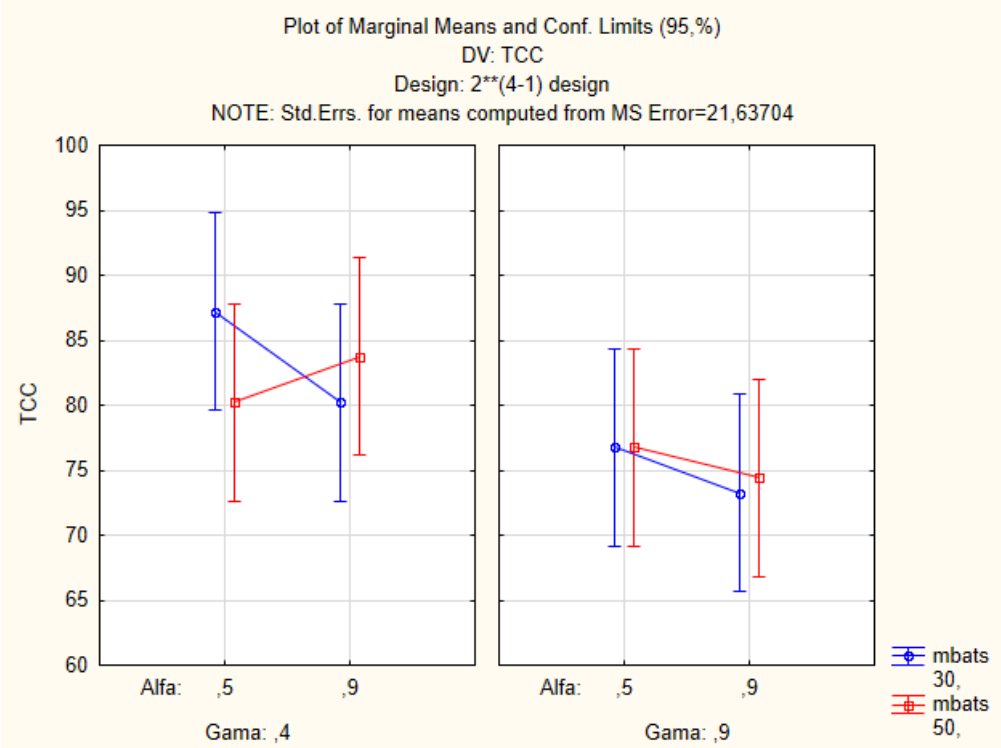
Figura 18- Gráfico de Pareto dos efeitos para os quatro fatores no planejamento fracionado 2^{4-1}.



Fonte: (própria).

Na **Figura 18**, é possível perceber que apenas o fator *Gama* (γ) foi significativo, e quando esse fator passou do nível inferior para o nível superior ocorreu um decréscimo da resposta (-3,25) indicando que a melhor resposta é obtida com γ no nível inferior (γ=0,4). Também foi possível observar que a interação entre os fatores *Alfa* (α) e *mbats* apresentou maior influência que esses fatores independentes. Assim, apesar de não demostrarem significância estatística, para definir os melhores níveis para estes fatores é importante considerar o gráfico das médias (**Figura 19**).

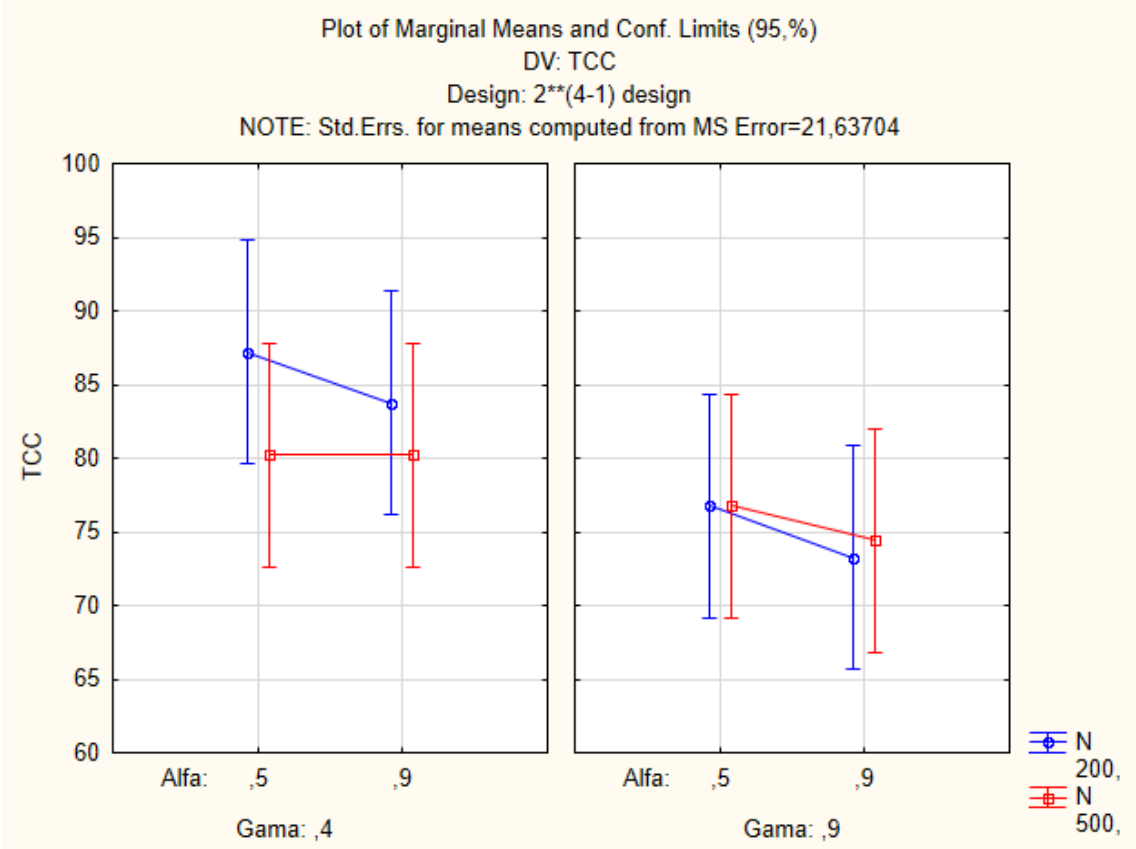
Figura 19- Gráficos de médias para interação dos efeitos entre os fatores *Gama*, *Alfa* e *mbats*.



Fonte: (própria).

Como se pode observar na **Figura 19**, o melhor desempenho foi obtido com alfa no nível inferior (α=0,5) e *mbats* também no nível inferior (*mbats* =30). Pelo gráfico de Pareto já podia-se observar que o melhor desempenho era obtido quando *mbats* e *α* estavam no nível inferior, devido ao decréscimo das respostas quando estes fatores passaram para o nível superior. O fato do melhor desempenho de γ no nível inferior também já havia sido observado no gráfico de Pareto. A partir desses resultados podemos considerar que o melhor desempenho foi obtido com γ = 0,4, α=0,5 e *mbats*= 30. A **Figura 20**, apresenta o gráfico das médias considerando o número de interações *N*.

Figura 20- Gráficos de médias para interação dos efeitos entre os fatores *Gama*, *Alfa* e *N*.



Fonte: (própria).

A partir da **Figura 20,** podemos inferir que o melhor desempenho é obtido com *N* no nível inferior, quando consideramos gama (γ) igual a 0,4. Assim, apesar dos efeitos α, *mbats* e *N* não serem significativos, os gráficos das médias foram usados para escolhermos os níveis dos fatores. A **Tabela 5** resume os parâmetros fixados pelo experimento.

Tabela 5: Parâmetros fixados pelo planejamento fatorial fracionado 2^{4-1}.

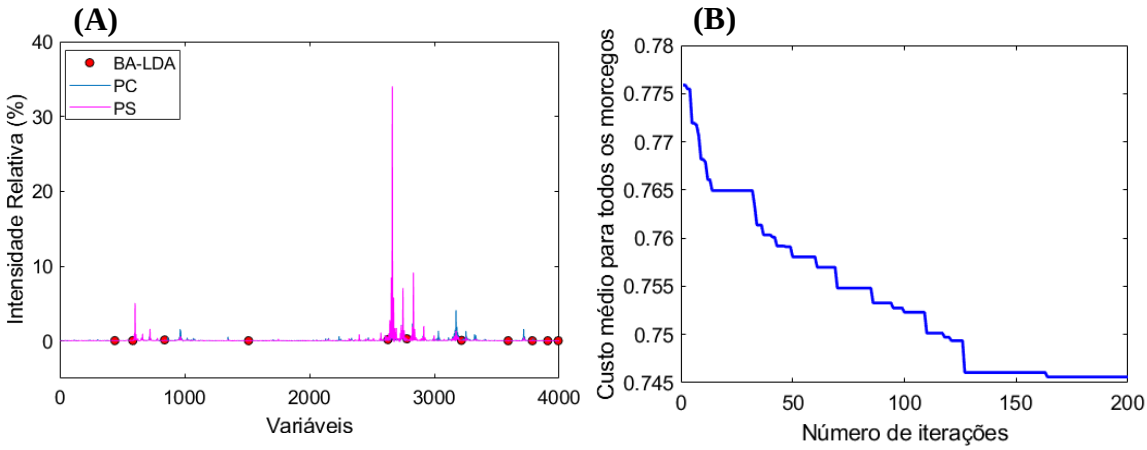
	α	***γ***	***mbats***	***N***
Parâmetros usados	0,5	0,4	30	200

Fonte: (própria).

5.1.2 Aplicação do BA-LDA

Para classificação dos dados em amostras de pacientes com a doença e amostras de pacientes saudáveis, o BA-LDA selecionou 11 variáveis como mostrado na **Figura 21A**. Como já mencionado, estes dados biológicos são complexos e uma interpretação detalhada é difícil de ser realizada. Conrads *et al.*, (2004) sugerem que a informação proteômica sérica diagnóstica, existe em pequenas proteínas e peptídeos (CONRADS *et al.*, 2004).

Figura 21- A) Espectros de massas de soro de pacientes com câncer de ovário (PC) e pacientes sem a doença (PS) e variáveis selecionadas pelo BA-LDA. B) Gráfico do custo médio para todos os morcegos ao longo das iterações.



Fonte: (própria).

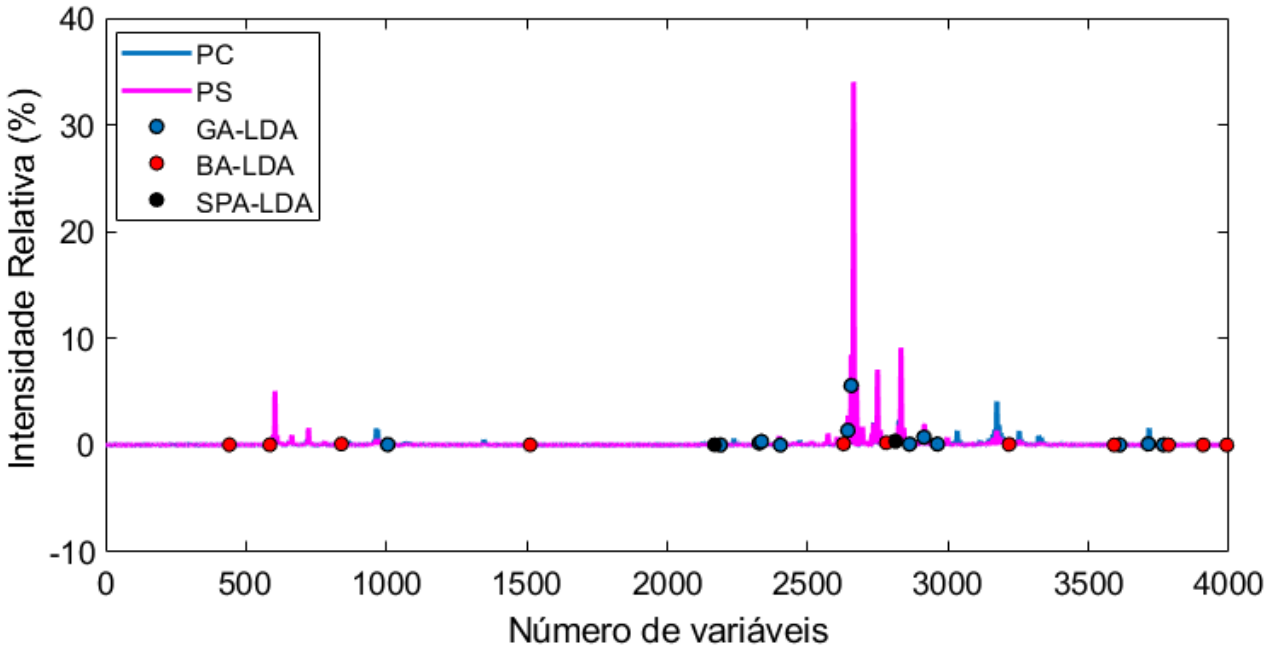
A **Figura 21B** mostra o gráfico do custo médio para todos os morcegos a cada iteração. A partir deste gráfico não foi possível verificar se o algoritmo convergiu para o melhor subconjunto de variáveis, ou se, com a elevação do número de iterações o custo ainda seria reduzido. Todavia, percebeu-se uma boa redução do custo e provavelmente o algoritmo já estaria próximo da melhor solução, tendo em vista que as reduções acentuadas do custo envolveram as cem primeiras iterações. Além disso, conforme o planejamento fatorial executado, foi possível constatar que um maior número de iterações não influenciava significativamente na obtenção de uma melhor resposta.

Com o subconjunto de variáveis selecionadas (**Figura 21A**) o BA-LDA obteve uma taxa de classificação correta de 93 %. As **Seções 5.1.3** e **5.1.4** comparam o desempenho do BA-LDA com outros métodos de classificação, assim, a **Tabela 6** resume os resultados obtidos com os diferentes métodos.

5.1.3 Comparação do desempenho do BA-LDA com outros algoritmos de seleção de variáveis

A **Figura 22**, mostra as variáveis selecionadas pelos algoritmos BA-LDA, GA-LDA e SPA-LDA. O BA-LDA selecionou 11 variáveis, o GA-LDA selecionou 14 e o SPA-LDA apenas 2. O BA-LDA apresentou o melhor desempenho de classificação com 93% de TCC para as amostras externas, seguido do GA-LDA com 88,37 %. O SPA-LDA apresentou uma TCC de apenas 79,07%. Esses resultados são apresentados na **Tabela 6**.

Figura 22- Variáveis selecionadas pelos algoritmos BA-LDA, GA-LDA e SPA-LDA.



Fonte: (própria).

É importante destacar que o SPA-LDA demorou cerca de 10 minutos para executar esses dados. Isso pode ser justificado devido à enorme quantidade de variáveis presentes. Devido a heurística estocástica, os algoritmos GA-LDA e BA-LDA levaram apenas alguns segundos para executarem os dados, fornecendo o resultado da classificação.

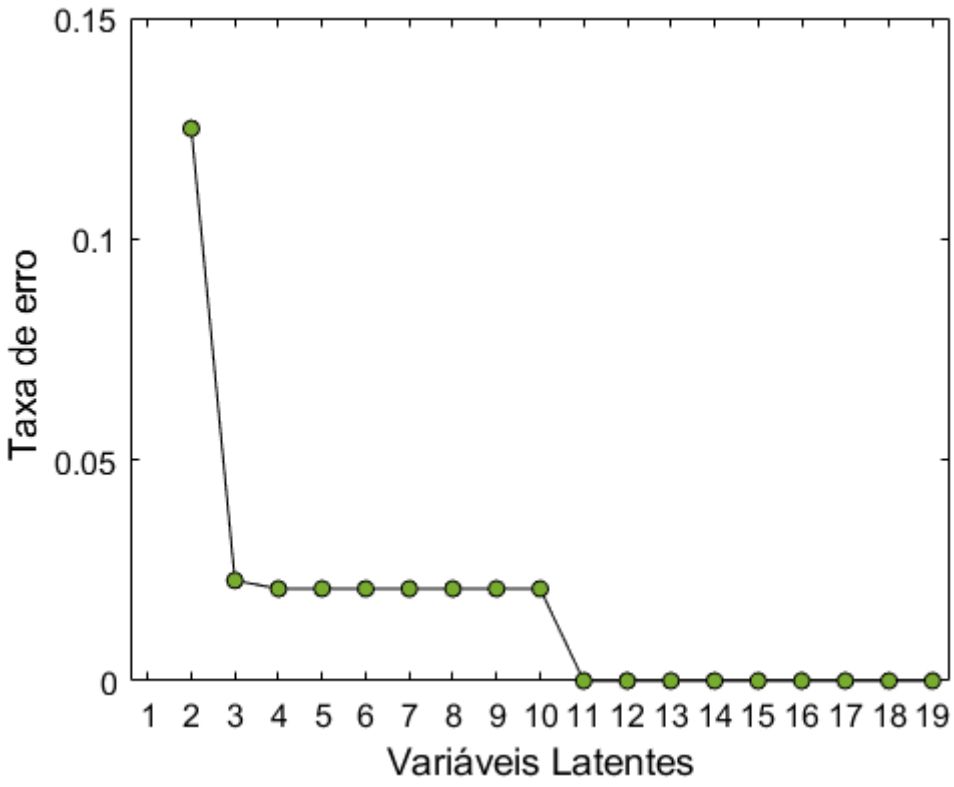
5.1.4 Desempenho do PLS-DA e da classificação SIMCA

A modelagem SIMCA e o PLS-DA são métodos com grande aplicabilidade para realizar a classificação multivariada. Nesta Tese, esses métodos foram usados para comparação de desempenho com o algoritmo proposto.

5.1.4.1 PLS-DA

A **Figura 23** apresenta o número ótimo de variáveis latentes usado para construir o modelo PLS-DA. Como pode ser observado, com onze variáveis latentes a taxa de erro foi nula para validação interna do modelo. Assim, esse número de variáveis latentes foi usado para avaliar a classificação de amostras externas.

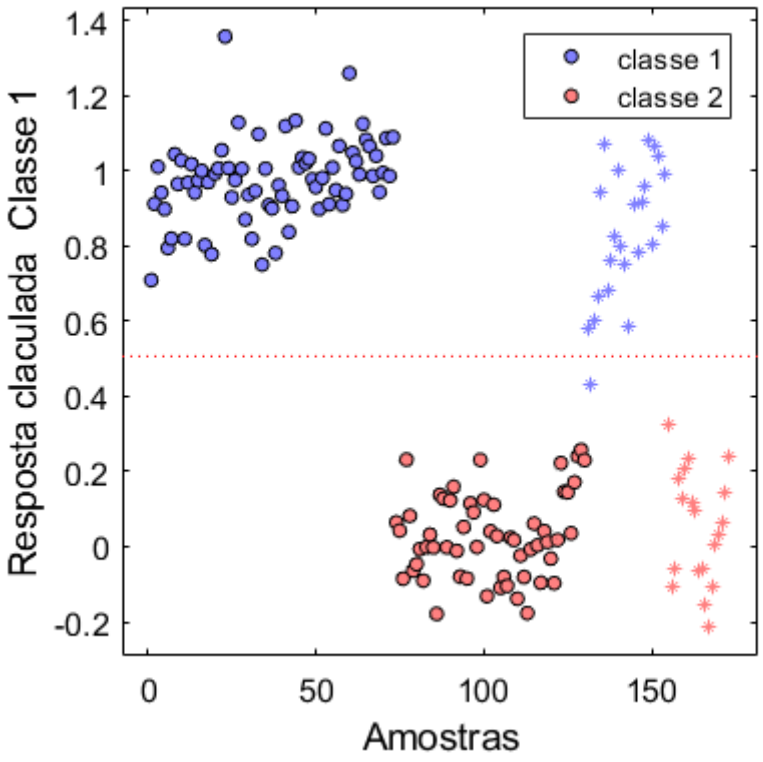
Figura 23- Número ótimo de variáveis latentes pelo PLS-DA.



Fonte: (própria).

A **Figura 24** mostra o resultado da predição das amostras de teste no modelo PLS-DA construído com as onze variáveis latentes.

Figura 24- Predição das amostras externas no modelo PLS-DA.



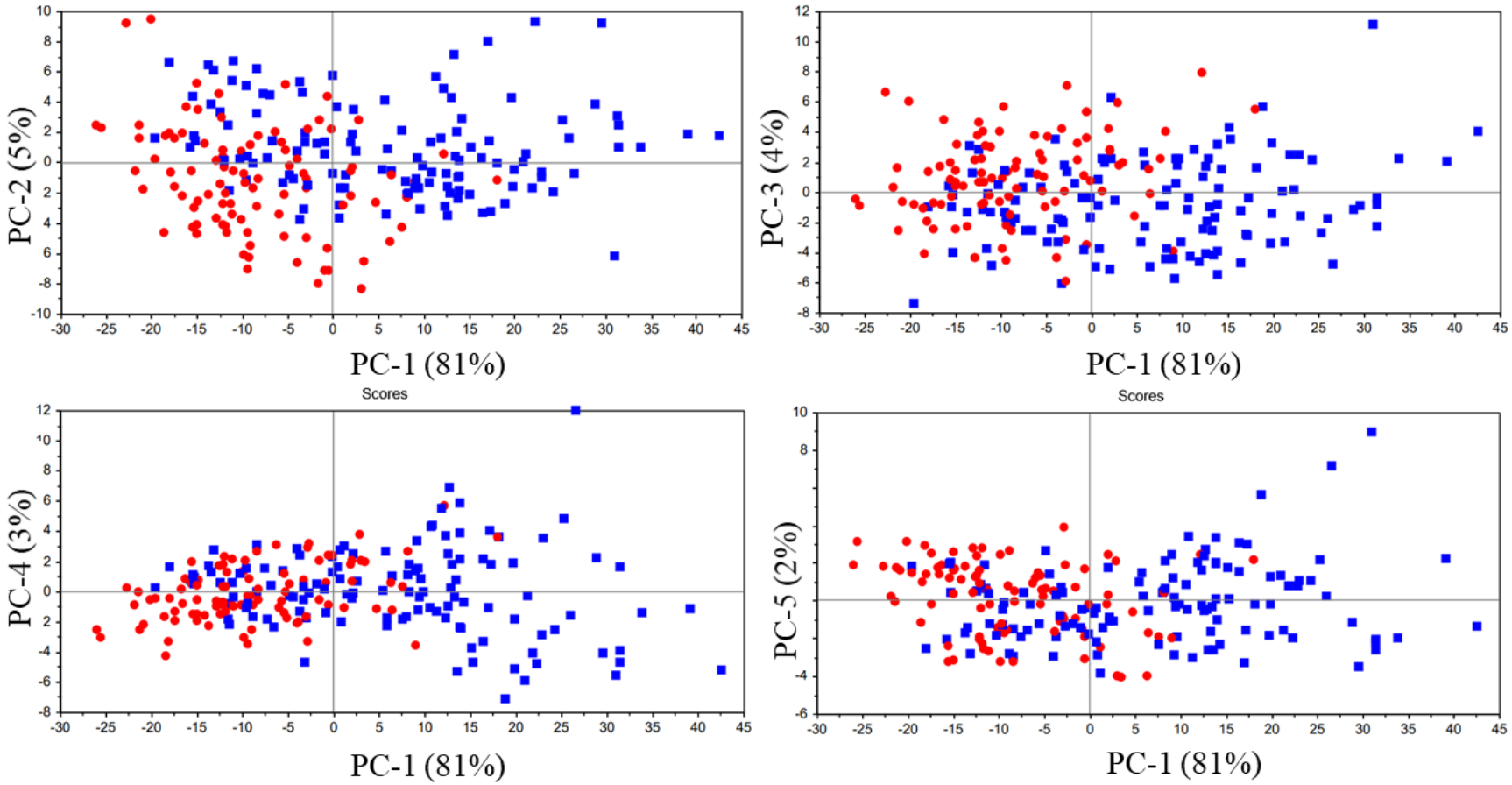
Fonte: (própria).

Na **Figura 24** os "asteriscos" azuis representam as amostras de teste pertencentes a classe 1 (PC) e os em vermelho representam as amostras de teste pertencentes a classe 2 (PS). Como pode ser observado, uma das amostras de teste da classe 1 encontra-se dentro do limite para a classe 2. De fato, uma amostra de teste pertencente a classe 1 foi atribuída incorretamente à classe 2, resultando em uma TCC de 98% (**Tabela 6**). A **Seção 5.1.4.2** apresenta os resultados da classificação SIMCA.

5.1.4.2 SIMCA

Antes de realizar a classificação SIMCA, o comportamento dos dados foi avaliado a partir da análise exploratória. Para isso, foi realizada uma PCA no conjunto total de amostras, o número de PCs usadas foi igual a cinco. A **Figura 25** mostra os gráficos de escores das PCs (PC1 versus a PC2, PC3, PC4 e PC5, que explicam 95% da variância dos dados).

Figura 25- Gráficos dos escores da PCA das amostras de soro de pacientes com câncer de ovário (em azul) e sem a doença (em vermelho).



Fonte: (própria).

Como pode ser observado na **Figura 25**, houve sobreposição entre as classes de amostras de pacientes com e sem a doença. Assim, apenas com o modelo PCA não foi possível fazer uma distinção entre as classes. Para a classificação SIMCA, usou-se modelos PCA individuais construídos para cada classe e o desempenho da classificação foi avaliado em quatro níveis de significância do *teste-F*. A classe 1 (PC- pacientes com câncer) foi modelada com um número de PCs igual a três (PC1-79% , PC2-5% e PC3-5%) e a classe 2 (PS- pacientes sem a doença) foi modelada com um número de PCs igual a quatro (PC1-70%, PC2-13%, PC3-6%, PC4-2%). A **Tabela 7** na **Seção 5.1.5** mostra a matriz de confusão e o desempenho da classificação SIMCA.

A partir do método SIMCA, **Tabela 7**, foi possível observar que para os diferentes níveis de significância houveram amostras que foram incorretamente classificadas. Para o nível de 1%, praticamente todas as amostras PC foram classificadas em sua respectiva classe e também na classe PS e todas as amostras PS foram classificadas em sua respectiva classe e também na classe PC. O melhor desempenho da classificação SIMCA foi obtido com 25% de significância do *teste-F*. Para esse nível de significância, 8 amostras PC foram incorretamente atribuídas a classe PS e 13 amostras da classe PS foram incorretamente atribuídas a classe PC (**Tabela 7)**. A classificação incorreta dessas amostras provavelmente ocorreu devido à complexidade desses dados biológicos e a possibilidade da informação discriminante estar no nível do ruído dos dados. Quando aplicada a seleção de variáveis, características que discriminam as amostras podem ter sido ressaltadas favorecendo o melhor desempenho dos algoritmos de seleção de variáveis para classificação (**Tabela 6** na **Seção 5.1.5**). A redução da dimensionalidade dos dados pelo PLS-DA também favoreceu o desempenho da classificação.

5.1.5 Avaliação geral dos métodos empregados na classificação dos dados de soro de pacientes

Como pode ser observado na **Tabela 6**, o BA-LDA e o PLS-DA apresentaram os melhores desempenhos, 93% e 98% de TCC, respectivamente; O GA-LDA apresentou 88,37% de taxa de classificação correta; O SPA-LDA classificou incorretamente nove amostras da classe PC como sendo PS (79,07 % de TCC); E o SIMCA (na **Tabela 7**) nos diferentes níveis de significância, foi o método que mais classificou incorretamente as amostras. Como já mencionado, a dificuldade na classificação correta das amostras pode ser justificada pela complexidade da matriz biológica. Conforme Conrads *et al.*, (2004) é possível que os espectros

de massa associados à doença compreendam espécies de peptídeo / proteína que podem variar em massa em apenas alguns daltons, dificultando a classificação. Outra hipótese de Conrads *et al*., (2004) é que padrões de diagnóstico do soro são constituídos por marcadores discretos que são um produto do complexo microambiente tumor-hospedeiro. Outra suposição dada pelo autor foi que provavelmente a informação proteômica diagnóstica era parcialmente derivada de proteínas do hospedeiro clivadas, em vez de proteínas que estão diretamente relacionadas à biologia do próprio tumor.

Na **Tabela 6**, foi possível verificar também que nenhum dos algoritmos usados classificou incorretamente amostras de pacientes sem câncer (PS), como sendo de pacientes com câncer (PC). Essa é uma característica importante para o diagnóstico. Além disso, foi possível observar que as variáveis selecionadas pelos métodos BA-LDA, GA-LDA e SPA-LDA quase não apresentaram colinearidade, onde os números de condição obtidos foram baixos.

Em resumo, o BA-LDA demonstrou um bom desempenho, selecionando poucas variáveis e obtendo uma TCC de 93%, superando o desempenho do GA-LDA e SPA-LDA. Sendo o método do PLS-DA o que levou ao melhor resultado. Esses métodos citados superaram o desempenho do SIMCA nos quatro níveis de significância avaliados. Assim, o algoritmo proposto é promissor para ser aplicado a dados biológicos para realizar a seleção de variáveis para classificação de soro de pacientes com câncer de ovário e amostras de soro de pacientes sem a doença. A **Seção 5.1.6** apresenta o estudo da robustez realizado com os algoritmos estocásticos BA-LDA e GA-LDA.

Tabela 6: Resultados obtidos pelos diferentes métodos para a classificação de soro de pacientes com câncer de ovário de sem a doença (série de teste).

		BA-LDA		GA-LDA		SPA-LDA		PLS-DA	
		PC	PS	PC	PS	PC	PS	PC	PS
Classes	PC	21	3	19	5	15	9	23	1
	PS	0	19	0	19	0	19	0	19
Sensibilidade (%)		87.5		79.0		62.5		96	
Seletividade (%)		100		100		100		100	
TCC (%)		93		88.4		79.1		98	
Número de variáveis selecionadas ou variáveis latentes		11		14		2		11	
Número de condição		11.4		53.9		3.9		----	

PC-pacientes com câncer de ovário; PS- pacientes sem a doença.

Fonte: (própria).

Tabela 7: Resultados obtidos pelo método SIMCA para classificação de soro de pacientes com câncer de ovário e sem a doença.

		SIMCA 1%		SIMCA 5%		SIMCA 10%		SIMCA 25%	
		PC	PS	PC	PS	PC	PS	PC	PS
Duas classes	PC	24	22	24	17	24	15	24	8
	PS	19	19	19	19	19	19	13	19
Nenhuma		---		---		---		---	

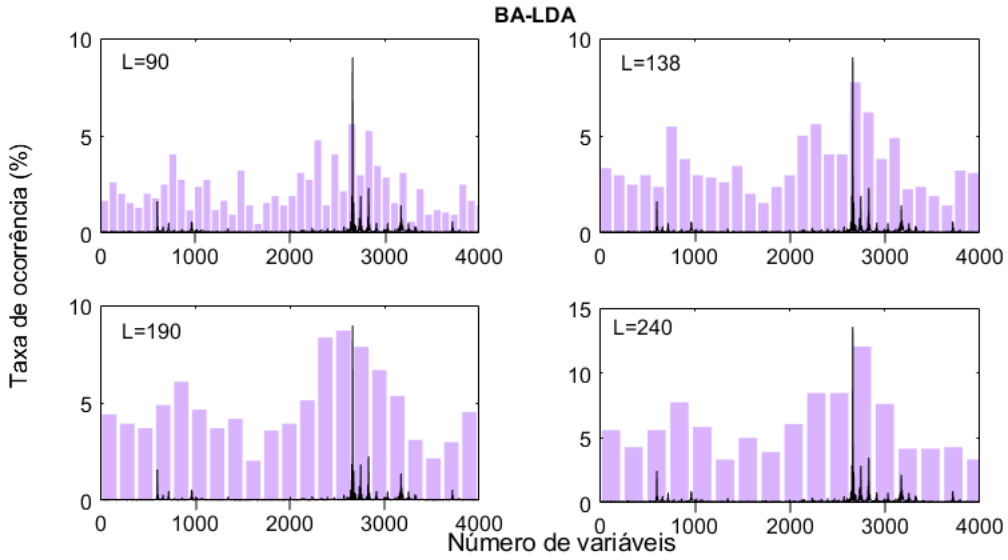
Fonte: (própria).

5.1.6 Avaliação da Robustez

5.1.6.1 BA-LDA

A **Figura 26** apresenta a taxa de ocorrência das variáveis selecionadas pelo BA-LDA para cem execuções do código. Foram empregadas quatro larguras de faixas distintas (90, 138, 190 e 240). O primeiro valor de largura de faixa foi escolhido dividindo-se o valor da amplitude pelo número de classes. Os outros valores foram escolhidos arbitrariamente.

Figura 26- Histogramas com diferentes larguras de faixas (90, 138, 190 e 240) para as variáveis selecionadas pelo BA-LDA em cem repetições do algoritmo.



Fonte: (própria).

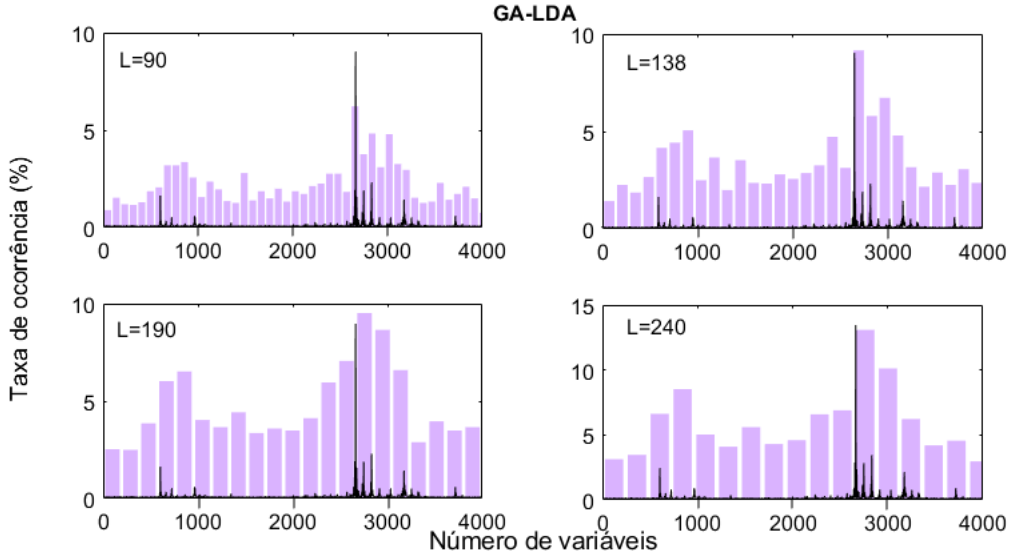
Conforme a **Figura 26**, foi possível verificar que independente da largura de faixa empregada o algoritmo convergiu para a região próxima a variável 3000. Como se pode observar a região de convergência possui o pico base, porém, próximo a este encontram-se picos de baixa intensidade que podem conter a informação discriminante. A **Seção 5.1.6.2** apresenta um estudo de robustez realizado com o algoritmo GA-LDA.

5.1.6.2 GA-LDA

A **Figura 27** mostra a taxa de ocorrência das variáveis selecionadas pelo GA-LDA para as diferentes larguras de faixa empregadas. É possível verificar que da mesma maneira que o BA-LDA, o GA-LDA também convergiu para a seleção de variáveis próximo ao pico de maior abundancia. Os histogramas do BA-LDA e GA-LDA ficaram bem parecidos, isso corrobora

com a afirmativa de Yang (2010), de que os algoritmos estocásticos podem convergir para soluções únicas com uma dada precisão.

Figura 27- Histogramas com diferentes larguras de faixas (90, 138, 190 e 240) para as variáveis selecionadas pelo GA-LDA em cem repetições do algoritmo.



Fonte: (própria).

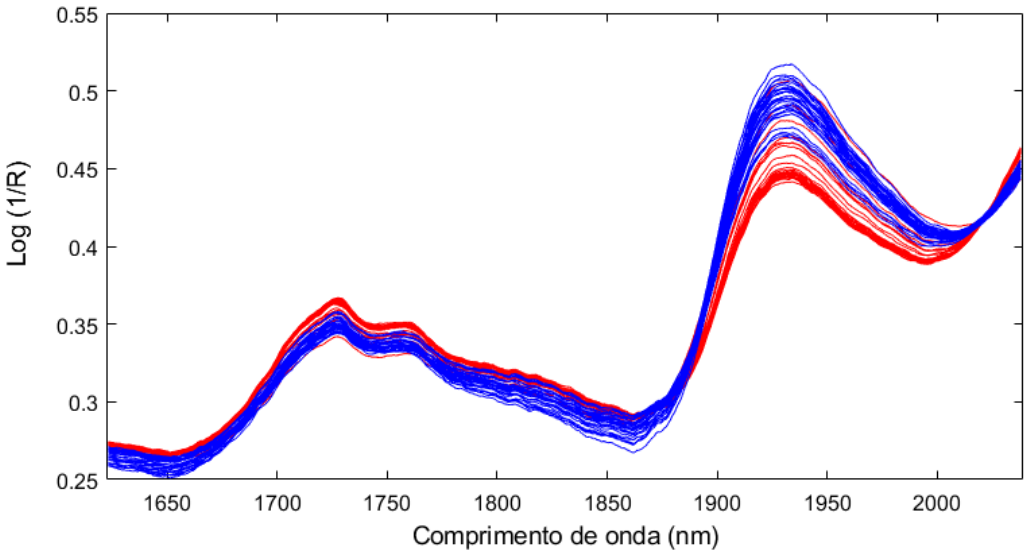
Considerando que os dois algoritmos estocásticos (GA-LDA e BA-LDA) conduziram para a seleção de variáveis na mesma região, é possível que esta contenha informação discriminante o que levou aos resultados de classificação discutidos. Com estas análises, foi possível verificar que o algoritmo proposto pode ser aplicado para classificar matrizes complexas com dados de amostras biológicas, superando o desempenho de algoritmos como GA-LDA e SPA-LDA. A **Seção 5.2** apresenta os resultados para um estudo de caso envolvendo a classificação de amostras de cafés.

5.2 Estudo de caso envolvendo a classificação NIR de cafés

A **Figura 28** mostra os espectros NIR de amostras de cafés pertencentes a duas classes (30 amostras de cafés *gourmet*- espectros em vermelho; 30 amostras de cafés tradicionais – espectros em azul). Os espectros para as duas classes de cafés são muito semelhantes, o que torna necessária a seleção de comprimentos de onda mais discriminativos. Devido ao alto ruído presente no espectro de refletância, o pré-processamento de suavização Savitzky-Golay usando uma janela de 31 pontos e polinômio de segunda ordem foi aplicado. Além disso, o espalhamento da luz pelas partículas do pó de café foi reduzido usando a técnica de correção baseada em multiplicação de sinais de espalhamento (MSC). A banda presente na região entre

1900 e 2000 nm pode ser atribuída ao primeiro sobreton de C=O da cafeína e ácido clorogênico (FERNANDES, D. *et al.*, 2014; OKUBO; KURATA, 2019).

Figura 28- Espectros de refletância NIR para as duas classes de cafés (cafés gourmet em vermelho e cafés tradicionais em azul).



Fonte: (própria).

Antes da aplicação do BA-LDA nestes dados, realizou-se a otimização dos parâmetros a partir de um planejamento fatorial fracionado.

5.2.1 Otimização dos parâmetros do BA-LDA

Para avaliação da significância dos parâmetros do BA-LDA um planejamento fatorial fracionado 2^{4-1} com uma repetição, foi realizado. A **Tabela 8** mostra as respostas obtidas em termos de %TCC para cada um dos experimentos.

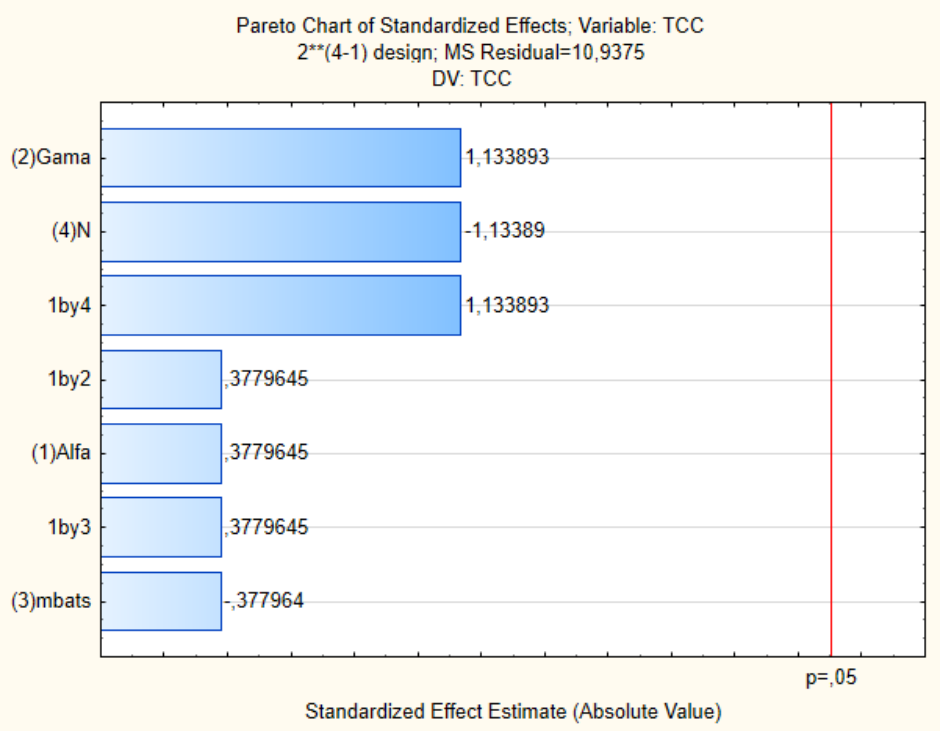
Tabela 8: Matriz de respostas no planejamento fatorial fracionado 2 $^{4-1}$.

Replicadas	α	γ	*mbats*	*N*	*%TCC*
1	0,5	0,4	30	200	100
1	0,9	0,4	30	500	100
1	0,5	0,9	30	500	95
1	0,9	0,9	30	200	100
1	0,5	0,4	50	500	100
1	0,9	0,4	50	200	95
1	0,5	0,9	50	200	100
1	0,9	0,9	50	500	100
2	0,5	0,4	30	200	100
2	0,9	0,4	30	500	95
2	0,5	0,9	30	500	100
2	0,9	0,9	30	200	100
2	0,5	0,4	50	500	90
2	0,9	0,4	50	200	100
2	0,5	0,9	50	200	100
2	0,9	0,9	50	500	100

Fonte: (própria).

Na **Tabela 8** foi possível verificar que a maioria dos experimentos levaram a uma taxa de classificação correta de 100%. A **Figura 29** mostra o gráfico de Pareto dos efeitos.

Figura 29- Gráfico de Pareto dos efeitos para os quatro fatores no planejamento fracionado 2$^{4-1}$.



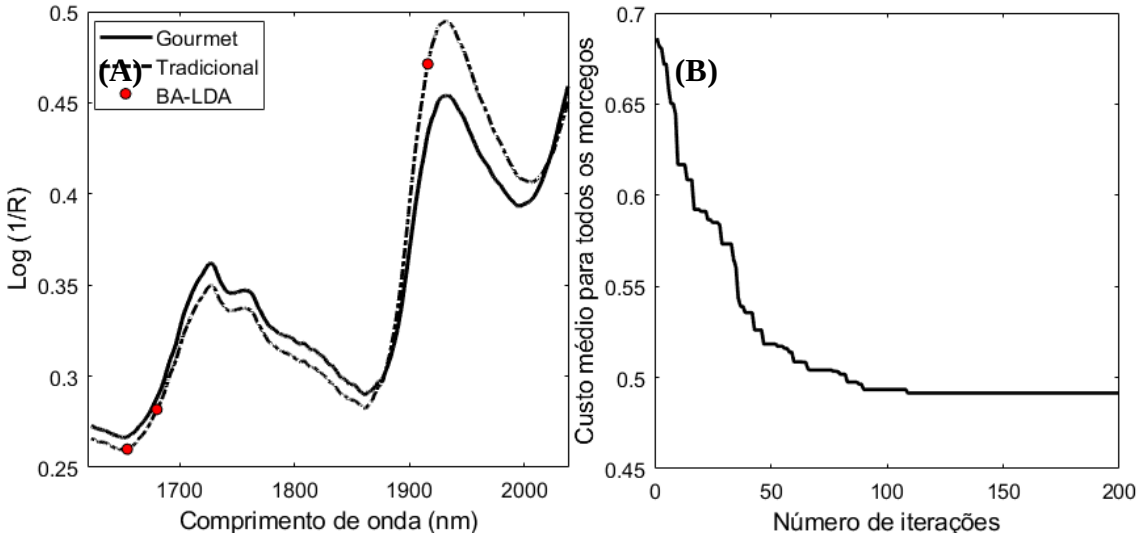
Fonte: (própria).

Na **Figura 29** foi possível observar que nenhum dos fatores nos níveis estudados foram significativos, portanto, adotou-se arbitrariamente os parâmetros α=0,5, γ=0,4, *mbats*= 30, *N*=200, conforme destacado em cinza na **Tabela 8**. A **Seção 5.2.2** apresenta o desempenho da aplicação do BA-LDA nos dados de cafés, empregando estes parâmetros.

5.2.2 Aplicação do BA-LDA

O BA-LDA foi aplicado ao conjunto de dados de espectros NIR de amostras de cafés *gourmet* e tradicional. As **Figuras 30A** e **30B** apresentam, respectivamente, as variáveis selecionadas pelo BA-LDA e o custo médio para todos os morcegos ao longo das iterações que levaram à melhor classificação LDA em cinco repetições executadas.

Figura 30- A) Espectros médios das classes de cafés *gourmet* e tradicionais e variáveis selecionadas pelo BA-LDA. B) Gráfico do custo médio para todos os morcegos ao longo das iterações.



Fonte: (própria).

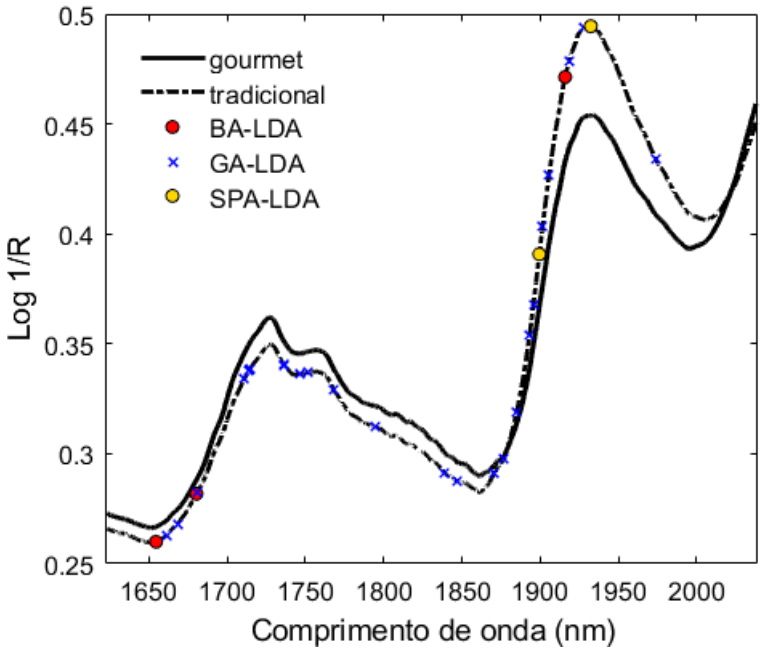
Na **Figura 30A**, pode-se observar que as variáveis selecionadas pelos morcegos virtuais encontram-se em regiões informativas com características químicas das amostras, a exemplo, a variável selecionada entre 1900 - 2000 nm que está relacionada ao primeiro sobreton de C=O da cafeína e ácido clorogênico (FERNANDES, D. *et al.*, 2014; OKUBO; KURATA, 2019). Como pode ser observado na **Figura 30B**, na primeira iteração o custo médio do erro de classificação é alto para as variáveis selecionadas. No decorrer das iterações, esse custo vai reduzindo e isso se justifica pelas atualizações (**Equações 34-36**), as quais conduzem a novas posições para os morcegos virtuais e melhores subconjuntos de variáveis selecionadas. Após executadas cerca de 110 iterações para esse conjunto de dados, foi possível perceber que os

morcegos convergiram para uma única solução (**Figura 30B**). Isso significa que as variáveis selecionadas na referida iteração levaram ao menor custo e a partir daí não foi encontrado nenhum subconjunto de variáveis capaz de superar o desempenho destas. Assim, com esse melhor subconjunto de variáveis o modelo LDA foi obtido. Como principal resultado, todas as amostras desconhecidas (amostras de teste) foram classificadas corretamente.

5.2.3 Comparação do desempenho do BA-LDA com outros algoritmos de seleção de variáveis

As variáveis selecionadas pelo BA-LDA e GA-LDA que levaram à melhor classificação em cinco repetições são apresentadas na **Figura 31**, bem como as variáveis selecionadas pelo SPA-LDA e os espectros médios para as duas classes de cafés.

Figura 31- Espectros médios de cafés e variáveis selecionadas pelos algoritmos BA-LDA, GA-LDA e SPA-LDA.



Fonte: (própria).

Pode-se verificar (**Figura 31**) que os espectros para as duas classes de cafés são muito semelhantes e possuem um grande número de comprimentos de onda (um total de 1301 variáveis), sendo importante selecionar os mais discriminantes. Como já discutido, o BA-LDA selecionou três variáveis, uma variável na região entre 1850 e 1950 nm que pode ser relacionada aos primeiros sobretons de C = O de cafeína e ácido clorogênico (FERNANDES, D. *et al.*, 2014; OKUBO; KURATA, 2019). O SPA-LDA selecionou duas variáveis também na região

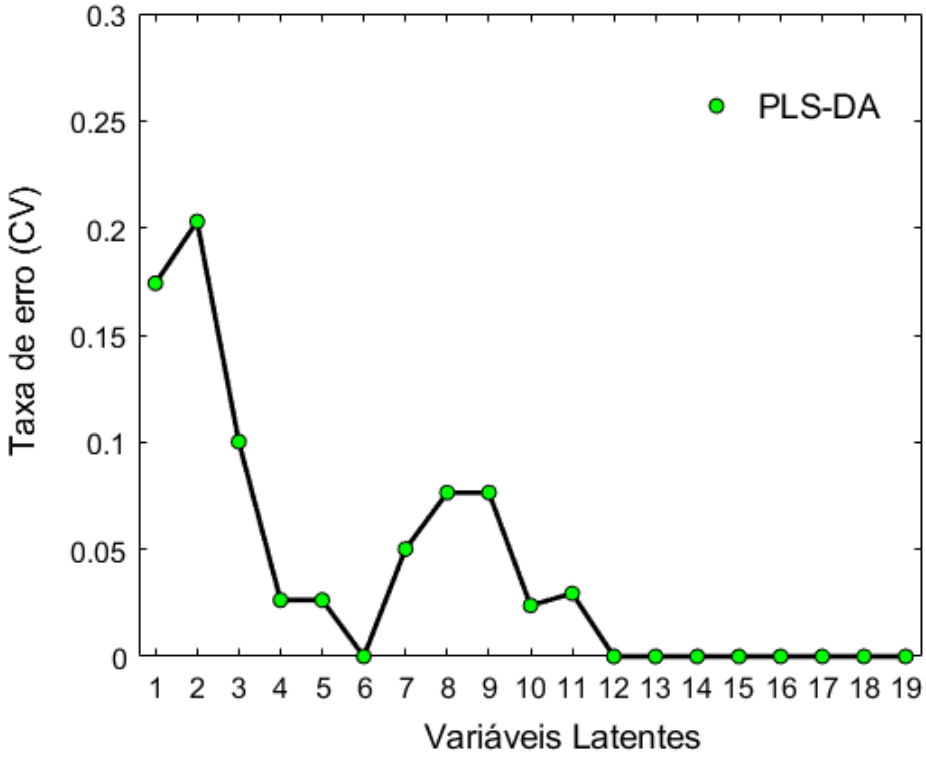
entre 1850 e 1950 nm e o GA-LDA selecionou vinte e quatro variáveis em todo o espectro. Os três algoritmos de seleção de variáveis obtiveram uma TCC de 100%, conforme apresentado na **Tabela 9** na **Seção 5.2.5**. A **Seção 5.2.4** apresenta os resultados obtidos usando o PLS-DA e o SIMCA.

5.2.4 Desempenho do PLS-DA e da classificação SIMCA

5.2.4.1 PLS-DA

A **Figura 32** apresenta o gráfico com as variáveis latentes utilizadas para construir o modelo PLS-DA. Para um modelo PLS-DA otimizado, o número de variáveis latentes foi escolhido com base no menor erro obtido. Assim, o número de variáveis latentes igual a seis foi escolhido e utilizado após validação cruzada nas amostras de treinamento.

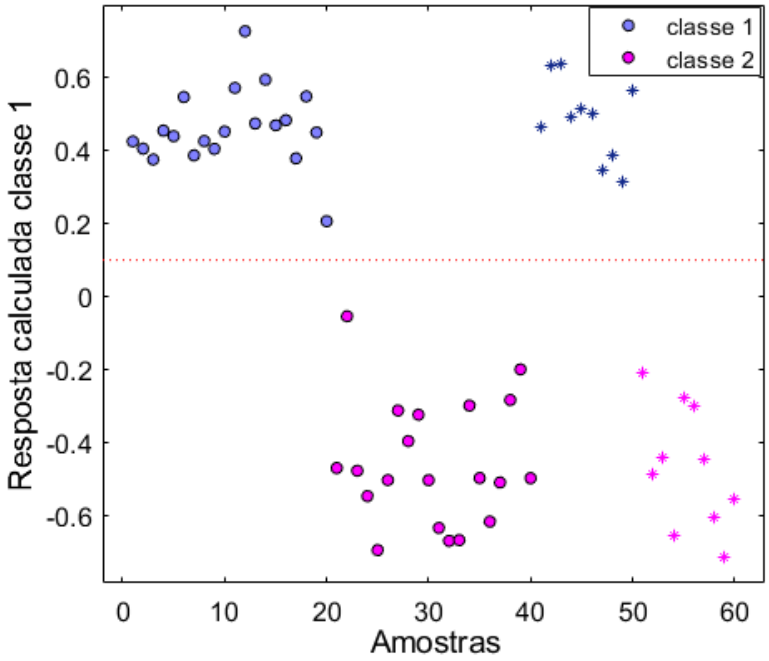
Figura 32- Número ótimo de variáveis latentes pelo PLS-DA.



Fonte: (própria).

Com as seis variáveis latentes o algoritmo PLS-DA obteve uma TCC de 100% para as amostras de cafés *gourmet* e tradicionais. A **Figura 33** mostra o resultado da predição das amostras de teste no modelo PLS-DA construído com as seis variáveis latentes.

Figura 33- Predição das amostras externas no modelo PLS-DA.



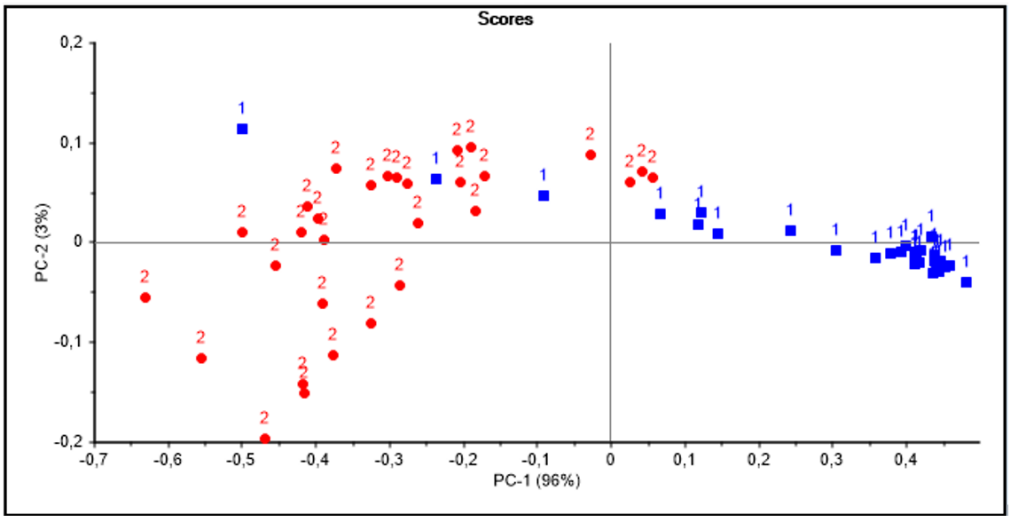
Fonte: (própria).

Na **Figura 33** as amostras na cor azul pertencem a classe 1 (cafés *gourmet*) e as amostras na cor rosa representam a classe 2 (cafés tradicionais). Como pode ser observado, todas as amostras externas (asteriscos azuis pertencentes a classe 1 e rosas pertencente a classe 2) forma atribuídas as suas respectivas classes. A **Seção 5.2.4.2** apresenta os resultados da classificação SIMCA para este conjunto de dados.

5.2.4.2 SIMCA

Antes da classificação SIMCA, inicialmente foi realizada uma PCA geral para observar o comportamento dos dados. A **Figura 34**, apresenta a PCA realizada com um número de 2 PCs que explicaram 99% da variância dos dados.

Figura 34- Gráfico dos escores da PCA das amostras de cafés *gourmet* (em azul) e tradicionais (em vermelho).



Fonte: (própria).

Na **Figura 34** foi possível verificar que três amostras da classe 1 encontravam-se dispersas na classe 2 e que para as demais amostras ocorreu uma separação entre as classes. Modelos de PCA para as classes individuais foram construídos. Para a classe 1 usou-se uma componente principal e para a classe 2 o número de componentes foi igual a dois. Pela **Tabela 10**, na classificação SIMCA, foi possível verificar que para 1% de significância do *teste-F* três amostras da classe 1 foram atribuídas como pertencentes a classe 2. Esse comportamento poderia ser esperando levando em consideração a disposição das amostras observadas na PCA geral. Ainda sobre o nível de significância de 1%, uma amostra da classe 2 foi atribuída a classe 1. Para os níveis de 5 e 10 % duas amostras da classe 1 foram atribuídas a classe 2 e o melhor desempenho foi obtido para o nível de significância de 25% onde apenas uma amostra da classe 1 foi atribuída a classe 2 e uma amostra da classe 2 foi atribuída a classe 1.

5.2.5 Avaliação geral dos métodos empregados na classificação dos dados de cafés

Na **Tabela 9** foi apresentada uma matriz de confusão com os resultados da classificação das amostras de café em um conjunto de teste independente. Para todos os algoritmos de seleção de variáveis e para o PLS-DA, 100% das taxas de sensibilidade, especificidade e classificação correta foram obtidas para o conjunto de teste mencionado. Assim, todas as amostras foram classificadas corretamente em suas respectivas classes. Isso significa que o desempenho do BA-LDA é comparável ao algoritmo estocástico tradicionalmente empregado, GA-LDA, e ao

algoritmo determinístico SPA-LDA, bem como ao PLS-DA. Em relação ao número de variáveis selecionadas os algoritmos foram parcimoniosos, sendo o GA-LDA o que selecionou o maior número de variáveis. Além disso, o BA-LDA selecionou variáveis com menor valor para o número de condição quando comparadas às selecionadas pelo GA-LDA, indicando menor presença de multicolinearidade. Quando comprado a classificação SIMCA, nos quatro níveis de significância do *teste-F*, o BA-LDA demonstrou desempenho superior. Dessa forma, o algoritmo proposto nesta Tese demonstrou potencial para ser aplicado em dados NIR para classificação de cafés. A **Seção 5.2.6** apresenta um estudo da robustez realizado com os algoritmos estocásticos BA-LDA e GA-LDA.

Tabela 9: Resultados obtidos pelos diferentes métodos para a classificação de cafés *gourmet* (C1) e tradicionais (C2).

		BA-LDA		GA-LDA		SPA-LDA		PLS-DA	
		C1	C2	C1	C2	C1	C2	C1	C2
Duas classes	C1	10	0	10	0	10	0	10	0
	C2	0	10	0	10	0	10	0	10
Nenhuma									
Sensibilidade (%)		100		100		100		100	
Seletividade (%)		100		100		100		100	
Número de variáveis selecionadas ou número ótimo de variáveis latentes		3		24		2		6	
TCC (%)		100		100		100		100	
Número de condição		280.3		4.9×10^4		125.3		-------	

Fonte: (própria).

Tabela 10: Resultados obtidos pelo método SIMCA para classificação de cafés *gourmet* (C1) e tradicionais (C2).

		SIMCA 1%		**SIMCA 5%**		**SIMCA 10%**		**SIMCA 25%**	
		C1	**C2**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
Duas classes	**C1**	10	**3**	10	**2**	10	**2**	10	**1**
	C2	**1**	10	**1**	10	**1**	10	**1**	10
	Nenhuma								

C1 – *gourmet*; C2- tradicional.

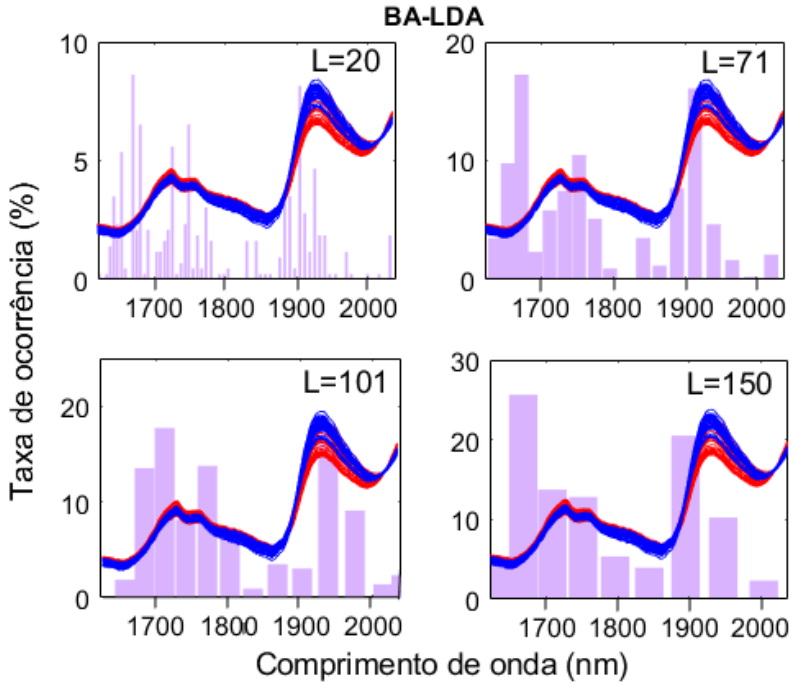
Fonte: (própria).

5.2.6 Avaliação da Robustez

5.2.6.1 BA-LDA

A **Figura 35** apresenta os resultados da avaliação da reprodutibilidade do BA-LDA na seleção das variáveis. Cada histograma tem diferentes larguras de intervalos de variáveis (L = 20, 71, 101 e 150) e representa a taxa de ocorrência das variáveis selecionadas para cem repetições de uso do algoritmo. A partir dos histogramas, pode-se verificar que o BA-LDA convergiu para a seleção de variáveis em duas regiões principais, independentemente da largura do intervalo observado. Como pode ser visto na **Figura 35**, o BA-LDA selecionou mais variáveis nessas duas regiões, correspondendo aos comprimentos de onda próximo a 1900nm e entre 1650 e 1700nm.

Figura 35- Histogramas com diferentes larguras de faixas (20, 71, 101 e 150) para as variáveis selecionadas pelo BA-LDA em cem repetições do algoritmo.



Fonte: (própria).

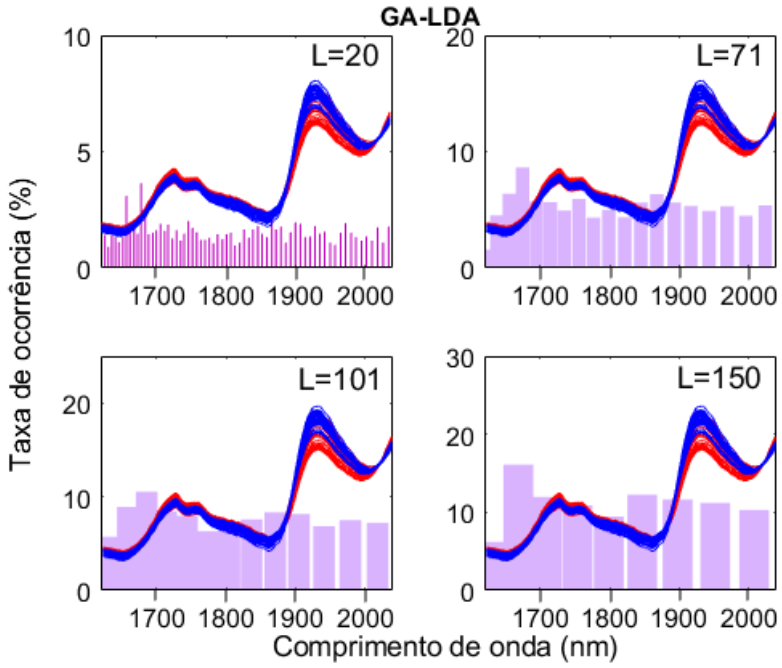
A região entre 1850 e 2000 nm pode estar relacionada aos primeiros tons C = O de cafeína e ácido clorogênico (BARDIN *et al.*, 2014; OKUBO; KURATA, 2019) e a região entre 1650 e 1700 nm pode estar associada aos primeiros sobretons CH de cafeína e ácido clorogênico (RIBEIRO, J. S.; FERREIRA, M. M. C; SALVA, 2011). Assim, o algoritmo proposto converge para a seleção de variáveis associadas às principais informações químicas do espectro. Na

Seção 5.2.6.1, os resultados do estudo de robustez usando o algoritmo GA-LDA foi apresentado.

5.2.6.2 GA-LDA

A **Figura 36** mostra os histogramas com diferentes larguras de intervalo (L = 20, 71, 101 e 150) para a taxa de ocorrência das variáveis selecionadas com cem repetições do algoritmo GA-LDA. Como pode ser visto, o GA-LDA selecionou variáveis espalhadas por todo o espectro. Ou seja, para este conjunto de dados o GA-LDA não demonstrou convergir para regiões informativas dos dados. Em contrapartida, o algoritmo BA-LDA proposto apresentou uma maior densidade de variáveis selecionadas na região próxima a 1700 e na região entre 1900-2000 nm, o que pode indicar certa robustez. Assim, para este conjunto de dados, BA-LDA apresentou melhor reprodutibilidade de variáveis selecionadas quando comparado ao GA-LDA.

Figura 36- Histogramas com diferentes larguras de faixas (20, 71, 101 e 150) para as variáveis selecionadas pelo BA-LDA em cem repetições do algoritmo.



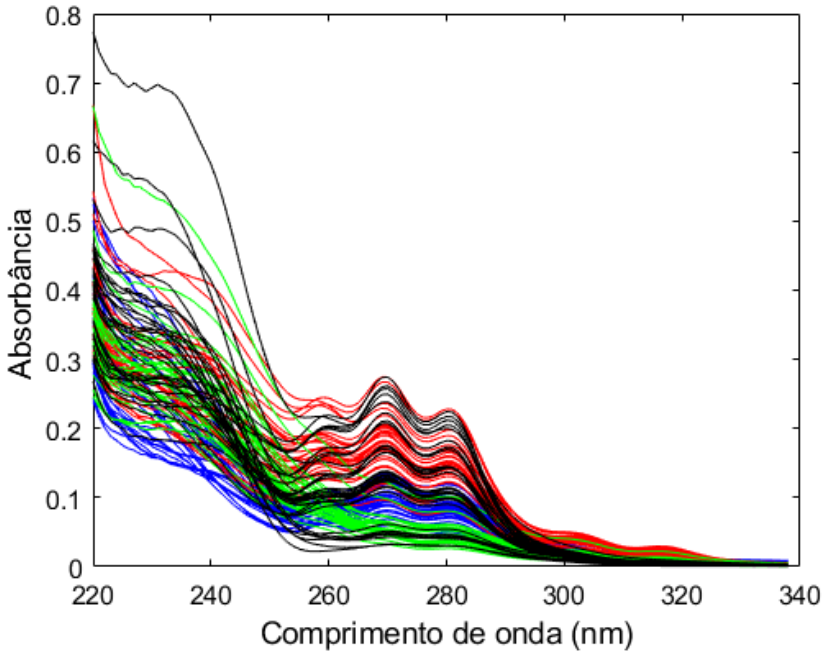
Fonte: (própria).

A **Seção 5.3** apresenta os resultados de outro estudo de caso analisado.

5.3 Estudo de caso envolvendo a classificação UV-Vis de óleos vegetais

Este conjunto de dados foi utilizado para verificar o desempenho do BA-LDA na seleção de variáveis para a discriminação de diferentes tipos de óleos vegetais. Os dados obtidos por Pontes (2009) foram provenientes de medidas no espectrofotômetro ultravioleta-visível (UV-Vis) de quatro classes de óleos vegetais (milho, canola, soja e girassol)(PONTES, M. J. C., 2009). A **Figura 37** apresenta os espectros das quatro classes de óleos vegetais (em azul estão os espectros dos óleos vegetais de milho, em verde os espectros dos óleos vegetais de canola, em vermelho os espectros dos óleos de soja e em preto os espectros dos óleos de girassol).

Figura 37- Espectros UV-Vis das quarto classes de óleos vegetais (óleos vegetais de milho - em azul; os óleos vegetais de canola- em verde; óleos de soja- em vermelho; óleos de girassol- em preto).



Fonte: (própria).

Antes de iniciar a análise destes dados a **Seção 5.3.1** apresenta o estudo da otimização dos parâmetros do BA-LDA.

5.3.1 Otimização dos parâmetros do BA-LDA

A **Tabela 11** apresenta as respostas obtidas em termos de %TCC para os experimentos realizados no planejamento fatorial fracionado 2^{4-1} com uma repetição.

Tabela 11: Matriz de respostas no planejamento fatorial fracionado 2^{4-1}.

Replicadas	**α**	**γ**	***mbats***	***N***	***%TCC***
1	0,5	0,4	30	200	100
1	0,9	0,4	30	500	100
1	0,5	0,9	30	500	100
1	0,9	0,9	30	200	100
1	0,5	0,4	50	500	100
1	0,9	0,4	50	200	100
1	0,5	0,9	50	200	100
1	0,9	0,9	50	500	100
2	0,5	0,4	30	200	100
2	0,9	0,4	30	500	100
2	0,5	0,9	30	500	100
2	0,9	0,9	30	200	100
2	0,5	0,4	50	500	100
2	0,9	0,4	50	200	100
2	0,5	0,9	50	200	100
2	0,9	0,9	50	500	100

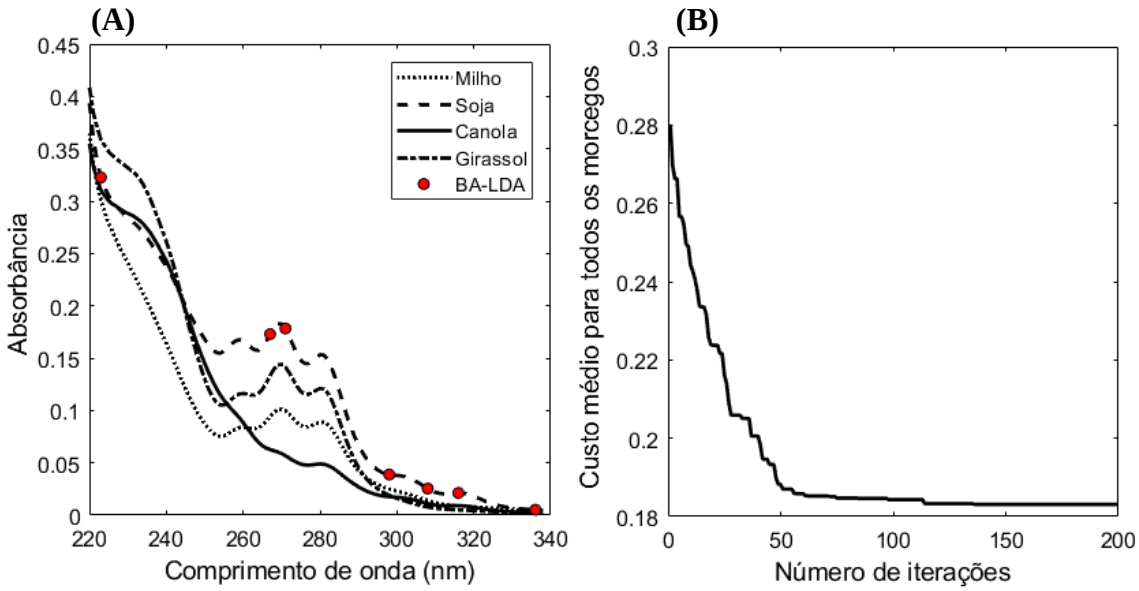
Fonte: (própria).

Na **Tabela 11** foi possível verificar que as respostas obtidas em todos os experimentos foram 100% de taxa de classificação correta. Assim, os fatores, nos níveis estudados, e suas interações não tiveram significância para obtenção das respostas. Com isso, qualquer conjunto de fatores avaliados poderia ser escolhido. Nesta Tese, foram escolhidos os parâmetros α=0,5, γ=0,4, *N*=30, *mbats*=200, que também foram usados nos conjuntos de dados avaliados nas **Seções 5.1** e **5.2**.

5.3.2 Aplicação do BA-LDA

As **Figuras 38A** e **38B** apresentam as variáveis selecionadas pelo BA-LDA e o gráfico do custo médio para todos os morcegos ao longo das iterações, respectivamente.

Figura 38- A) Espectros médios das classes de óleos vegetais e variáveis selecionadas pelo BA-LDA. B) Gráfico do custo médio para todos os morcegos ao longo das iterações.



Fonte: (própria).

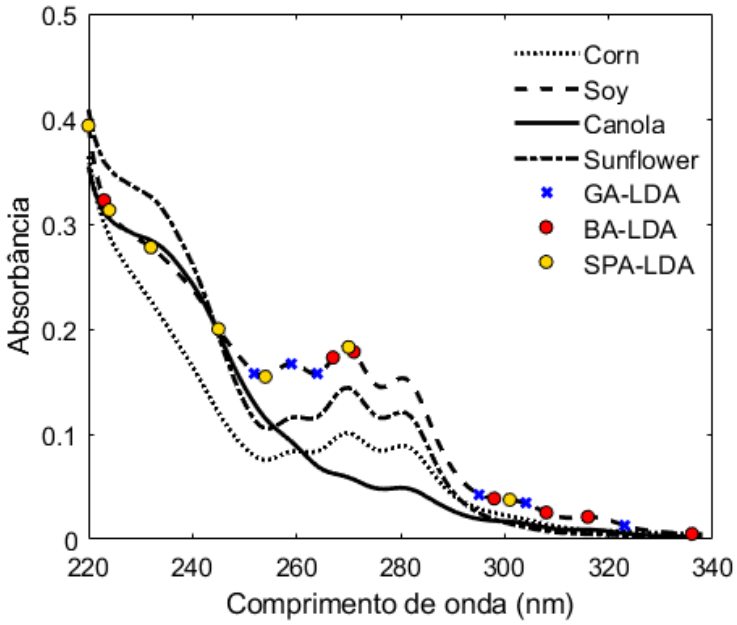
Na **Figura 38A**, foi possível verificar que o BA-LDA selecionou sete variáveis, sendo duas variáveis selecionadas na região entre 260 e 280 nm, que apresentam as bandas mais nítidas para os óleos de girassol, milho e soja quando comparadas ao espectro do óleo de canola. A banda intensa em torno de 232 nm pode estar associada à absorção de dienos conjugados (TOLENTINO *et al.*, 2014) e as demais bandas podem estar associadas a absorção de alguns ácidos graxos que absorvem em regiões abaixo de 375nm (FERREIRA, L. S. *et al.*, 2017). As variáveis selecionadas pelo BA-LDA levaram a um desempenho de 100% de TCC. A **Figura 38B** apresenta o gráfico do custo médio para todos os morcegos a cada iteração do algoritmo. Como o algoritmo classificou corretamente todas as amostras de teste, é provável que tenha alcançado a convergência para o melhor subconjunto de variáveis por cerca da iteração 150, conforme **Figura 38B**. A **Seção 5.3.3** mostra os resultados de outros algoritmos de seleção de variáveis e compara com os resultados do BA-LDA.

5.3.3 Comparação do desempenho do BA-LDA com outros algoritmos de seleção de variáveis

A **Figura 39** apresenta os espectros médios para as quatro classes de óleos vegetais e as variáveis selecionadas pelos algoritmos BA-LDA, GA-LDA e SPA-LDA. Com essas variáreis selecionadas os três respectivos algoritmos resultaram em 100% de sensibilidade,

especificidade e taxas de classificação corretas para o conjunto de teste. Esses resultados significam que todas as amostras foram classificadas corretamente em suas respectivas classes, conforme mostrado na **Tabela 11**.

Figura 39- Espectros médios das classes de óleos vegetais e variáveis selecionadas pelos algoritmos BA-LDA, GA-LDA e SPA-LDA.



Fonte: (própria).

Ainda na **Figura 39**, pode-se verificar que todos os algoritmos selecionaram um pequeno número de variáveis (sete variáveis foram selecionadas pelo BA-LDA, oito variáveis foram selecionadas pelo GA-LDA e sete variáveis foram selecionadas pelo SPA-LDA) demonstrando a capacidade de reduzir a dimensionalidade dos dados e a parcimônia.

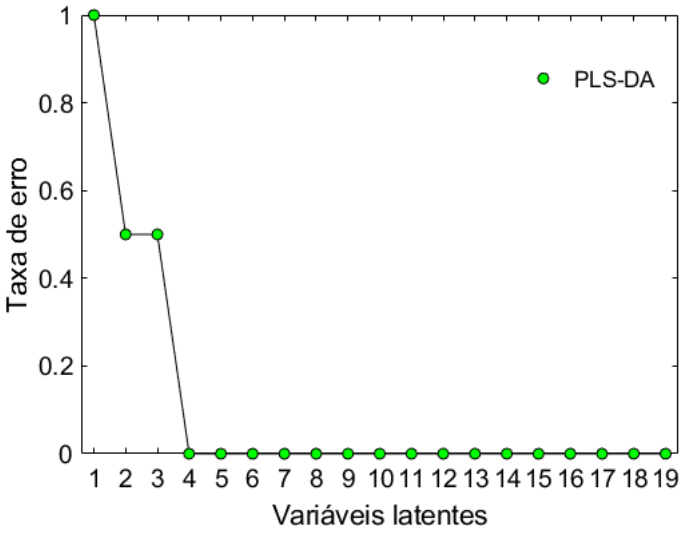
No Trabalho de Pontes (2009), os algoritmos GA-LDA e SPA-LDA também foram aplicados aos dados. A **Tabela 13**, na **Seção 5.3.5** apresenta uma comparação entre os resultados do BA-LDA e os obtidos no trabalho de Pontes (2009). Antes desta comparação, a **Seção 5.3.4** apresenta os resultados de classificação obtidos pelo PLS-DA e pela classificação SIMCA.

5.3.4 Desempenho do PLS-DA e da classificação SIMCA

5.3.4.1 PLS-DA

A **Figura 40** apresenta o gráfico da taxa de erros pelo número de variáveis latentes. A partir desta figura determinou-se um número ótimo de quatro variáveis latentes e construiu-se o modelo PLS-DA.

Figura 40- Número ótimo de variáveis latentes pelo PLS-DA.



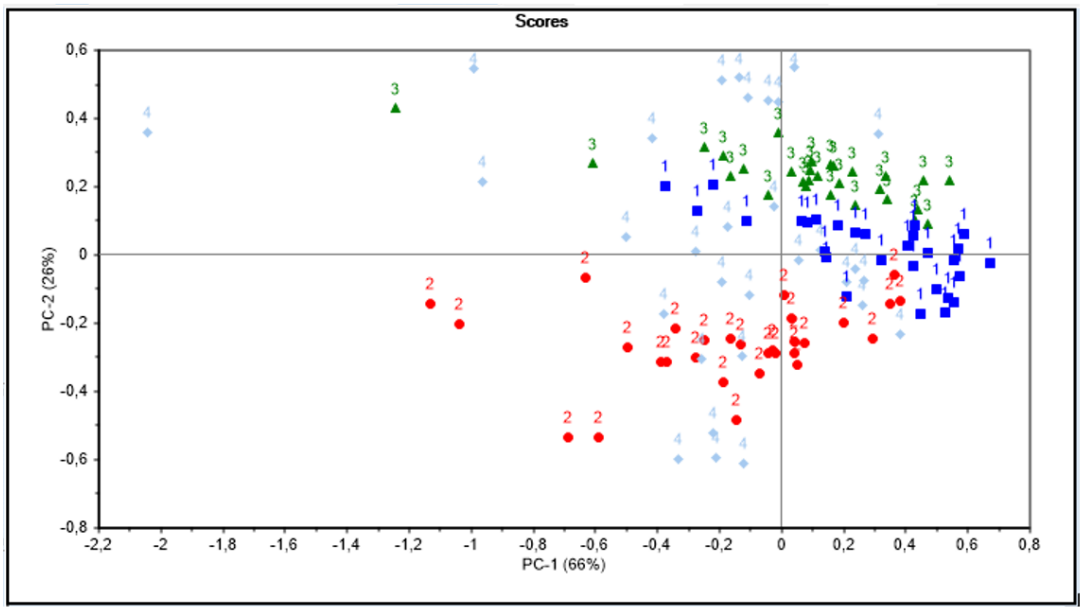
Fonte: (própria).

A partir do modelo PLS-DA construído com as quatro variáveis latentes, foi possível verificar que apenas 95% das amostras de teste foram classificadas corretamente. A **Tabela 12** na **Seção 5.3.5**, mostra os resultados da classificação das amostras de teste obtidos a partir do modelo PLS-DA e a **Tabela 14** apresenta uma comparação com os resultados obtidos no trabalho de Pontes (2009)**.**

5.3.4.2 SIMCA

A **Figura 41** apresenta a PCA geral realizada nos dados de óleos vegetais. A PC1 explicou 66% da variância dos dados e a PC2 explicou 26%. Após avaliar a PCA geral, modelos individuais para cada classe foram construídos antes de realizarmos a classificação SIMCA.

Figura 41- Gráfico dos escores da PCA das amostras de óleos de milho (em azul), óleos de soja (em vermelho), óleos de canola (em verde) e óleos de girassol (em azul claro).



Fonte: (própria).

Na **Figura 41** foi possível verificar que as classes das amostras estavam sobrepostas, assim uma distinção entre elas não foi observada na PCA. Para a classificação SIMCA, a classe 1 (óleos de milho) foi modelada com quatro PCs e as classes 2 (óleos de soja), 3 (óleos de canola) e 4(óleos de girassol) foram modeladas com três PCs. A **Tabela 13** na **Seção 5.3.5** apresenta os resultados da classificação SIMCA. Para os quatro níveis de significância do *teste-F* avaliados, amostras foram incorretamente classificadas. Para o nível de significância de 1%, uma amostra da classe milho foi atribuída na sua verdadeira classe e também na classe de óleos de soja, além disso, as seis amostras da classe óleos de soja também foram classificadas como pertencentes a classe milho; Para os níveis de 5 e 10% de significância, cinco amostras da classe soja foram atribuídas a classe milho e uma amostra da classe canola não foi atribuída a nenhuma das classes; Para o nível de 25% de significância, uma amostra da classe soja foi atribuída a classe milho e duas amostras da classe canola não foram atribuídas a nenhuma das classes. A **Seção 5.3.5** apresenta uma avaliação geral dos métodos empregados para classificação destes dados.

5.3.5 Avaliação geral dos métodos empregados na classificação dos dados de óleos vegetais

A **Tabela 12** resume os resultados obtidos pela aplicação dos algoritmos BA-LDA, GA-LDA, SPA-LDA e PLS-DA para o conjunto de dados de óleos vegetais. A **Tabela 13** apresenta os resultados da classificação SIMCA para quatro níveis de significância do *teste-F* avaliados. Como pode ser observado os algoritmos de seleção de variáveis classificaram corretamente todas amostras do conjunto de teste (**Tabela 12**). No entanto, o PLS-DA resultou em 95% de TCC e apenas 92% de especificidade para óleos de canola e 50% de sensibilidade para óleos de milho, conforme mostrado na **Tabela 12**. Na classificação SIMCA (**Tabela 13**), para os quatro níveis de significância, amostras foram atribuídas incorretamente. Assim, o BA-LDA supera o PLS-DA e o SIMCA para esses dados de óleos vegetais e é semelhante ao GA-LDA e SPA-LDA.

A **Tabela 14** apresenta os resultados obtidos por Pontes (2009) para este mesmo conjunto de dados usando os algoritmos GA-LDA, SPA-LDA e a classificação SIMCA. Para comparação com os resultados obtidos por Pontes (2009), o BA-LDA foi novamente aplicado aos dados, porém, usando a mesma divisão dos conjuntos de treinamento, validação e teste. Pontes (2009) usou a denominação das amostras não classificadas em sua verdadeira classe como erros do tipo I, e as amostras classificadas em uma classe errada como erros do tipo II (**Tabela 14**). Como pode ser observado na **Tabela 14**, usando as mesmas condições, o BA-LDA demonstrou excelente desempenho classificando todas as amostras de teste corretamente, sendo superior ao SPA-LDA e ao SIMCA.

A partir das **Tabelas 12** e **14** foi possível verificar que na obtenção dos resultados desta Tese o SPA-LDA classificou todas amostras de teste corretamente enquanto no trabalho de Pontes (2009) ocorreram dois erros de classificação. Isso pode ser justificado pela divisão dos conjuntos de dados. Nesta Tese os dados foram divididos pelo KS com 60% das amostras para treinamento, 20% para validação e 20% para teste. Assim, foram usadas quantidades menores de amostras nos conjuntos de teste (25 amostras) quando comparado ao trabalho de Pontes (2009) (47 amostras). Possivelmente a classificação de todas as amostras externas, nesta Tese, pelo algoritmo SPA-LDA foi obtida devido a isso. Ou seja, as amostras que estavam sendo atribuídas incorretamente em Pontes (2009) foram usadas como amostras de treinamento do modelo nesta Tese. Quando observado os resultados do GA-LDA percebeu-se o mesmo desempenho de classificação que o obtido por Pontes (2009).

Ainda sobre a **Tabela 12**, também foi possível verificar que o algoritmo BA-LDA selecionou um subconjunto de variáveis com menor número de condição (187,4) que o GA-LDA (366,5). Assim, as variáveis selecionadas pelo BA-LDA resultaram em uma maior redução da multicolinearidade presente nos dados. Em suma, de acordo com os resultados observados (**Tabelas 12, 13** e **14**), o BA-LDA apresentou excelente desempenho, sendo promissor para aplicações envolvendo a classificação de óleos vegetais a partir de dados UV-Vis. A **Seção 5.3.6** apresenta um estudo da robustez realizado com os algoritmos estocásticos BA-LDA e GA-LDA.

Tabela 12: Resultados obtidos pelos diferentes métodos para a classificação de óleos de milho (C1), de soja (C2), de canola (C3) e de girassol (C4).

		BA-LDA (Série de teste)				**GA-LDA (Série de teste)**				**SPA-LDA (Série de teste)**				**PLS-DA (Série de teste)**				
		c1	***C2***	***c3***	***C4***	***c1***	***C2***	***c3***	***C4***	***c1***	***C2***	***c3***	***C4***	***c1***	***C2***	***c3***	***C4***	***Não atribuído***
Quatro classes	***c1***	**6**	0	0	0	**6**	0	0	0	**6**	0	0	0	**6**	0	0	0	0
	c2	0	**6**	0	0	0	**6**	0	0	0	**6**	0	0	1	**1**	0	0	4
	c3	0	0	**6**	0	0	0	**6**	0	0	0	**6**	0	0	0	**6**	0	0
	c4	0	0	0	**7**	0	0	0	**7**	0	0	0	**7**	0	0	0	**5**	**2**
Sensitividade (%)		100	100	100	100	100	100	100	100	100	100	100	100	100	50	100	100	
Seletividade (%)		100	100	100	100	100	100	100	100	100	100	100	100	92	100	100	100	
Número de variáveis selecionadas ou número ótimo de variáveis latentes		**7**				**8**				**7**				4				
TCC (%)		**100**				**100**				**100**				**95**				
Número de condição		**187.4**				**366.5**				***247.4***				**-----**				

C1-Milho; C2-Soja; C3- Canola; C4-Girassol.

Fonte: (própria).

Tabela 13: Resultados obtidos pelo método SIMCA para classificação de óleos de milho (C1), de soja (C2), de canola (C3) e de girassol (C4).

		SIMCA 1% (Série de teste)				**SIMCA 5% (Série de teste)**				**SIMCA 10% (Série de teste)**				**SIMCA 25% (Série de teste)**			
		c1	***C2***	***c3***	***C4***	***c1***	***c2***	***c3***	***c4***	***c1***	***c2***	***c3***	***c4***	***c1***	***c2***	***c3***	***c4***
Quatro classes	***c1***	6	6	0	0	6	5	0	0	6	5	0	0	6	1	0	0
	c2	1	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0
	c3	0	0	6	0	0	0	5	0	0	0	5	0	0	0	4	0
	c4	0	0	0	7	0	0	0	7	0	0	0	7	0	0	0	7
Nenhuma								1				1				2	

Fonte: (própria).

Tabela 14: Comparação do desempenho do BA-LDA com os resultados obtidos por Pontes (2009) para classficação de óleos vegetais.

	O Autor	Pontes (2009)	Pontes (2009)	Pontes (2009)	Pontes (2009)	Pontes (2009)	Pontes (2009)
	BA-LDA	SPA-LDA	GA-LDA	SIMCA 1%	SIMCA 5%	SIMCA 10%	SIMCA 25%
Tipo I	-	1	-	-	-	1	9
Tipo II	-	1	-	19	7	4	-
Total	-	2	-	19	7	5	9

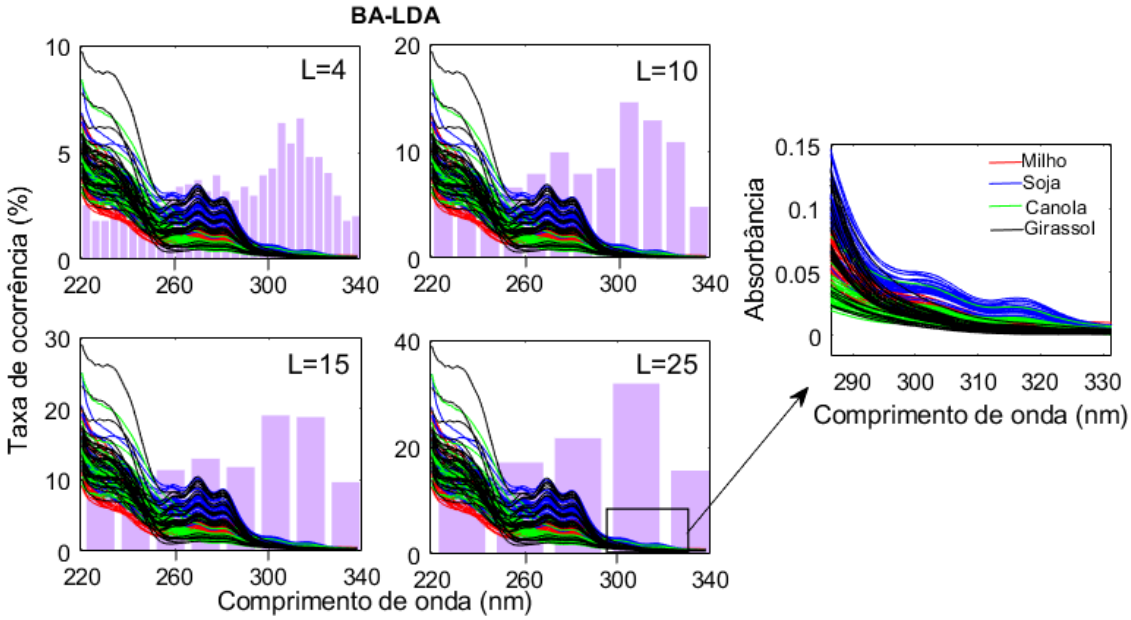
Fonte: adaptada (Pontes ,2009).

5.3.6 Avaliação da Robustez

5.3.6.1 BA-LDA

Para o conjunto de dados de amostras de óleos vegetais, a robustez do algoritmo proposto também foi avaliada. A **Figura 42** mostra os histogramas da taxa de ocorrência das variáveis selecionadas para cem repetições. Esses histogramas foram obtidos usando as larguras de intervalo (L) das variáveis iguais a 4, 10, 15 e 25.

Figura 42- Histogramas com diferentes larguras de faixas (4, 10, 15 e 25) para as variáveis selecionadas pelo BA-LDA em cem repetições do algoritmo.



Fonte: (própria).

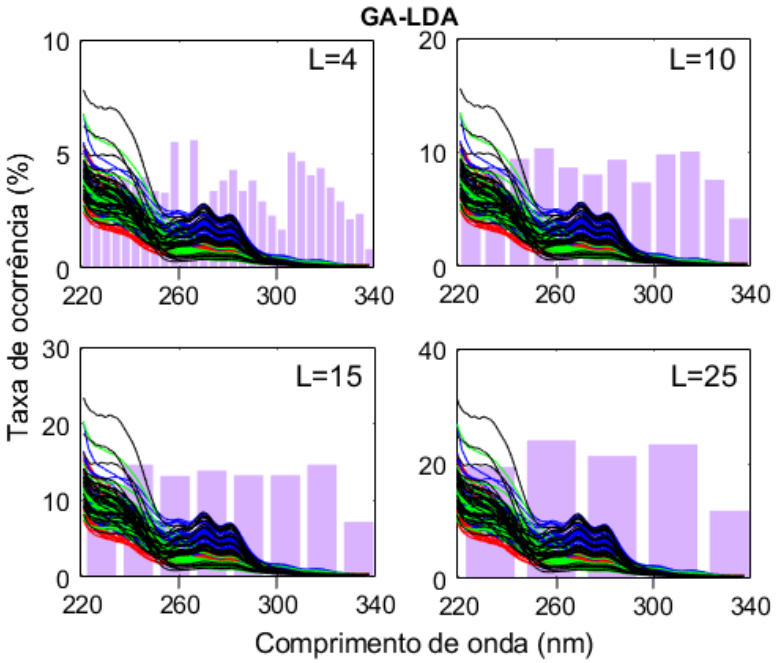
Na **Figura 42**, foi possível perceber que para as cem repetições, as variáveis selecionadas pelo BA-LDA convergiram na região aproximadamente entre 300 e 320 nm. Este comportamento pode ser observado independente da largura da faixa de variáveis utilizada, indicando a robustez do algoritmo. Na região de convergência das variáveis selecionadas pelo BA-LDA, os espectros apresentam a maior separação entre as classes, conforme observado no destaque da **Figura 42**. Ou seja, entre todas as variáveis do espectro (119 variáveis), o algoritmo baseado em morcegos virtuais, apesar de atuar de forma aleatória, foi direcionado para a região com as variáveis que levam à melhor separação das classes. Isso pode estar relacionado com a penalidade inserida no código do algoritmo para que os morcegos só atualizem suas posições se as variáveis selecionadas levarem ao menor erro de classificação. Assim, a cada atualização, os morcegos convergiam nas variáveis que reduzem o custo (G_{cost}). Com isso, pode-se inferir

que as variáveis selecionadas na região entre 300 e 320 nm levaram ao melhor desempenho de classificação. A **Seção 5.3.6.2** apresenta a avaliação da robustez do GA-LDA.

5.3.6.2 GA-LDA

A robustez do GA-LDA em realizar a seleção de variáveis para a classificação dos conjuntos de óleos vegetais também foi avaliada para fins de comparação. A **Figura 43** mostra os histogramas da taxa de ocorrência das variáveis selecionadas para cem repetições. Esses histogramas foram obtidos usando larguras de intervalo (L) das variáveis iguais a 4, 10, 15 e 25.

Figura 43- Histogramas com diferentes larguras de faixas (4, 10, 15 e 25) para as variáveis selecionadas pelo GA-LDA em cem repetições do algoritmo.



Fonte: (própria).

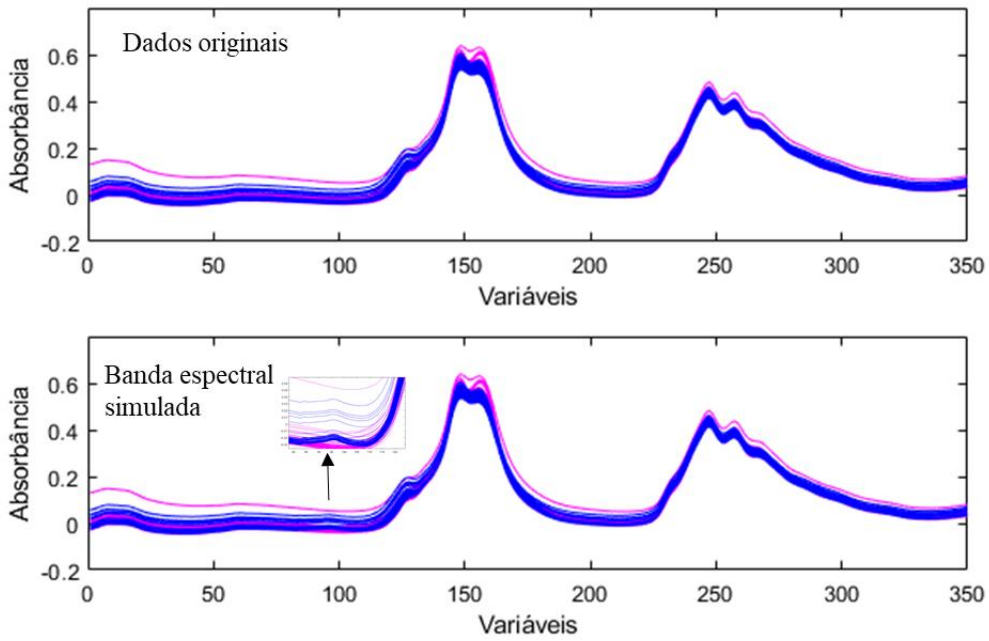
Na **Figura 43**, quando os resultados foram obtidos para largura de faixa com apenas 4 variáveis (L = 4), houve uma reprodutibilidade aparente especialmente próxima ao comprimento de onda de 260 nm (duas barras mais altas). Porém, quando a largura do intervalo das variáveis foi aumentada (L = 10, 15 e 25), o comportamento aleatório do algoritmo GA-LDA afetou a convergência das variáveis selecionadas, prejudicando sua robustez. Este algoritmo selecionou variáveis em todas as faixas agindo de forma aleatória, de acordo com sua natureza, sem nenhuma discriminação das variáveis mais significativas para o modelo de

classificação. Assim, o principal diferencial entre GA-LDA e BA-LDA foi a característica vantajosa do algoritmo dos morcegos na convergência para variáveis em regiões que favoreceram a discriminação de classes amostrais. A **Seção 5.4** apresenta um estudo de caso envolvendo dados simulados e dados reais de diesel.

5.4 Estudo de caso envolvendo dados simulados e dados espectrométricos NIR de diesel

Para este estudo de caso, dados espectrométricos reais de amostras de diesel foram modificados. Para isso, incialmente uma banda espectral simulada foi adicionada entre as variáveis 84 e 108 em cem das duzentas amostras de diesel. O propósito do estudo foi simular uma informação química que não estaria presente nas outras cem amostras do conjunto de dados. A utilização de dados reais teve como principal objetivo a manutenção de toda a variabilidade (flutuação e ruído) dos dados originais. Assim, pretendia-se verificar se o algoritmo conseguiria distinguir como pertencente a uma outra classe os espectros contendo a informação de baixa intensidade adicionada em uma pequena faixa espectral. A **Figura 44** mostra os espectros NIR das duzentas amostras de diesel antes e depois da adição dos dados simulados.

Figura 44- Espectros NIR reais e com informação simulada adicionada entre as variáveis 84 e 108.



Fonte: (própria).

Antes da avaliação do desempenho do algoritmo proposto foi realizado o estudo da otimização dos parâmetros para esses dados.

5.4.1 Otimização dos parâmetros do BA-LDA

Inicialmente, foi realizado um planejamento fatorial fracionado para otimizar os parâmetros do algoritmo BA-LDA. O intuito foi avaliar a significância dos fatores (α, γ, *mbats* e *N*) e suas respectivas interações. A **Tabela 15** mostra as respostas obtidas em termos de %TCC para cada um dos experimentos executados no planejamento fracionado 2^{4-1} com uma repetição.

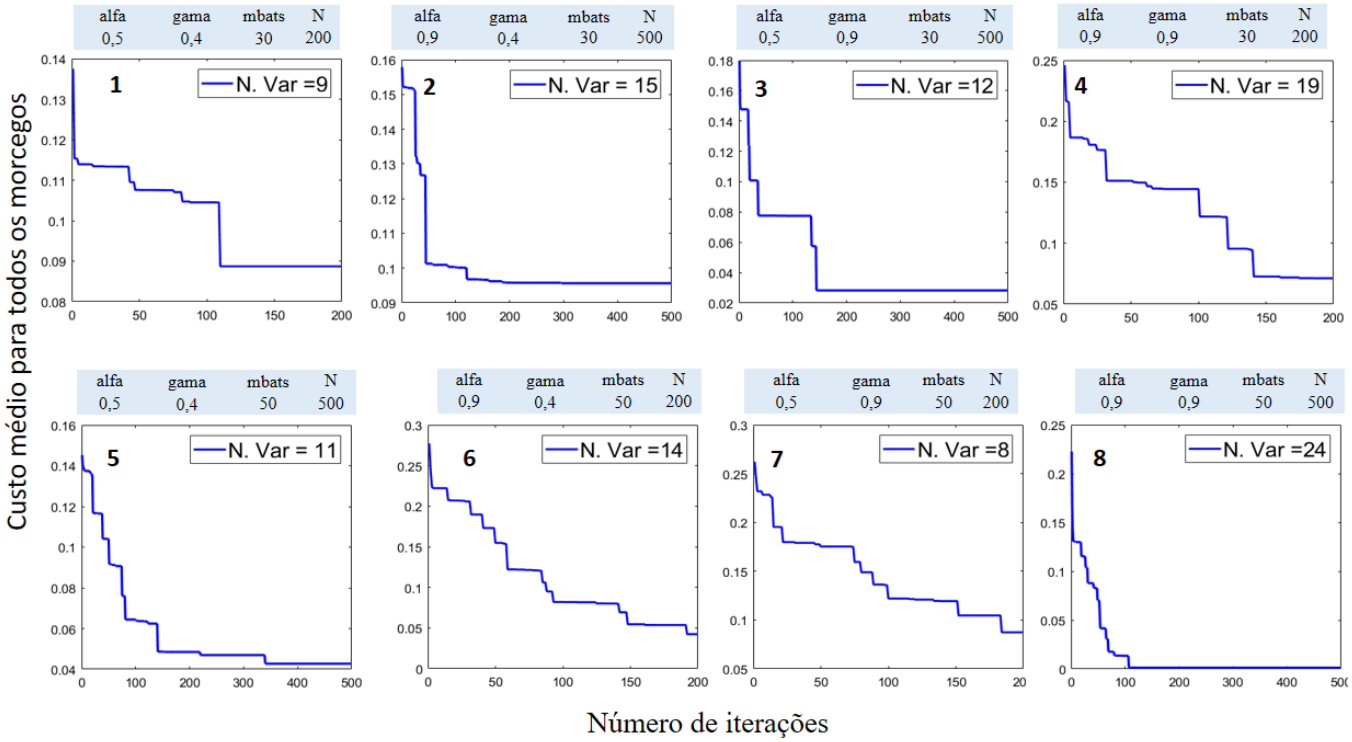
Tabela 15: Matriz de respostas no planejamento fatorial fracionado 2^{4-1}.

Replicadas	**α**	**γ**	***mbats***	***N***	***%TCC***
1	0,5	0,4	30	200	100
1	0,9	0,4	30	500	100
1	0,5	0,9	30	500	100
1	0,9	0,9	30	200	100
1	0,5	0,4	50	500	100
1	0,9	0,4	50	200	100
1	0,5	0,9	50	200	100
1	0,9	0,9	50	500	100
2	0,5	0,4	30	200	100
2	0,9	0,4	30	500	100
2	0,5	0,9	30	500	100
2	0,9	0,9	30	200	100
2	0,5	0,4	50	500	100
2	0,9	0,4	50	200	100
2	0,5	0,9	50	200	100
2	0,9	0,9	50	500	100

Fonte: (própria).

Como pode ser observado pela **Tabela 15**, independentemente do nível dos fatores avaliados a resposta sempre foi 100% de taxa de classificação correta. Dessa forma, os fatores, nos níveis estudados, e suas interações não teriam significância para obtenção da resposta. Assim, qualquer combinação de parâmetros avaliados nos experimentos poderia ser escolhida. Para escolha dos parâmetros decidiu-se então verificar os gráficos do custo médio para todos os morcegos ao longo das interações e o número de variáveis selecionadas (**Figura 45**).

Figura 45- Avaliação do custo médio para todos os morcegos com o emprego de diferentes combinações de fatores.



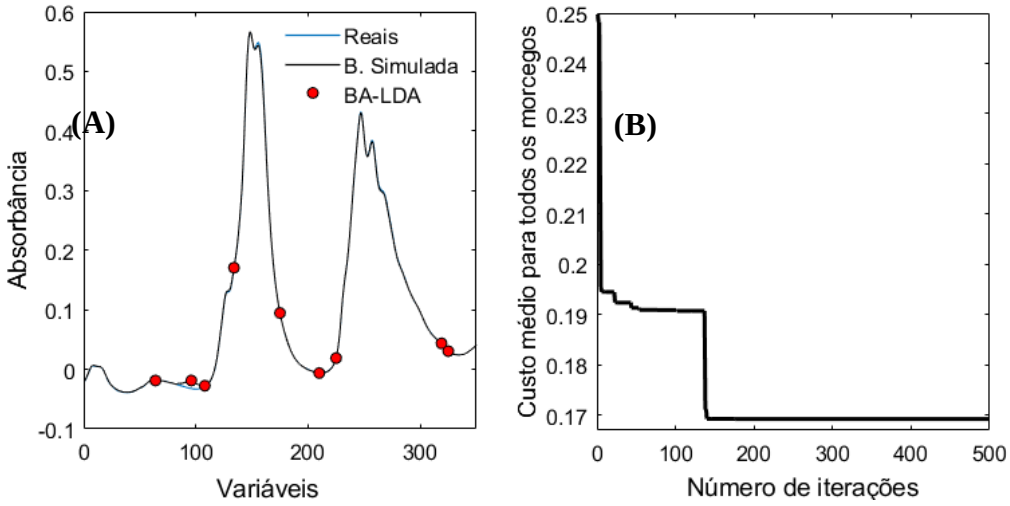
Fonte: (própria).

Como pode ser observado na **Figura 45**, na maioria dos casos em que se utilizou um número de iterações igual a 500, foi possível verificar a convergência do algoritmo, com exceção do quinto experimento. Para determinação dos parâmetros ideais selecionamos, então, o experimento realizado com 500 iterações que obteve o menor número de variáveis selecionadas. Dessa forma o terceiro experimento foi selecionado (α=0,5 γ =0,9 *mbats*=30 N=500). A **seção 5.4.2** apresenta o desempenho do BA-LDA aplicado aos dados.

5.4.2 Aplicação do BA-LDA

Para classificação dos dados com a informação simulada e os dados reais, o BA-LDA foi executado cinco vezes. Em todas as execuções o BA-LDA classificou corretamente todas as amostras. O melhor desempenho foi considerado em termos do menor número de variáveis selecionadas. A **Figura 46A** mostra as nove variáveis selecionadas pelo BA-LDA na segunda execução dos dados, para classificação das amostras. A **Figura 46B** mostra o gráfico do custo médio para todos os morcegos a cada iteração.

Figura 46- A) Espectros médios das classes e variáveis selecionadas pelo BA-LDA. B) Gráfico do custo médio para todos os morcegos ao longo das iterações.



Fonte: (própria).

Na **Figura 46A** foi possível perceber que duas variáveis foram selecionadas na região que porta a informação discriminante (região com a banda simulada - entre as variáveis 84 e 108). No estudo da robustez iremos verificar se o BA-LDA é realmente capaz de convergir para esta região que favorece a classificação das amostras. Como pode ser observado na **Figura 46B**, o custo médio para todos os morcegos na geração inicial é alto para as variáveis selecionadas. Quando os morcegos atualizam as posições pelas **Equações 34-36** esse custo reduz bastante logo nas primeiras iterações. No decorrer das próximas iterações, esse custo vai reduzindo e isso se justifica pelas atualizações (**Equações 34-36**), as quais conduzem a novas posições para os morcegos virtuais e melhores subconjuntos de variáveis selecionadas. Após executadas cerca de 150 iterações para esse conjunto de dados, é possível perceber que os morcegos convergem para uma única solução (**Figura 46B**). Isso significa que as variáveis selecionadas na referida iteração levaram ao menor custo e a partir daí não foi encontrado nenhum subconjunto de variáveis capaz de superar o desempenho destas. Assim, com esse melhor subconjunto de variáveis o modelo LDA foi obtido.

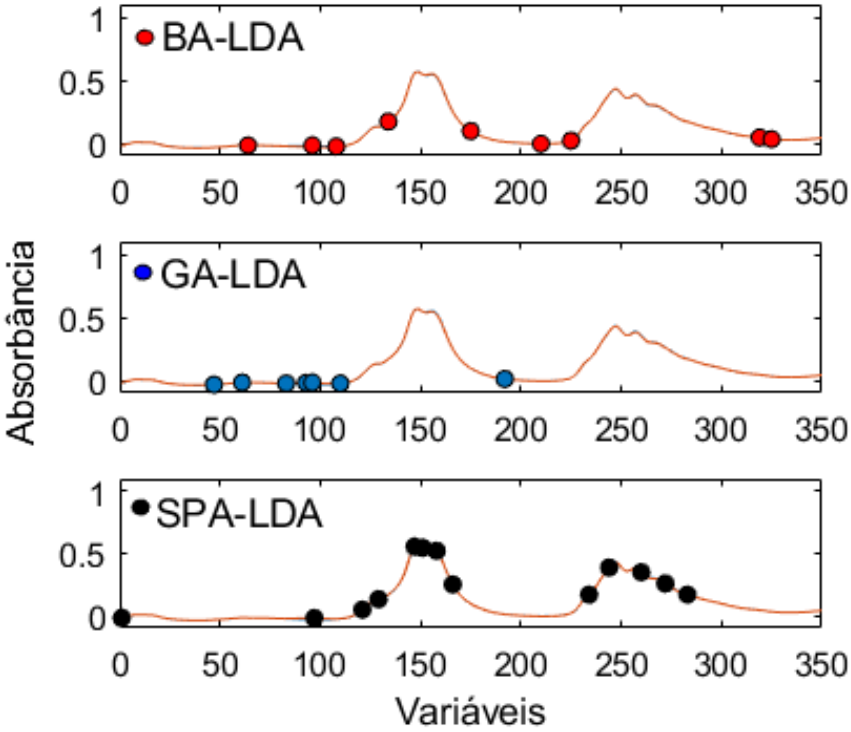
Com esse subconjunto de variáveis selecionadas o BA-LDA conseguiu distinguir com 100% de TCC, as amostras originais das amostras com a banda espectral simulada. Assim, pode-se inferir que o BA-LDA foi capaz de realizar a classificação de amostras com espectros muito semelhantes captando informações de baixa intensidade, em meio as flutuações presentes nos dados. As **Seções 5.4.3** e **5.4.4** comparam o desempenho do BA-LDA com outros métodos

de classificação, assim, as **Tabelas 17** e **18** resumem os resultados obtidos com os diferentes métodos.

5.4.3 Comparação do desempenho do BA-LDA com outros algoritmos de seleção de variáveis

A **Figura 47** mostra as variáveis selecionadas pelos algoritmos BA-LDA, GA-LDA e SPA-LDA. Como pode ser observado, todos os algoritmos selecionaram ao menos uma variável na região onde a banda espectral simulada foi adicionada aos dados, região responsável pela discriminação das classes.

Figura 47- Espectro médio das classes e variáveis selecionadas pelos algoritmos BA-LDA, GA-LDA e SPA-LDA.



Fonte: (própria).

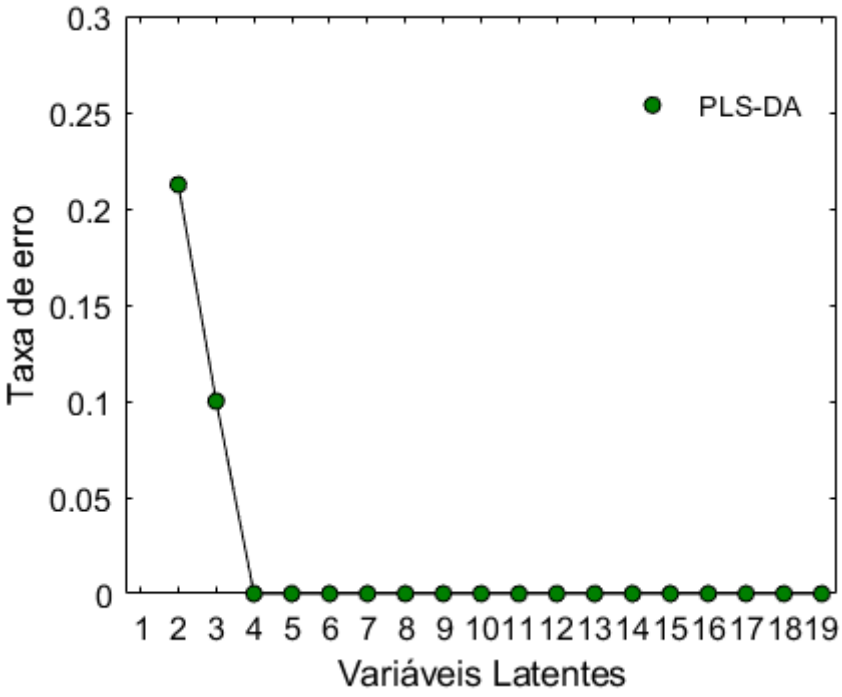
Ainda sobre a **Figura 47**, foi possível verificar que todos os algoritmos de seleção de variáveis foram parcimoniosos. Como desempenho de classificação os três algoritmos conseguiram classificar as amostras. A **Tabela 18**, resume os resultados obtidos pelos diferentes métodos de classificação empregados. Na **Seção 5.4.4** é apresentado o desempenho dos métodos PLS-DA e SIMCA.

5.4.4 Desempenho do PLS-DA e da classificação SIMCA

5.4.4.1 PLS-DA

Para utilização do algoritmo PLS-DA inicialmente calculou-se o número ótimo de variáveis latentes utilizando um conjunto de amostras de validação. A **Figura 48** mostra que com apenas quatro variáveis latentes foi possível obter a menor taxa de erro.

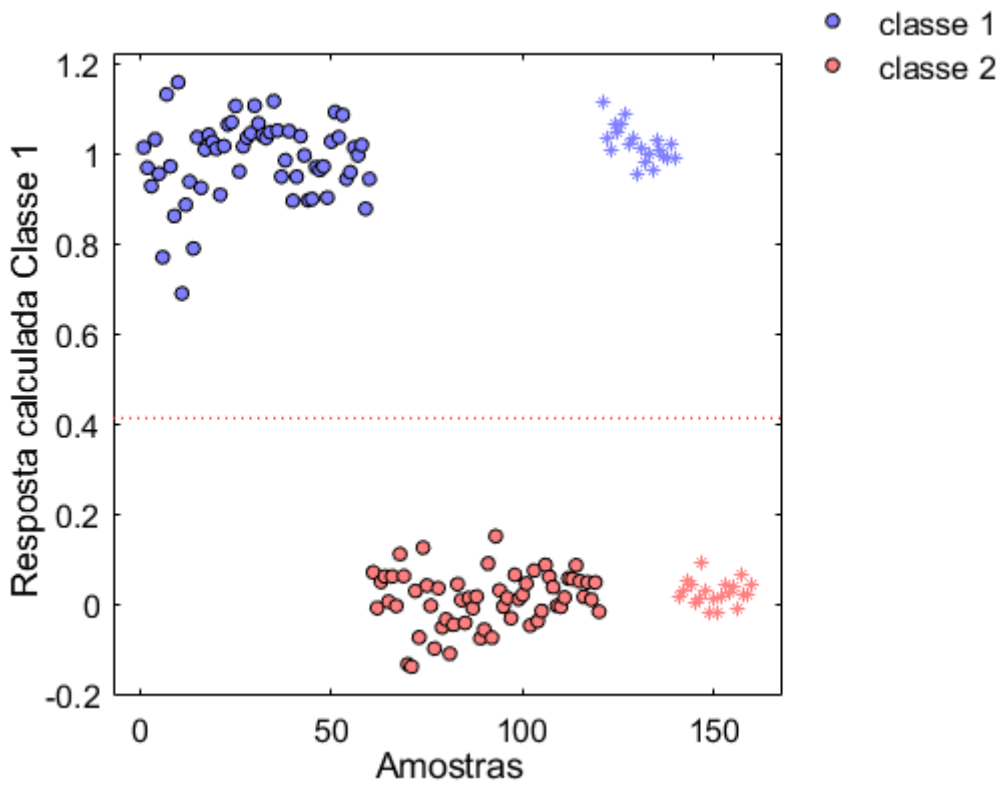
Figura 48- Número ótimo de variáveis latentes pelo PLS-DA.



Fonte: (própria).

Com o número ideal de variáveis latentes o modelo foi calibrado e posteriormente as amostras de teste foram preditas neste modelo. Com as quatro variáveis latentes (VL), o PLS-DA classificou as amostras com 100% de TCC, conforme apresentado na **Figura 49**.

Figura 49- Predição das amostras externas no modelo PLS-DA.



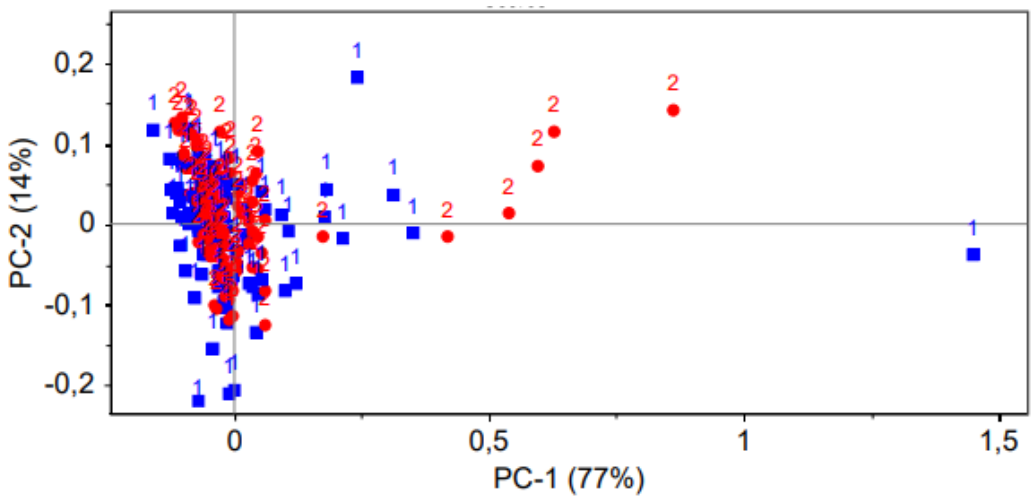
Fonte: (própria).

Na **Figura 49**, as amostras sem a informação simulada são representadas na cor azul e as amostras com a informação simulada são representadas em rosa. As amostras externas aos modelos (amostras de teste) são representadas como asteriscos e as amostras de treinamento e validação são representadas em círculos. Como pode ser observado, as amostras das diferentes classes estão bem distantes e isso também é refletido no excelente desempenho de classificação do modelo PLS-DA. A **Seção 5.4.4.2** apresenta o desempenho da classificação SIMCA para estes dados.

5.4.4.2 SIMCA

Para a classificação SIMCA, inicialmente foi realizada uma PCA geral para avaliar a distribuição das amostras. O número de PCs utilizado foi igual a dois e as duas PCs explicaram cerca de 91% da variabilidade dos dados. A **Figura 50** mostra os escores da PC1 x PC2 e nos eixos a variância explicada por cada PC.

Figura 50- Gráfico dos escores da PCA das amostras reais e com a informação simulada.



Fonte: (própria).

Como pode ser observado na **Figura 50**, as amostras das classe 1 (sem informação simulada) e 2 (amostras com informação simulada) estavam sobrepostas, dessa forma distinção entre elas não foi observada na PCA. Para classificação SIMCA foram construídos modelos de PCs individuais para cada classe, cada um com número de PCs igual a dois. A **Tabela 16** mostra a matriz de confusão para os quatro níveis de significância de *teste-F* avaliados.

Tabela 16: Resultados obtidos pelo método SIMCA modelos individuais das classes construídos com 2 PCs.

	SIMCA 1%		**SIMCA 5%**		**SIMCA 10%**		**SIMCA 25%**	
	C1	**C2**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
C1	**20**	**19**	**20**	**18**	**20**	**15**	**17**	**0**
C2	**20**	**20**	**19**	**20**	**19**	**20**	**2**	**16**
Nenhuma	---		---		---		**7**	

Fonte: (própria).

Como se pode observar na **Tabela 16**, nos diferentes níveis de significância do *teste-F*, amostras foram classificadas incorretamente. Com 1, 5 e 10% de significância quase todas as amostras foram atribuídas as duas classes estudadas. O melhor resultado da classificação SIMCA foi obtido com 25% de significância, conforme **Tabela 16**, onde duas amostras da

classe 2 foram atribuídas como pertencentes a classe 1, e 7 amostras (sendo três da classe 1 e quatro da classe 2) não foram atribuídas a nenhuma das classes.

Um estudo interessante foi a adição de ruído nestes dados para avaliação do desempenho dos métodos de classificação (**Seção 5.4.6**). Antes disso, a **Seção 5.4.5** compara o desempenho dos diferentes métodos de classificação nos dados sem adição do ruído.

5.4.5 Avaliação geral dos métodos empregados na classificação dos dados reais de diesel e dados com a informação simulada

A **Tabela 17** resume o desempenho de classificação de todos os métodos empregados, com exceção da classificação SIMCA apresentado na **Tabela 16**.

Tabela 17: Resultados obtidos pelos diferentes métodos para a classificação dos dados reais e dados com a informação simulada.

		BA-LDA (série de teste)		**GA-LDA (série de teste)**		**SPA-LDA (série de teste)**		**PLS-DA (série de teste)**	
		C1	**C2**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
Duas classes	**C1**	**20**	0	**20**	0	**20**	0	**20**	0
	C2	0	**20**	0	**20**	0	**20**	0	**20**
Nenhuma		---		---		---		---	
Sensibilidade (%)		100		100		100		100	
Seletividade (%)		100		100		100		100	
Número de variáveis selecionadas ou número ótimo de variáveis latentes		**9**		**7**		**13**		**4**	
TCC (%)		**100**		**100**		**100**		**100**	
Número de condição		**1.0431e+03**		**1.0815e+03**		**1.2224e+03**		---	

Fonte: (própria).

Analisando as **Tabelas 16** e **17**, foi possível verificar que apenas na modelagem SIMCA amostras foram classificadas incorretamente. Assim, apesar da região discriminante simulada possuir pequena intensidade os demais métodos avaliados conseguiram classificar as amostras. Em um experimento com dados reais, dentre os métodos empregados, os algoritmos de seleção de variáveis seriam preferíveis pois operam no domínio original dos dados e favorecem a interpretação química direta dos espectros. Assim, caso a banda simulada correspondesse a um estiramento de algum grupo funcional, a seleção de variáveis nessa região seria de grande importância. Dessa forma o BA-LDA apresentou desempenho semelhante aos métodos tradicionalmente empregados (GA-LDA, PLS-DA), ao algoritmo determinístico SPA-LDA e superior ao SIMCA em todos os níveis de significância do *teste-F*. A **Seção 5.4.6** apresenta um estudo para avaliar a sensibilidade, dos métodos de classificação, ao ruído adicionado nos dados. Na **Seção 5.4.7** um estudo da robustez do BA-LDA e GA-LDA para os dados é realizado.

5.4.6 Estudo de sensibilidade ao ruído

Nesse estudo, foi adicionado ruído aos espectros das amostras de teste nos dados de diesel com e sem a informação simulada. As amostras externas com a adição do ruído foram aplicadas aos modelos de classificação construídos. O ruído adicional foi avaliado com desvios padrão de 0,001 e 0,005. Quando aplicado com desvio padrão do ruído adicional de 0,001 todos os algoritmos classificaram todas as amostras corretamente (**Tabela 18**). Quando se elevou o nível de ruído adicional (**Tabela 19**) o BA-LDA apresentou o mesmo desempenho que o GA-LDA e ambos foram superiores ao SPA-LDA. O melhor resultado de sensibilidade ao ruído foi obtido com o PLS-DA. As **Tabelas 20** e **21** apresentam o desempenho da classificação SIMCA para quatro níveis de significância do Teste F com ruído adicional com desvio de padrão de 0,001 (**Tabela 20**) e 0,005 (**Tabela 21**).

Tabela 18: Resultados obtidos para a classificação dos dados reais e dados com a informação simulada com adição de ruído com desvio padrão 0,001.

		BA-LDA		GA-LDA		SPA-LDA		PLS-DA	
		C1	**C2**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
Duas classes	**C1**	**20**	0	**20**	0	**20**	0	**20**	0
	C2	0	**20**	0	**20**	0	**20**	0	**20**
Nenhuma		---		---		---		---	
Sensibilidade (%)		100		100		100		100	
Seletividade (%)		100		100		100		100	
Número de variáveis selecionadas ou número ótimo de variáveis latentes		**10**		**6**		**13**		**4**	
TCC (%)		**100**		**100**		**100**		**100**	

Fonte (própria).

Tabela 19: Resultados obtidos para a classificação dos dados reais e dados com a informação simulada com adição de ruído com desvio padrão 0,005.

		BA-LDA		GA-LDA		SPA-LDA		PLS-DA	
		C1	**C2**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
Duas classes	**C1**	**20**	1	**19**	1	**18**	2	**20**	0
	C2	2	**20**	2	**18**	4	**16**	0	**20**
Nenhuma		---		---		---		---	
Sensibilidade (%)		95		95		88,9		100	
Seletividade (%)		90		90		81,8		100	
Número de variáveis selecionadas ou número ótimo de variáveis latentes		**16**		**15**		**13**		**4**	
TCC (%)		**92,5%**		**92,5%**		**85%**		**100**	

Tabela 20: Resultados obtidos pelo método SIMCA para classificação de dados reais e com informação simulada com adição de ruído com desvio padrão 0,001.

		SIMCA 1%		**SIMCA 5%**		**SIMCA 10%**		**SIMCA 25%**	
		C1	**C2**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
Duas classes	**C1**	20	19	20	11	15	0	0	0
	C2	20	20	16	18	3	13	0	0
	Nenhuma				1	5	7	20	20

Fonte: (própria).

Tabela 21: Resultados obtidos pelo método SIMCA para classificação de dados reais e com informação simulada com adição de ruído com desvio padrão 0,005.

		SIMCA 1%		**SIMCA 5%**		**SIMCA 10%**		**SIMCA 25%**	
		C1	**C2**	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
Duas classes	**C1**	0	0	0	0	0	0	0	0
	C2	0	0	0	0	0	0	0	0
	Nenhuma	20	20	20	20	20	20	20	20

Fonte: (própria).

A partir da **Tabela 19** foi possível verificar que o desempenho de previsão dos modelos de seleção de variáveis (BA, GA, SPA acoplados a LDA) e o PLS-DA não foram afetados pelo ruído adicionado com desvio padrão de 0,001. Em contrapartida, na **Tabela 20**, para os quatro níveis de significância do *teste-F* a classificação SIMCA atribuiu amostras incorretamente. Para os nível de significância de 25% as amostras com adição de ruído não foram atribuídas a nenhuma das classes. A informação de baixa intensidade adicionada aos dados (que simulou uma nova classe) provavelmente estava no nível do ruído dos dados. Quando mais ruído foi adicionado as amostras externas, os modelos SIMCA não conseguiram classifica-las. Isso indica que a variância residual das amostras externas era superior a todas as variâncias residuais das amostras de treinamento delimitadas nos modelos das classes. Quando o desvio padrão do ruído adicional foi de 0,005 (**Tabela 21**) em nenhum dos níveis de significância do Teste F avaliados o SIMCA conseguiu classificar as amostras externas. Assim, para dados ruidosos e com informações discriminantes de baixa intensidade, os modelos SIMCA podem não ser a melhor opção.

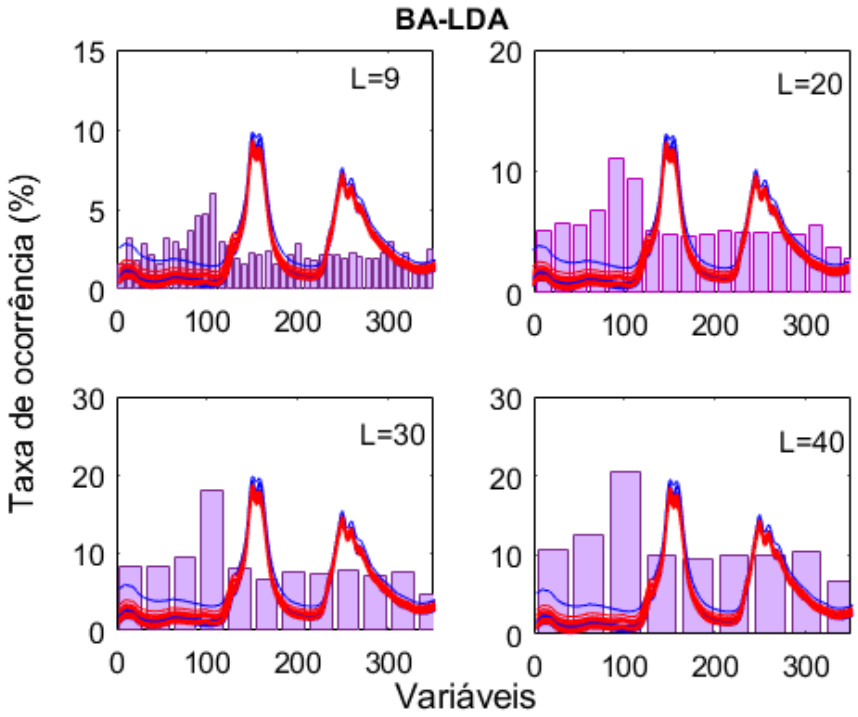
O desempenho superior dos algoritmos de seleção de variáveis pode ter ocorrido devido a seleção de variáveis na região discriminatória das classes. Assim, além de reduzir a dimensionalidade dos dados esses algoritmos demostraram excelentes resultados de classificação. Quando o desvio padrão do ruído adicional foi de 0,005 o BA-LDA foi semelhante ao GA-LDA e superior ao SPA-LDA e ao SIMCA. Assim, o algoritmo proposto (BA-LDA) demostrou potencialidade para ser aplicado a dados ruidosos e com informações discriminantes de baixa intensidade. A **Seção 5.4.7** apresenta um estudo da robustez realizado com os algoritmos estocásticos BA-LDA e GA-LDA.

5.4.7 Avaliação da Robustez

5.4.7.1 BA-LDA

O estudo da robustez objetivou avaliar uma característica observada logo que o BA-LDA foi implementado e testado. Nos primeiros testes com o código, observou-se que a cada execução, variáveis eram repetidamente selecionadas em regiões específicas dos espectros. Para esse estudo, o algoritmo foi executado cem vezes e as variáveis selecionadas em cada execução foram armazenadas e posteriormente histogramas de frequências foram construídos com essas variáveis. A **Figura 51** apresenta os histogramas com diferentes larguras de faixas.

Figura 51- Histogramas com diferentes larguras de faixas (9, 20, 30 e 40) para as variáveis selecionadas pelo BA-LDA em cem repetições do algoritmo.



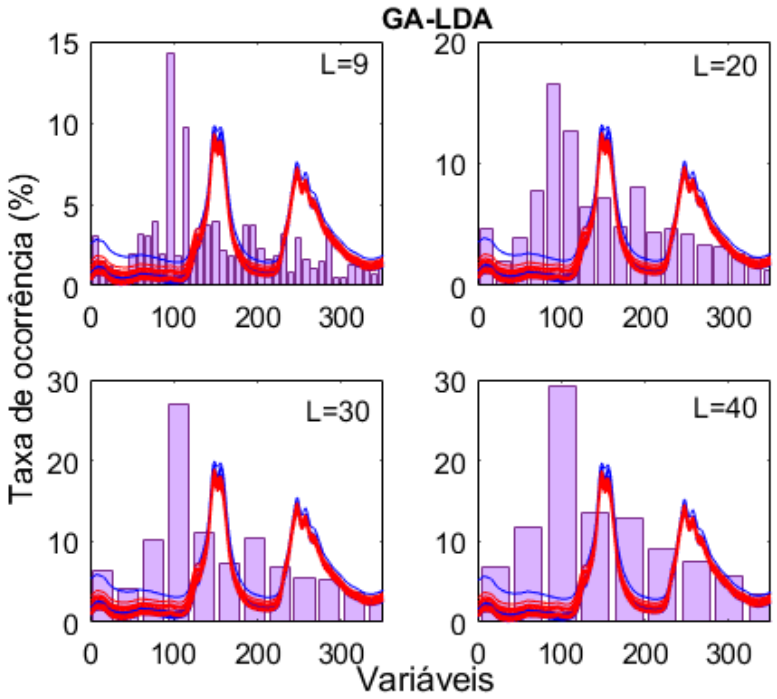
Fonte: (própria).

A partir da **Figura 51** foi possível verificar a convergência para a seleção de variáveis na região discriminante dos dados. Ou seja, essa foi mais uma aplicação que comprovou a robustez do BA-LDA. A **Seção 5.4.7.2** apresenta um estudo da robustez com o algoritmo GA-LDA.

5.4.7.2 GA-LDA

O GA-LDA foi empregado como método estocástico comparativo, assim, o estudo da robustez também realizado com este algoritmo. A **Figura 52** mostra os histogramas construídos com diferentes larguras de faixas para as variáveis selecionadas pelo GA-LDA em cem execuções do código do algoritmo.

Figura 52- Histogramas com diferentes larguras de faixas (9, 20, 30 e 40) para as variáveis selecionadas pelo GA-LDA em cem repetições do algoritmo.



Fonte: (própria).

Na **Figura 52** foi possível verificar que para estes dados o GA-LDA assim como o BA-LDA, também demostrou robustez convergindo para a região com a informação discriminante nos espectros dos dados. Provavelmente, a convergência se deu devido a presença da informação simulada em uma região específica, que apesar de possuir pequena intensidade favoreceu a obtenção de melhores resultados para os algoritmos de seleção de variáveis. Outra característica observada nos estudos desta Tese, foi que o GA-LDA apresentava os melhores desempenhos quando o número de iterações era elevado. Assim, para outras aplicações aumentar o número de iterações do GA-LDA pode favorecer um melhor desempenho.

Comparando ao GA-LDA, a principal vantagem observada para o BA-LDA foi a convergência para regiões específicas (talvez regiões discriminantes das classes) independente do conjunto de dados e/ou técnica ao qual foi aplicado. Quando comparado ao PLS-DA, o principal diferencial do BA-LDA está focado na interpretação química diretamente dos espectros, pois usa o domínio original dos dados. E quando comprado ao SPA-LDA o desempenho do BA-LDA foi semelhante. Assim, o BA-LDA pode ser superior a outros algoritmos estocásticos em termos de robustez e pode ser capaz de gerar resultados semelhantes à métodos não probabilísticos como o SPA-LDA e o PLS-DA.

CONCLUSÕES E PERSPECTIVAS

Capítulo 6

6 CONCLUSÕES E PERSPECTIVAS

Neste trabalho, um novo algoritmo inspirado em morcegos foi proposto para selecionar variáveis para classificação multivariada via LDA. Denominado de BA-LDA, o algoritmo foi avaliado usando quatro conjuntos de dados multivariados envolvendo espectros de massas MS, espectros UV-Vis e espectros NIR reais e com informação simulada. Seu desempenho foi comparado ao do algoritmo genético (GA-LDA), ao algoritmo das projeções sucessivas (SPA-LDA), a análise discriminante de mínimos quadrados parciais (PLS-DA) e a classificação SIMCA.

Como principais vantagens, o BA-LDA demonstrou desempenho satisfatório de classificação, sendo comparável ao algoritmo estocástico GA-LDA, ao algoritmo determinístico SPA-LDA e ao PLS-DA. Quando comparado a classificação SIMCA, o BA-LDA apresentou desempenho superior para os conjuntos de dados avaliados. Quanto ao estudo da sensibilidade ao ruído, no quarto conjunto de dados avaliado, o BA-LDA foi menos sensível quando comparado ao SIMCA.

Ainda sobre as vantagens, em todos os conjuntos de dados avaliados o BA-LDA foi parcimonioso. Quando avaliada a robustez, o BA-LDA foi superior ao GA-LDA em dois dos quatro conjuntos de dados estudados. Ou seja, independe dos dados analisados o BA-LDA convergiu em diferentes repetições para subconjuntos de variáveis em regiões favoráveis ao melhor desempenho de classificação.

Por fim, o algoritmo BA-LDA proposto também apresenta vantagens em relação a outros estocásticos, pois possui um mecanismo de busca local que permite a busca por melhores soluções quando o morcego virtual se depara com uma solução indesejável. Além disso, como funciona no domínio dos dados originais, permite uma interpretação química direta dos espectros. Considerando que o desempenho de classificação do algoritmo proposto é tão bom quanto o dos algoritmos tradicionais estudados, o BA-LDA se destaca como um novo algoritmo de classificação com potencial para operar com robustez em dados complexos como dados de espectros de massas, NIR e UV-VIS.

Como perspectivas o algoritmo BA-LDA pode ser aplicado a outros conjuntos de dados envolvendo outras técnicas, a exemplo, imagens hiperespectrais. Também sugere-se a adaptação do algoritmo com uma função quadrática (BA-QDA). Além disso, pretende-se avaliar o algoritmo em dados de 2° ordem. Assim, espera-se que o algoritmo BA-LDA seja utilizado por muitos pesquisadores auxiliando nas metodologias de classificação multivariada.

REFERÊNCIAS

AMJAD, A. *et al.* Raman spectroscopy based analysis of milk using random forest classification. **Vibrational Spectroscopy**, 2018. v. 99, p. 124–129. Disponível em: <https://doi.org/10.1016/j.vibspec.2018.09.003>.

ANDERSEN, C. M.; BRO, R. Variable selection in regression — a tutorial. **J. Chemometrics**, 2010. v. 24, p. 728–737.

ARAÚJO, T. K. L. *et al.* Non-destructive authentication of Gourmet ground roasted coffees using NIR spectroscopy and digital images. **Food Chemistry**, 2021. v. 364, p. 130452.

ATTIA, K. A. M. *et al.* Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: A comparative study. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, 2016. v. 170, p. 117–123. Disponível em: <http://dx.doi.org/10.1016/j.saa.2016.07.016>.

AZCARATE, Silvana Mariela *et al.* Classification of argentinean sauvignon blanc wines by UV spectroscopy and chemometric methods. **Journal of Food Science**, 2013. v. 78, n. 3, p. 432–436.

BAQUETA, M. R. *et al.* Multivariate classification for the direct determination of cup profile in coffee blends via handheld near-infrared spectroscopy. **Talanta**, 2021. v. 222, p. 121526. Disponível em: <https://doi.org/10.1016/j.talanta.2020.121526>.

BARBOSA, T. M. *et al.* A novel use of infra-red spectroscopy (NIRS and ATR-FTIR) coupled with variable selection algorithms for the identification of insect species (Diptera: Sarcophagidae) of medico-legal relevance. **Acta Tropica**, 2018. p. 1–12. Disponível em: <https://doi.org/10.1016/j.actatropica.2018.04.025>.

BARDIN, F. D. *et al.* Application of infrared spectral techniques on quality and compositional attributes of coffee : An overview. **Food Research International**, 2014. v. 61, p. 23–32.

BONIFAZI, G. *et al.* Contaminant detection in pistachio nuts by different classification methods applied to short-wave infrared hyperspectral images. **Food Control**, 2021. v. 130, p. 108202. Disponível em: <https://doi.org/10.1016/j.foodcont.2021.108202>.

CAI, T.; LIU, W. A direct estimation approach to sparse linear discriminant analysis. **Journal**

of the American Statistical Association, 2011. v. 106, n. 496, p. 1566–1577.

CEBI, N. *et al.* A rapid ATR-FTIR spectroscopic method for classification of gelatin gummy candies in relation to the gelatin source. **Food Chemistry**, 2019. v. 277, n. June 2018, p. 373–381.

CENTNER, V. *et al.* Elimination of Uninformative Variables for Multivariate Calibration. **Analytical Chemistry**, 1996. v. 68, n. 21, p. 3851–3858.

CERQUEIRA, E. O. *et al.* Utilização de filtro de transformada de fourier para a minimização de ruídos em sinais analíticos. **Química Nova**, 2000. v. 23, n. 5, p. 690–698.

CHAUHAN, R. *et al.* On the discrimination of soil samples by derivative diffuse reflectance UV – vis-NIR spectroscopy and chemometric methods. **Forensic Science International journal**, 2021. v. 319.

CHEN, H.; TAN, C.; LIN, Z. Identification of ginseng according to geographical origin by near-infrared spectroscopy and pattern recognition. **Vibrational Spectroscopy**, 2020. v. 110, p. 103149. Disponível em: <https://doi.org/10.1016/j.vibspec.2020.103149>.

CHENG, J.; SUN, D.; PU, H. Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen – thawed fish muscle. **Food Chemistry**, 2016. v. 197, p. 855–863.

CHENG, R.; JIN, Y. A social learning particle swarm optimization algorithm for scalable optimization. **Information Sciences**, 2015. v. 291, p. 43–60.

COIMBRA, L. J. M. *et al.* Conseqüências da multicolinearidade sobre a análise de trilha em canola. **Ciência Rural**, 2005. v. 35, n. 2, p. 347–352.

CONRADS, T. P. *et al.* High-resolution serum proteomic features for ovarian cancer detection. **Endocrine-Related Cancer**, 2004. v. 11, p. 163–178.

CORMEN, T. **Desmistificando algoritmos**. [S.l.]: Elsevier Brasil, 2017.

DORIGO, M.; STÜTZLE, T. **Optimization**. [S.l.]: A Bradford Book, 2004.

FARID, S. *et al.* Exploring ATR Fourier transform IR spectroscopy with chemometric analysis and laser scanning microscopy in the investigation of forensic documents fraud. **Optics and Laser Technology**, 2021. v. 135.

FERNANDES, D. *et al.* Application of infrared spectral techniques on quality and compositional attributes of coffee : An overview. **FRIN**, 2014. p. 1–10.

FERREIRA, L. S. *et al.* Caracterização de óleos e resinas vegetais da Amazônia por espectroscopia de absorção. **Scientia Plena**, 2017. v. 13, p. 1–7.

FERREIRA, Márcia M. C. *et al.* Quimiometria I: Calibração Multivariada, Um tutorial. **Química Nova**, 1999. v. 22, n. 5, p. 724–731. Disponível em: <http://www.scielo.br/scielo.php?pid=S0100-40421999000500016&script=sci_arttext>.

FERREIRA, Márcia M. C. Análise exploratória de dados. 2008. Disponível em: <https://lqta.iqm.unicamp.br/portugues/downloads/ANAL EXPL2008.pdf>. Acesso em: 23 nov. 2021.

FERREIRA, Márcia M. C. **Quimiometria - Conceitos, Métodos e Aplicações**. Campinas-SP: Editora da Unicamp, 2015.

FILHO, A. C.; POPPI, R.J. Algoritmo genético em química. **Quimica Nova**, 1999. v. 22, n. 3, p. 405–411.

GALVÃO, R. K. H.; ARAÚJO, M. C. U. Variable Selection. *In*: TAULER, R.; WALCZAK, B.; BROWN, S. (Org.). **Comprehensive chemometrics: chemical and biochemical data analysis.** 1 st ed. [S.l.]: Elsevier, 2009, p. 233–283.

GOMES, A. A. *et al.* Variable selection in the chemometric treatment of food data: A tutorial review. **Food Chemistry**, 2021. v. 370, p. 131072. Disponível em: <https://doi.org/10.1016/j.foodchem.2021.131072>.

GU, S.; CHENG, R.; JIN, Y. Feature selection for high-dimensional classification using a competitive swarm optimizer. **Soft Computing**, 2016.

HAASE, E.; ARROYO, L.; TREJOS, T. Classification of printing inks in pharmaceutucal packages by Laser-Induced Breakdown Spectroscopy and Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy. **Spectrochimica Acta - Part B Atomic Spectroscopy**, 2020. v. 172.

HAIR, J. F. *et al.* **Análise multivariada de dados**. São Paulo: Bookman editora, 2009.

HU, L. *et al.* Vis-NIR spectroscopy Combined with Wavelengths Selection by PSO Optimization Algorithm for Simultaneous Determination of Four Quality Parameters and

Classification of Soy Sauce. **Food Analytical Methods**, 2019. v. 12, p. 633–643.

JAKOBSEN, L.; BRINKLØV, S.; SURLYKKE, A. Intensity and directionality of bat echolocation signals. **Frontiers in Physiology**, 2013. v. 4, p. 1–10.

JAMWAL, R. *et al.* Rapid detection of pure coconut oil adulteration with fried coconut oil using ATR-FTIR spectroscopy coupled with multivariate regression modelling. **LWT - Food Science and Technology**, 2020. v. 125, n. February.

JAMWAL, R. *et al.* Recent trends in the use of FTIR spectroscopy integrated with chemometrics for the detection of edible oil adulteration. **Vibrational Spectroscopy**, 2021. v. 113.

JANNAT, B. *et al.* Gelatin speciation using real-time PCR and analysis of mass spectrometry-based proteomics datasets. **Food Control**, 2018. v. 87, p. 79–87.

JOLAYEMI, O. S.; AJATTA, M. A.; ADEGEYE, A. A. Geographical discrimination of palm oils (Elaeis guineensis) using quality characteristics and UV- - visible spectroscopy. **Food Science & Nutrition**, 2018. v. 6, p. 773–782.

KARUNATHILAKA, S. R. *et al.* Rapid classification and quantification of marine oil omega-3 supplements using ATR-FTIR , **FT-NIR and chemometrics**. 2019. v. 77, n. December 2018, p. 9–19.

KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometrics**, 1969. v. 11, n. 1, p. 137–148.

KENNEDY, J.; EBERHART, R. Particle Swarm Optimisation. [S.l.]: [s.n.], 1995. v. 4, p. 1942–1948.

KHANMOHAMMADI, M.; GARMARUDI, A. B.; GUARDIA, M. De La. Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear discriminant analysis. **Talanta**, 2013. v. 104, p. 128–134. Disponível em: <http://dx.doi.org/10.1016/j.talanta.2012.11.032>.

KONZEN, P. H. De A. *et al.* Otimização de métodos de controle de qualidade de fármacos usando algoritmo genético e busca de tabu. **Pesquisa Operacional**, 2003. v. 23, n. 1, p. 189–207.

LEARDI, R. Genetic Algorithm. *In*: WALCZAK, B.; FERRÉ, R. T.; BROWN, S. (Org.).

Comprehensive chemometrics:chemical and biochemical data analysis. 1º ed. Amsterdã: Elsevier, 2009, p. 631–653.

LIN, S.; CHEN, S. PSOLDA : A Particle Swarm Optimization Approach for Enhancing Classification Accuracy Rate of Linear Discriminant Analysis. **Applied Soft Computing**, 2009. v. 9, p. 1008–1015.

LINDEN, R. **Algoritmos genéticos**. 2º edição ed. Rio de Janeiro: Brasport livros e multimídia Ltda, 2008.

LIU, Q. *et al.* A novel hybrid bat algorithm for solving continuous optimization problems. **Applied Soft Computing Journal**, 2018. v. 73, p. 67–82.

LOHUMI, S. *et al.* A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. **Trends in Food Science & Technology**, 2015. v. 46, n. 1, p. 85–98.

LÓPEZ-MAESTRESALAS, A. *et al.* Detection of minced lamb and beef fraud using NIR spectroscopy. **Food Control**, 2018.

MAJDA, A. *et al.* Hyperspectral imaging and multivariate analysis in the dried blood spots investigations. **Applied Physics A**, 2018. v. 124, p. 1–8.

MEHMOOD, T. *et al.* A Partial Least Squares based algorithm for parsimonious variable selection. **Algorithms for molecular biology**, 2011. v. 6, p. 1–12.

MESSAOUDI, I.; KAMEL, N. A multi-objective bat algorithm for community detection on dynamic social networks. **Applied Intelligence**, 2019.

MIRJALILI, S.; MOHAMMAD, S. Binary bat algorithm. **Neural Comput & Applic**, 2014. v. 25, p. 663–681.

NADERI, M.; KHAMEHCHI, E.; KARIMI, B. Novel statistical forecasting models for crude oil price, gas price, and interest rate based on meta-heuristic bat algorithm. **Journal of Petroleum Science and Engineering**, 2019. v. 172, p. 13–22. Disponível em: <https://doi.org/10.1016/j.petrol.2018.09.031>.

NEMA, S.; THAKUR, S. . Improved Particle Swarm Optimization approach for Classification by using LDA. **IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO)**, 2015. n. 2.

NIU, T. *et al.* Multi-step-ahead wind speed forecasting based on optimal feature selection and a modified bat algorithm with the cognition strategy. **Renewable Energy**, 2018. v. 118, p. 213–229.

OKUBO, N.; KURATA, Y. Nondestructive classification analysis of green coffee beans by using near-infrared spectroscopy. **Foods**, 2019. v. 82, n. 8, p. 1–7.

PANCHUK, V. *et al.* Analytica Chimica Acta Application of chemometric methods to XRF-data e A tutorial review. **Analytica Chimica Acta**, 2018. v. 1040, p. 19–32.

PAOLETTI, M. E. *et al.* A new deep convolutional neural network for fast hyperspectral image classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, 2018. v. 145, p. 120–147. Disponível em: <https://doi.org/10.1016/j.isprsjprs.2017.11.021>.

PAULA, L. C. M. *et al.* Modern metaheuristic with multi-objective formulation for the variable selection problem. **Journal of Computer Science**, 2017. v. 13, n. 11, p. 659–666.

PAULA, L. C. M. *et al.* Epistasis-based FSA : Two versions of a novel approach for variable selection in multivariate calibration. **Engineering Applications of Artificial Intelligence**, 2019. v. 81, p. 213–222.

PIERNA, J. A. F. *et al.* A Backward Variable Selection method for PLS regression (BVSPLS). **Analytica Chimica Acta journal**, 2009. v. 642, p. 89–93.

PONTES, A. S. *et al.* Ant colony optimization for variable selection in discriminant linear analysis. **Journal of Chemometrics**, 2020. p. 1–12.

PONTES, C. *et al.* The successive projections algorithm for spectral variable selection in classification problems. **Chemometrics and Intelligent Laboratory Systems**, 2005. v. 78, p. 11–18.

PONTES, M. J. C. **Algoritmo das Projeções Sucessivas para Seleção de Variáveis Espectrais em Problemas de Classificação**. [S.l.]: Universidade Federal da Paraíba, 2009.

RIBEIRO, J. S.; FERREIRA, M. M. C; SALVA, T. J. Chemometric models for the quantitative sensory analysis of Arabica coffee beverages using near infrared spectroscopy. **Talanta**, 2011. v. 83, p. 1352–1358.

RIBEIRO, L. A. *et al.* Multi-objective genetic algorithm for variable selection in multivariate classification problems: A case study in verification of biodiesel adulteration. **Procedia**

Computer Science, 2015. v. 51, p. 346–355.

RINNAN, Å. *et al.* Recursive weighted partial least squares (rPLS): An efficient variable selection method using PLS. **Journal of Chemometrics**, 2013. v. 28, p. 439–447.

RINNAN, Å.; BERG, F. Van Den; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **Trends in Analytical Chemistry**, 2009. v. 28, n. 10, p. 1201–1222.

SAFO, S. E.; AHN, J. General sparse multi-class linear discriminant analysis. **Computational Statistics and Data Analysis**, 2016. v. 99, p. 81–90. Disponível em: <http://dx.doi.org/10.1016/j.csda.2016.01.011>.

SANTANA, F. B. De *et al.* Experimento didático de quimiometria para classificação de óleos vegetais comestíveis por espectroscopia no infravermelho médio combinado com análise discriminante por mínimos quadrados parciais: Um tutorial, parte V. **Quim. Nova**, 2020. v. 43, n. 3, p. 371–381.

SAPTORO, A.; TADÉ, M. O.; VUTHALURU, H. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. **Chemical Product and Process Modeling**, 2012. v. 7, n. 1.

SEM, V. Interpretability of selected variables and performance comparison of variable selection methods in a polyethylene and polypropylene NIR classification task. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, 2021. v. 258, p. 119850. Disponível em: <https://doi.org/10.1016/j.saa.2021.119850>.

SHAMSIPUR, M. *et al.* Ant colony optimisation : a powerful tool for wavelength selection. **Journal of chemometrics**, 2007. v. 20, p. 146–157.

SHARMA, V. *et al.* Multivariate analysis for forensic characterization , discrimination , and classification of marker pen inks. **Spectroscopy Letters**, 2018. v. 51, n. 5, p. 205–215.

SHENG, C.; MIAW, W.; SENA, M. M.; *et al.* Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification strategies. **Food Chemistry**, 2018.

SHENG, C.; MIAW, *et al.* Variable selection for multivariate classification aiming to detect individual adulterants and their blends in grape nectars. **Talanta**, 2018. v. 190, p. 55–61.

SHEYKHIZADEH, S.; NASERI, A. An efficient swarm intelligence approach to feature selection based on invasive weed optimization: Application to multivariate calibration and classification using spectroscopic data. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, 2018. Disponível em: <https://doi.org/10.1016/j.saa.2018.01.028>.

SILVA, N. C. D. *et al.* Classification of Brazilian and foreign gasolines adulterated with alcohol using infrared spectroscopy. **Forensic Science International**, 2015. v. 253, p. 33–42. Disponível em: <http://dx.doi.org/10.1016/j.forsciint.2015.05.011>.

SOARES, S. F. C. *et al.* The successive projections algorithm. **TrAC - Trends in Analytical Chemistry**, 2013. v. 42, p. 84–98.

SOUTO, U. T. C. P. *et al.* UV – Vis spectrometric classification of coffees by SPA – LDA. **Food Chemistry**, 2010. v. 119, p. 368–371.

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

TAYLOR, H. M.; KARLIN, S. **An Introduction To Stochastic Modeling**. 3ed ed. ed. San Diego: Elsevier, 1998.

TOLENTINO, M. C. *et al.* Avaliação da estabilidade foto-oxidativa dos óleos de canola e de milho em presença de antioxidantes sintéticos. **Ciencia Rural**, 2014. v. 44, n. 4, p. 728–733.

VALINGER, D. *et al.* Detection of honey adulteration – The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis. **LWT - Food Science and Technology**, 2021. v. 145.

VARMUZA, K.; FILZMOSER, P. Introduction to Multivariate Statistical Analysis in Chemometrics. **Applied Spectroscopy**. [S.l.]: CRC Press, 2009, V. 64, p. 112A-112A.

VISCONTI, L. G.; RODRÍGUEZ, M. S.; ANIBAL, C. V. Di. Determination of grated hard cheeses adulteration by near infrared spectroscopy (NIR) and multivariate analysis. **International Dairy Journal**, 2020. v. 104.

WANG, G.; GUO, L. A Novel Hybrid Bat Algorithm with Harmony Search for Global Numerical Optimization. **Journal of Applied Mathematics**, 2013. v. 2013, p. 1–21.

WENG, S. *et al.* Rapid detection of adulteration of minced beef using Vis / NIR re fl ectance spectroscopy with multivariate methods. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, 2020. v. 230.

WITTEN, D. M.; TIBSHIRANI, R. Penalized classification using Fisher's linear discriminant. **Journal of the Royal Statistical Society. Series B: Statistical Methodology**, 2011. v. 73, n. 5, p. 753–772.

XIAOWEI, H. *et al.* Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models. **Food Chemistry**, 2014. v. 164, p. 536–543.

XU, S. *et al.* Near infrared fluorescent dual ligand functionalized Au NCs based multidimensional sensor array for pattern recognition of multiple proteins and serum discrimination. **Biosensors and Bioelectronic**, 2017. v. 97, p. 203–207.

XUE, B.; ZHANG, M.; BROWNE, W. N. Particle swarm optimisation for feature selection in classification : Novel initialisation and updating mechanisms. **Applied Soft Computing**, 2014. v. 18, p. 261–276.

YANG, M. *et al.* Portable spectroscopy system determination of acid value in peanut oil based on variables selection algorithms. **Measurement**, 2017. v. 103, p. 179–185. Disponível em: <http://dx.doi.org/10.1016/j.measurement.2017.02.037>.

YANG, Q.; DONG, N.; ZHANG, J. An enhanced adaptive bat algorithm for microgrid energy scheduling. **Energy**, 2021. v. 232, p. 121014. Disponível em: <https://doi.org/10.1016/j.energy.2021.121014>.

YANG, Xin-She. A New Metaheuristic Bat-Inspired Algorithm. **Nature Inspired Cooperative Strategies for Optimization (NISCO 2010)**, 2010a. v. 284, p. 65–74.

YANG, Xin-She. Firefly algorithm, stochastic test functions and design optimisation. **Int. J. Bio-Inspired Computation**, 2010b. v. 2, n. 2, p. 78–84.

YUE, S.; ZHANG, H. A hybrid grasshopper optimization algorithm with bat algorithm for global optimization. **Multimedia Tools and Applications**, 2020. v. 80, p. 3863–3884.

YVES, I. *et al.* Starch adulteration in turmeric samples through multivariate analysis with infrared spectroscopy. **Food Chemistry**, 2021. v. 340.

ZHANG, D. *et al.* Nondestructive evaluation of soluble solids content in tomato with different stage by using Vis/NIR technology and multivariate algorithms. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, 2021. v. 248, p. 119139. Disponível em: <https://doi.org/10.1016/j.saa.2020.119139>.

ZHANG, Y. *et al.* Non-destructive recognition and classi fi cation of citrus fruit blemishes based on ant colony optimized spectral information. **Postharvest Biology and Technology**, 2018. v. 143, p. 119–128.

ZHANG, Y. *et al.* Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. **Geoderma**, 2019. v. 333, p. 23–34.

ZHU, B. *et al.* A Novel Quantum-Behaved Bat Algorithm with Mean Best Position Directed for Numerical Optimization. **Computational Intelligence and Neuroscience**, 2016. v. 2016, p. 1–17.

ANEXO A- Artigo publicado

Microchemical Journal 187 (2023) 108382

Contents lists available at ScienceDirect

Microchemical Journal

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

Bat algorithm for variable selection in multivariate classification modeling using linear discriminant analysis

Juliana da Cruz Souza [a], Sófacles F.C. Soares [b], Lauro Cássio M. de Paula [c], Clarimar J. Coelho [d], Mário César Ugulino de Araújo [a], Edvan Cirino da Silva [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] Universidade Federal da Paraíba, CT, Departamento de Engenharia Química, CEP 58051-900 João Pessoa, PB, Brazil
[c] Instituto Federal da Bahia, IFBA, Campus Santo Antônio de Jesus, Bahia, Brazil
[d] Pontifícia Universidade Católica de Goiás, Escola Politécnica, Goiás, Brazil

ARTICLE INFO

Keywords:
BA-LDA algorithm
Variable selection
Multivariate classification
Linear discriminant analysis

ABSTRACT

Variable selection is an efficient and powerful tool for reducing the dimensionality of multivariate data and multicollinearity, enabling the successful classification of samples by Linear Discriminant Analysis (LDA). This paper describes a bat-inspired algorithm as an alternative to performing variable selection in multivariate classification by LDA. Named BA-LDA, this algorithm simulates the echolocation behavior of bats when moving in search of prey. It was implemented with a cost function associated with the average risk of misclassification in LDA. The performance of BA-LDA was evaluated on mass spectrometry (MS), near-infrared (NIR), and ultraviolet–visible (UV–vis) spectrometric data sets of serum from unaffected and affected women with ovarian cancer, coffee, and vegetable oil samples, respectively. Its performance was compared with the genetic algorithm (GA-LDA) and successive projection algorithm (SPA-LDA). As the main results, BA-LDA presented a classification performance similar to the GA-LDA and SPA-LDA, classifying all coffee and vegetable oil samples. For the ovarian cancer dataset, BA-LDA (93%

ANEXO B

A seguir constam os certificados de apresentação de trabalhos em eventos científicos e de premiações.

Certificado

19º Encontro Nacional de Química Analítica e
7º Congresso Iberoamericano de Química Analítica

Inovação para a Química Analítica Sustentável

16 a 19 de Setembro de 2018

Centro de Eventos e Convenções DiRoma, Caldas Novas - GO

Certificamos que

o trabalho intitulado: UM ALGORITMO BIO-INSPIRADO NOS MORCEGOS PARA SELEÇÃO DE VARIÁVEIS ESPECTRAIS EM PROBLEMAS DE CLASSIFICAÇÃO, de autoria de Juliana da Cruz Souza, Sófacles Figueredo C. Soares, Lauro Cássio M. Paula, Clarimar J. Coelho, Edvan Cirino da Silva, Mario Cesar U. Araujo, foi apresentado na modalidade 9. Quimiometria, durante o 19º ENQA - Encontro Nacional de Química Analítica e o 7º CIAQA - Congresso Iberoamericano de Química Analítica, realizado de 16 a 19 de setembro de 2018 no Centro de Eventos e Convenções do DiRoma em Caldas Novas/GO.

Caldas Novas, 19 de setembro de 2018.

Wendell Karlos Tomazelli Coltro
Presidente do 19º ENQA e 7º CIAQA

IV ESCOLA DE INVERNO DE QUIMIOMETRIA

20 a 23 AGOSTO 2019 PORTO ALEGRE/RS

Certificamos que o trabalho **CLASSIFICAÇÃO DE ÓLEOS VEGETAIS E CAFÉS USANDO O ALGORITMO INSPIRADO NOS MORCEGOS PARA SELEÇÃO DE VARIÁVEIS EM LDA** foi apresentado na forma de painel durante as sessões de pôster da IV Escola de Inverno de Quimiometria.

Autores: **Juliana da Cruz Souza, Sófacles F. C. Soares, Lauro Cássio M. Paula, Clarimar J. Coelho, Edvan Cirino da Silva, Mario Cesar U. Araujo**

Porto Alegre, 23 de agosto de 2019.

Prof. Marco Flôres Ferrão
Presidente da Comissão Organizadora

Realização: PPGQ UFRGS · Instituto de Química UFRGS · UFRGS · CAPES

IV ESCOLA DE INVERNO DE QUIMIOMETRIA

20 a 23 AGOSTO 2019 PORTO ALEGRE/RS

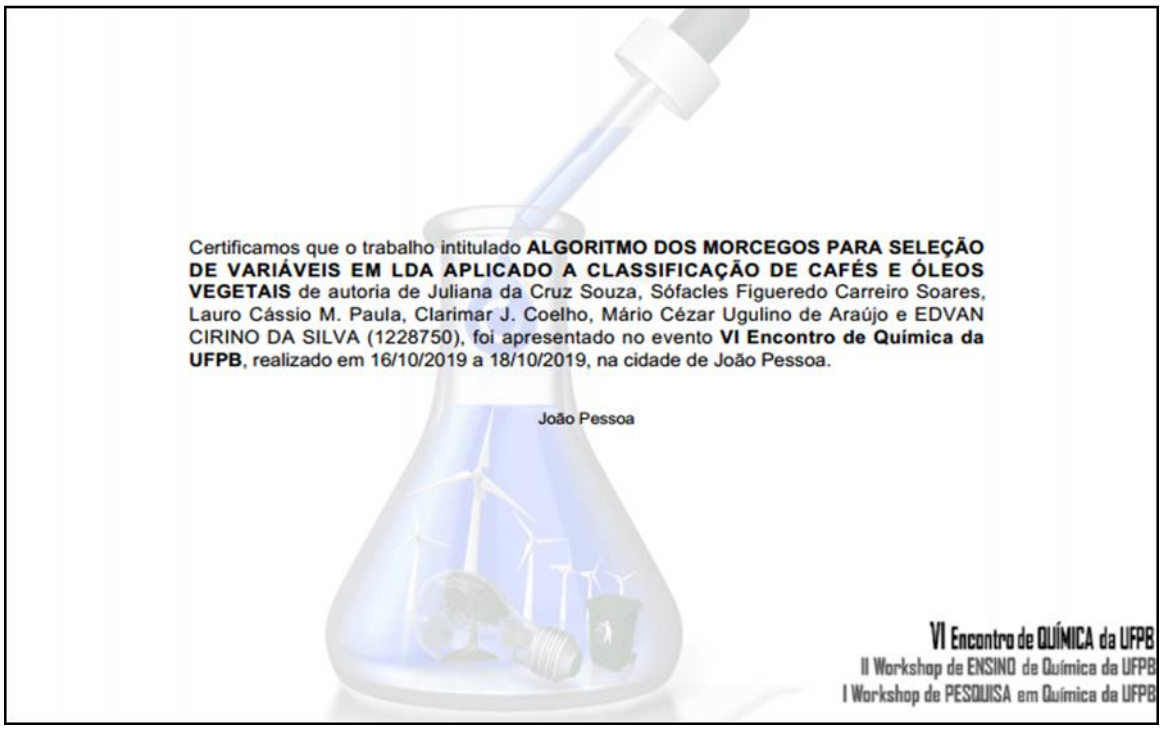
Certificamos que o trabalho intitulado **CLASSIFICAÇÃO DE ÓLEOS VEGETAIS E CAFÉS USANDO O ALGORITMO INSPIRADO NOS MORCEGOS PARA SELEÇÃO DE VARIÁVEIS EM LDA** foi premiado pela comissão da **IV Escola de Inverno de Quimiometria.**

Autores: **Juliana da Cruz Souza, Sófacles F. C. Soares, Lauro Cássio M. Paula, Clarimar J. Coelho, Edvan Cirino da Silva, Mario Cesar U. Araujo**

Porto Alegre, 23 de agosto de 2019.

Prof. Marco Flôres Ferrão
Presidente da Comissão Organizadora

Realização: PPGQ UFRGS · Instituto de Química UFRGS · UFRGS · CAPES



Certificamos que o trabalho intitulado **ALGORITMO DOS MORCEGOS PARA SELEÇÃO DE VARIÁVEIS EM LDA APLICADO A CLASSIFICAÇÃO DE CAFÉS E ÓLEOS VEGETAIS** de autoria de Juliana da Cruz Souza, Sófacles Figueredo Carreiro Soares, Lauro Cássio M. Paula, Clarimar J. Coelho, Mário Cézar Ugulino de Araújo e EDVAN CIRINO DA SILVA (1228750), foi apresentado no evento **VI Encontro de Química da UFPB**, realizado em 16/10/2019 a 18/10/2019, na cidade de João Pessoa.

João Pessoa

VI Encontro de QUÍMICA da UFPB
II Workshop de ENSINO de Química da UFPB
I Workshop de PESQUISA em Química da UFPB

ANEXOS C- Código fonte do BA-LDA

O código do BA-LDA é apresentado a seguir.

```
function [var_sel,class_train,class_val,class_test] =
batclassification_v2(Train,Group_Train,Val,Group_Val,Test,Group_Test,n_min,
n_max,validation)
%
% batclassification - Selection of decision variables for Linear
Discriminant Analysis (LDA) employing
% the metaheuristic bat-inspired algorithm
%
% [var_sel,class_val,class_test] =
batclassification_v2(Train,Group_Train,Val,Group_Val,Test,Group_Test,n_min,
n_max,validation)
%
% INPUT
%
% Train       --> Matrix of training objects (#training objects x
#variables)
% Val         --> Matrix of validation objects (#validation objects x
#variables)
% Test        --> Matrix of test objects (#test objects x #variables)
% Group_Train --> Column vector with the class indexes (1, 2, ...) for each
training object
% Group_Val   --> Column vector with the class indexes (1, 2, ...) for each
validation object
% Group_Test  --> Column vector with the class indexes (1, 2, ...) for each
test object
% n_min       --> Lower bound on the number of decision variables
% n_max       --> Upper bound on the number of decision variables
% validation  --> 1-Test set, 2-Internal validation, and 3-cross-validation
%
% OUTPUT
% var_sel     --> Selected variables
% class_val   --> Columns vector with the classification of the
corresponding row of Val set
% class_test  --> Columns vector with the classification of the
corresponding row of Test set
% class_test  --> Columns vector with the classification of the
corresponding row of Training set
%
% Author: Sófacles Figueredo Carreiro Soares
% Email: sofacles@gmail.com
% Last modified on 18 July 2018
% Versin: 2.0
%
% References:
%
% [1] PONTES, M. J. C.; GALVÃO, R. K. H. ; ARAÚJO, M. C. U.; MOREIRA, P. N.
T. ; PESSOA NETO, O. D.; JOSÉ, G. E.; SALDANHA, T. C. B.
%      The Successive Projections Algorithm for Spectral Variable Selection
in Classification Problems.
%      Chemometrics and Intelligent Laboratory Systems, v. 78, n. 1-2, p.
11-18, 2005.
% [2] Soares, S. F. C; Galvão, R. K. H.; Pontes, M. J. C.; Araújo, M. C. U;
A New Validation Criterion for Guiding the Selection of
%     Variables by the Successive Projections Algorithm in Classification
%     Problems. J. Braz. Chem. Soc., Vol. 25, No. 1, 176-181, 2014
%
```

```
[mtrain,Nlambdas] = size(Train);      % Xtrain dimensions(number of
training objects and variables)
[mval,nval]       = size(Val);        % Xval dimensions(number of training
objects and variables)
C                 = max(Group_Train); % Number of classes

alpha = 5/10; % Loudness control parameter
gamma = 4/10; % Emission rate control parameter
mbats = 30;       % Number of bats
numberofiterations = 200; % Number of iterations

costallbats = zeros(1, numberofiterations);  % Average cost for all bats
sigma = 0.5;     % Threshold for binarization of variables

limit_lower = 0; % Lower limit of the domain search
limit_upper = 1; % Upper limit of the domain search

fmin     = 0;     % Minimum frequency
fmax     = 5/100; % Maximum frequency

% initializing the population of the bats. The number of positions bigger
than sigma is between n_min and n_max
Xold = zeros(mbats,Nlambdas);
for i = 1:mbats
   N = n_min + sum(round(rand(1,n_max - n_min)));
   temp = randperm(Nlambdas); % Auxiliary matrix
   lh = temp([1:N]); % variables bigger than sigma;
   Xold(i,lh) = sigma + (1-sigma)*rand(N,1);
   ll = setdiff(1:Nlambdas,lh);
   Xold(i,ll) = sigma*rand(length(ll),1);
end

Xnew = zeros(mbats, Nlambdas); % News positions
v = zeros(mbats, Nlambdas); % Initial velocity
r = round(rand(mbats, 1) * 1000) / 1000; % Pulse emissition rate matrix
A = ones(mbats, 1);                      % Loudness matrix

r0 = r; % t0 emissition rate
Xbin = Xold > sigma; % Binarization of the initial position
% Choosing the validation
if (validation == 1)  % Test set
    disp('The test set will be used')
    Nobjects = mval;
    ObjectsVal = Val;
    Group_ObjectsVal = Group_Val;
    if n_max > mtrain - C
        error('The number of decision variables in LDA cannot be larger
than the number of training objects minus the number of classes');
    end
elseif (validation == 2 || validation == 3) % Internal validation
    disp('The internal validation will be used')
    Nobjects = mtrain;
    ObjectsVal = Train;
    Group_ObjectsVal = Group_Train;
    if n_max > mtrain - C - 1
        error('The number of decision variables in LDA cannot be larger
than the number of training objects minus the number of classes minus
one');
```

```
    end

else
    error('The validation was not chosen correctly')
end


% Initial G cost
[number_of_var,cost] =
gcost(Train,Group_Train,ObjectsVal,Group_ObjectsVal,Xbin,validation,n_min,n
_max);
[~, index] = min(cost); % lower cost
Xbest = Xold(index, :); % Position of bat with lower cost


for t = 1 : numberofiterations % critério de convergência seja atingido

    mean_loudness = mean(A); % Loudness average

    for i = 1 : mbats, % Loop for generation
        beta       = round(rand * 1000) / 1000; % beta = [0, 1]
        fi         = fmin + ((fmax - fmin) * beta); % Adjusting frequency
        v(i, :)    = v(i, :) + (Xold(i, :) - Xbest) * fi; % Updating
velocities
        Xnew(i, :) = v(i, :) + Xold(i, :); % Updating solutions
        if rand > r(i) % Generate a local solution
            epsilon    =  2 * round(rand * 1000) / 1000 - 1; % epsilon [-
1,1]
            Xnew(i, :) =  Xnew(i, :) + epsilon * exp(mean_loudness); %New
positions
        end
    end

    %Avaliation of limits
    Xnew(Xnew < limit_lower) = limit_lower;
    Xnew(Xnew > limit_upper) = limit_upper;

    Xbin = Xnew > sigma;
    [new_number_of_var, new_cost] =
gcost(Train,Group_Train,ObjectsVal,Group_ObjectsVal,Xbin,validation,n_min,n
_max);
    [~, index] = min(new_cost);

    for i = 1 : mbats, % loop sobre todas as soluções
        if (rand < A(i)) && (new_cost(i) < cost(i))
            Xold(i, :) = Xnew(i, :);
            number_of_var(i) = new_number_of_var(i);
            cost(i) = new_cost(i);
            A(i) = alpha * A(i); % Reduce Ai
            r(i) = r0(i) * (1 - exp(-gamma * t)); % Increase ri
        end

    end
    [~, index] = min(cost);
    Xbest = Xold(index,:);
    costallbats(t) = mean(cost); % Average cost associated to the all bats
in the instant t
end
%
```

```
figure(1),plot(1:t,costallbats,'b')
xlabel('Number of iterations')
ylabel('Average cost for all bats')
Xbin = Xold > sigma;

[number_of_var, cost] =
gcost(Train,Group_Train,ObjectsVal,Group_ObjectsVal,Xbin,validation,n_min,n
_max);
[~, c] = min(cost);
number_of_sel_var = number_of_var(c);
cost = cost(c);

var_sel = find(Xbin(c,:));
mt = mean(Train);
figure(2), plot(var_sel,mt(:,var_sel),'bo',1:Nlambdas,mt,'k-')
xlabel('Selected Variables')
ylabel('Signal')
Train2 = Train(:,var_sel);
ObjectsVal2 = ObjectsVal(:,var_sel);
Test2  = Test(:,var_sel);
[class_train,errors_train] =
multilda(Train2,Train2,Group_Train,Group_Train);
[class_test,errors_test] = multilda(Test2,Train2,Group_Train,Group_Test);
[class_val,errors_val] =
multilda(ObjectsVal2,Train2,Group_Train,Group_ObjectsVal);
disp(['Number of variables selected: ' num2str(number_of_sel_var)])
disp(['Number of errors in the training set: ' num2str(errors_train)])
disp(['Number of validation errors: ' num2str(errors_val)])
disp(['Number of errors in the test set: ' num2str(errors_test)])


function [Number_of_var,cost] =
gcost(Train,Group_Train,ObjectsVal,Group_ObjectsVal,Xbin,validation,n_min,n
_max)

C = max(Group_Train);   % Number of classes
Nobjects = size(ObjectsVal);
mtrain = size(Train,1);
for i = 1 : size(Xbin,1) % beginning from the i-th variable (xi)
    lambdas = find(Xbin(i, :));
    % Test for the number de variables out of [n_min n_max]
    Number_of_var(i) = length(lambdas);
    if ((size(lambdas,2) > n_max) | (size(lambdas,2) < n_min))
        cost(i,:) = 1.1;
    else
    custoaux = 0;
    if (validation ~= 3)
        [Xpooled,media] = mcdata(Train(:,lambdas),Group_Train);
        S = cov(Xpooled,1);% Pooled covariance matrix
        invS = inv(S); % Inverse of the pooled covariance matrix
    end

    for objecttest = 1:Nobjects % For each validation object
        x = ObjectsVal(objecttest,lambdas);
        groupx = Group_ObjectsVal(objecttest); % True class index
        if (validation == 3) % Cross-validation
            Objectscv = Train([1:objecttest-
1,objecttest+1:mtrain],lambdas);
            Group_Objectscv = Group_Train([1:objecttest-
1,objecttest+1:mtrain]);
```

```
            [Xpooled,media] = mcdata(Objectscv,Group_Objectscv);
            S = cov(Xpooled,1); % Pooled covariance matrix
            invS = inv(S); % Inverse of the pooled covariance matrix
        end
        for k = 1:C  % For each class
            mu = media{k};
            r(k) = (x - mu)*invS*(x - mu)';
        end
        num = r(groupx); % Mahalanobis distance to the correct class
        remaining = setdiff([1:C],groupx); % Remaining classes
        den = min(r(remaining)); % Smallest Mahalanobis distance to the
remaining classes
        custoaux = custoaux + num/den;
    end
     if (validation == 2) % Average cost associated to the subset of
variables for bat i
        cost(i,:) = custoaux/(size(Group_ObjectsVal,1)-Number_of_var(i)-C);
     else
        cost(i,:) = custoaux/size(Group_ObjectsVal,1);
     end
   end
end


% mean centering
function [Xpooled,media] = mcdata(Train,Group_Train)
C = max(Group_Train); % Number of classes
Xpooled = []; % Xpooled will be employed in the calculation of the Pooled
Covariance Matrix
for k = 1:C
    index{k} = find(Group_Train == k); % Training objects belong to k-th
class
    Traink = Train(index{k},:);
    media{k} = mean(Traink);
    Trainkc = Traink - repmat(media{k},size(Traink,1),1);
    Xpooled = [Xpooled;Trainkc]; % Xpooled contains the objects centered on
the mean of their respective classes
end
```